JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Performance Comparison of Machine Learning Techniques in Intrusion Detection using Rapid Miner

Sanjeet Choudhary[1*], Varsha Namdeo[2], Abhijit Dwivedi[3]

[1,2,3]Computer Science & Engineering, RKDF Institute of Science & Technology, Bhopal, India

[*]*Corresponding Author: sanjeetchoudhary87@gmail.com, Tel.: +91-8109836975*

Available online at: www.ijcseonline.org

Accepted: 26/Nov/2018, Published: 30/Nov/2018

***Abstract*—** Information security is becoming a more important issue for modern computer generation, progressively. Intrusion Detection System (IDS) as the main security defensive technique and is widely used against many category of attacks. Intrusion Detection Systems are used to detect various kinds of attack in very large datasets. Data Mining (DM) and Machine Learning (ML) techniques are powerful enough and proved useful in the Network Intrusion Detection research area. In Recent years, many ML methods have also been introduced by researchers, to obtain high accuracy and good detection rate. A potential drawback of all those methods is how to classify different intrusion attacks effectively. Looking at such inadequacies, the RapidMiner tool is tested for the few ML techniques in this work. As most of the research works using tools like MATLAB, WEKA etc. the purpose of this work is to test and evaluate the ML techniques on RapidMiner. This paper presents a performance comparison of three ML techniques including: K-NN, Decision tree, Naïve Bayes using RapidMiner tool. This paper will provide an insight for the future research. The techniques were tested based on Detection rate and False Alarm rate. The result analysis and evaluation obtained by applying these approaches to the KDD CUP'99 data set.

***Keywords*—** Intrusion Detection System (IDS), Data Mining, Classification, Data Science, Machine Learning, RapidMiner, Security, KDD CUP’99

I. INTRODUCTION

Secure information either in private or government sector has become an essential requirement. System vulnerabilities and valuable information catch the attention of most attackers’. All attack types are not prevented using traditional intrusion detection approaches like firewalls and encryption. The number of attacks through network and other medium has increased dramatically in recent years. Efficient intrusion detection is needed as a security cover against these malicious or suspicious and abnormal activities. Thus, classification in Intrusion Detection Systems (IDSs) has been introduced as a security technique to classify & detect various attacks. The misuse detection and anomaly detection are two different techniques identified in IDSs. Misuse detection techniques can detect known attacks by examining attack patterns, similar to virus detection by an antivirus application. This technique require updated attack pattern to detect unknown attacks. On the other hand, anomaly detection identifies any unusual activity pattern which deviates from the normal usage, as intrusion. Although anomaly detection has the ability to detect unknown attacks which cannot be addressed by misuse detection, it suffers from high false alarm rate.

In recent years, ML techniques received much attention to overcome the constraint of traditional IDSs by increasing accuracy and detection rates. In literature, numbers of IDSs are developed based on many different ML techniques such as neural networks, support vector machines and genetic algorithms etc. These techniques are developed as classifiers, which are used to classify or distinguish whether the incoming Internet access is the normal access or an attack. This work aims at implementing three different ML models for IDS using Naïve Bayes, Decision Tree and k-Nearest Neighbor in most advanced RapidMIner tool.

The rest of the paper is organized as follows. Section 2 presents the review of literature on the topic. Section 3 presents the dataset and tool description used for experiment. The experimental setup and major experiment steps are also shown in section 3. A performance evaluation of implemented techniques, done using famous RapidMiner tool, is meticulously shown in Section 4. Finally Section 5 concludes the paper giving future directions.

II. Related Work

Hua TANG et al. [1] proposed a new approach to detect network attacks, which aims to study the efficiency of the method based on ML in intrusion detection, including artificial neural networks and support vector machine. The experimental results obtained by applying this approach to the KDD CUP'99 data set demonstrate that the proposed approach performs high performance, especially to U2R attacks and R2L type.

According to the authors [2], neural networks, SVM and decision trees are the popular schemes borrowed from Machine learning community into IDS. In it these three techniques are compared by applying on KDD CUP'99 data set. The ML approaches are supposed to be fit to identify the anomalies detection, in an appropriate way by proper training but the performance may be variable in terms of different algorithms.

Chi Cheng et al. [3] proposed Extreme Learning Machines methods to classify binary and multi-class network traffic for intrusion detection. In this work the performance of ELM in both binary-class and multi-class scenarios are investigated, and compared to SVM based classifiers. Simulation results on KDD CUP'99 data set show that the proposed method can detect intrusions even in large datasets with short training and testing times.

The work [4] presents a neural-network-based active learning procedure for computer network intrusion detection. As applying DM and ML techniques to network intrusion detection often faces the problem of very large training dataset size the active learning procedure can noticeably reduce the size of the training data, without significantly sacrificing the classification accuracy of the intrusion detection model. A comparison of the with a C4.5 decision tree indicated that the actively learned model had better generalization accuracy.

The authors of [5] evaluated the performance of a ML algorithm called Decision Tree and compared with two other ML algorithms namely Neural Network and Support Vector Machines. The algorithms were tested on basis of accuracy, detection rate, false alarm rate and detection accuracy of all four attack types. From the experiments conducted, authors found that the Decision tree algorithm outperformed the other two algorithms. In this research, we intend to compare the efficiency of Neural Networks, Support Vector Machines and Decision Tree algorithms against KDD-cup dataset.

Jingbo Yuan et al. [6] first introduces the basic structure of the intrusion detection system, then analyzed the intrusion Detection Techniques Based on ML Method, including the Bayesian based method, the neural network based method, the DM based method and the SVM based method.

The authors in [7], aim to use DM techniques including classification tree and support vector machines for intrusion detection. As their results indicate, C4.5 algorithm is better than SVM in detecting network intrusions and false alarm rate in KDD CUP 99 dataset.

Anand Motwani et al. [8] proposed an intrusion classification framework based on Optimal Sampling for Class Balancing sampling to improve the classification performance. This framework is tested with three different ML algorithms along with optimal sampling. The proposed work is tested on basis of Accuracy, Error rate, Detection rate and False Alarm rate. The model is helpful in detecting intrusions even in large datasets with short training and testing times.

Based on the survey, in this paper we evaluate the performance of a comprehensive set of classifier algorithms using KDD CUP’99 dataset. Based on evaluation results, three classifier algorithms are compared in our work.

III. DATASET DESCRIPTION AND EXPERIMENTAL FRAMEWORK

A. KDD CUP 1999 DATASET

KDD CUP 1999 Data (KDD99) is the dataset used in the evaluate ML technique. In practice, we recognize that this dataset is more than a decade old and has many criticisms for Current research. But we believe that it is still sufficient for our experiment which aims to reflect the performance of distinct ML approaches in a general way and find out relevant issues and also give future research directions. In addition, the full KDD99 dataset Contain 4,898,431 records and each record contain 41 features [9]. Due to the computing power, we do not use the full dataset of KDD99 in the experiment but a 10% portion use of it. This 10% KDD99 dataset contains 494,069 records (each with 41 features) and 4 categories of attacks. The details of attack categories and specific types are shown in Table 1. The four attack types are [2, 5, 7]:

1) *Probing: Scan networks to gather deeper information*

2) *Denial of service (DoS): such attacks make computing or memory resource too busy or full that it denies legitimate users access to a machine.*

3) *User to Root (U2R): Illegal access to gain super user privileges*

4) *Remote to User (R2L): Illegal access from a remote machine.*

Table 1, Attack categories and types in KDD dataset

No.	**Four Attack Categories**			
	Probe	***DoS***	***U2R***	***R2L***
1	ipsweep	back	buffer overflow	ftp write
2	nmap	land	loadmodul	Guess passwd
3	Portswee	neptune	Perl	Imap
4	Satan	pod	Rootkit	Mutihop
5	---	smurf	---	Phf
6	---	teardrop	---	Spy
7	---	---	---	Warezmaster
8	---	---	---	warezmaster

Every attack categories contain some specific attack types [2]. For example, DoS has 6 specific attack types (e.g. back, land, neptune), R2L has 8 specific attack types (e.g. ftp write, guess password, imap). There are totally 22 specific attack types within the 10% KDD99 dataset, while the full KDD99 dataset has 39 specific attack types. Although the number of specific attack types is different between 10% KDD99 dataset and full KDD99 dataset, we believe that there are no negative effects on our evaluation purpose.

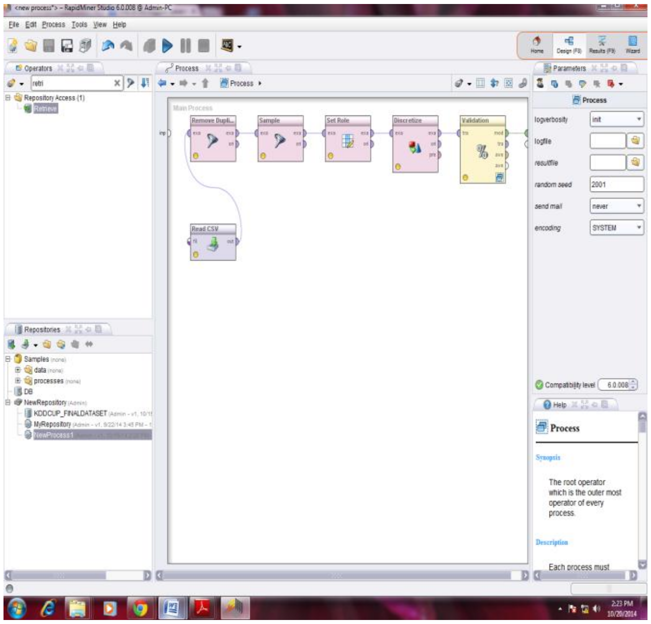
B. *RAPIDMINER TOOL DESCRIPTION*

The models based on DM techniques are demonstrated using variety of languages like Python, Java and tools like Weka and RapidMiner [10]. RapidMiner [11] is one the world-leading open source systems for data mining solutions, due to the blend of its functional range and applications. It serves as standalone application for data analysis and as DM solution for industries and researchers. A huge amount of visualization techniques and operators used in it gives insight into the progress for running experiments. Although the main application of RapidMiner lies in the area of inferential statistics, it also provides the best combination of numerous preprocessing and learning steps. One of the biggest advantages is without doubt the fact that RapidMiner is available for free download in the Community Edition. Students can therefore install it on their private computers in just the same way as the university can make RapidMiner installations available on institute computers.

C. *EXPERIMENTAL FRAMEWORK*

Firstly, we build the experiment environment in RapidMiner for evaluation, with major steps: environment setup, data preprocessing, choosing the classifier. Figure 1 and 2 shows the Experimental framework for this work. Secondly, we select a comprehensive set of most popular classifier algorithms, three distinct widely used classifier algorithms were selected so that they represent a wide variety of fields: Bayesian approaches, decision trees, and lazy functions. Finally, we come up with the performance comparison between the selected classifiers in next section.

In order to verify the effectiveness of different classifiers for the field of intrusion detection, we will use the KDD99 dataset to make relevant experiments step-by-step. For this purpose we used RapidMiner tool, a brief description of which is given below.



Process

logverbosity init

random seed 2001

send mail never

encoding SYSTEM

Process

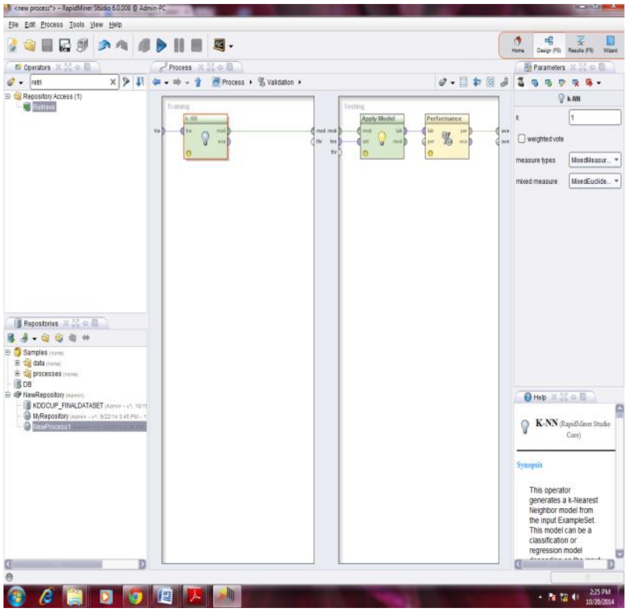
Synopsis

The root operator which is the outer most operator of every process.

Description

Each process must

Figure 1. Snapshot of experiment



k-NN

k 1

□ weighted vote

measure types MixedMeasur...

mixed measure MixedEuclide...

K-NN (RapidMiner Studio Core)

Synopsis

This operator generates a k-Nearest Neighbor model from the input ExampleSet. This model can be a classification or regression model

Figure 2. Snapshot of Experiment

IV. ANALYSIS AND EVALUATION

A representative and frequent task in the area of ML is comparing two or more learning procedures with one another. This can be done to study the improvements that can be obtained by new procedures, and also simply be used to select a suitable technique for IDSs. In this section we will show how this can be done with RapidMiner. The detection of attacks can be measured by following metrics [5, 7]:

A. PERFORMANCE METRICS

- **True Positive (TP):** When, the number of found instances for attacks is actually attacks.
- **False Positive (FP):** When, the number of found instances for attacks is normal.
- **True Negative:** When, the number of found instances is normal data and it is actually normal.
- **False Negative:** When, the number of found instances is detected as normal data but it is actually attack.
- The accuracy of IDS is measured generally on basis of following parameters:
- **Detection Rate:** Detection rate refers to the percentage of detected Attack among all attack data, and is defined as follows:

$$\text{Detection rate} = \frac{TP}{TP + TN} * 100$$

- With this formula detection rate for different types of Attacks can be calculated.
- **False Alarm rate:** It refers to the percentage of normal data which is wrongly recognized as attack. The formula represented below as:

$$\text{False Alarm rate} = \frac{FP}{FP + TN} * 100$$

B. COMPARISON WITH DIFFERENT MACHINE LEARNING TECHNIQUES

An overview of how specific values of these algorithms were identified as well as their detection performance is mentioned in Table 2. In results it is shown that no single algorithm could detect all attack classes with a high detection rate and a low false alarm rate. It reinforce our belief that different algorithms should be used to deal with dissimilar types of network attacks. Results also show that for a given attack category, certain algorithms shows superior detection performance compared to others. For DoS category, most algorithms provide very high TP rates – averagely 92%. NaïveBayes is the only one that lags as it gives a TP at 81.2%. But for Probe attacks, NaïveBayes outperforms the others with its TP at 95.3%; Decision Tree has impressive performance for this category at 97.6%. In U2R attacks, k-NN and Decision Tree are the best two classifiers with FP at 0.7 and 1.1 respectively. And for the case of R2L attacks, k-NN could produce about 9% of attacks while the others just lag behind with inappreciable results.

Table 2, Performance comparison of three Classifiers

Classifier Category	**Classifier Algorithm**		**DoS**	**Probe**	**U2R**	**R2L**
Bayes	NaïveBayes	TP	81.2	95.3	13.1	0.1
		FP	1.49	12.2	0.8	0.3
Trees	Decision Tree	TP	97.6	74.5	1.3	0.1
		FP	1.1	1.0	0.1	0.4
Lazy	k-NN	TP	96.8	73.5	22.9	8.1
		FP	0.7	0.2	0.1	0.5

V. CONCLUSION AND FUTURE SCOPE

Intrusion Detection System (IDS) as the main security defensive technique in Information security. In recent years ML techniques proved useful and attracted increasing attention in the network intrusion detection research area. Looking at the need of classification of correct attack types, the RapidMiner tool is tested for the few ML techniques including: K-NN, Decision tree and Bayesian is presented in this work. The performance comparison of three ML techniques on KDD Cup 1999 Intrusion Data (KDD99) using RapidMiner tool is presented. Also, as most of the research works using tools like MATLAB, WEKA etc. are available. The purpose of this work is to test and evaluate the ML techniques on RapidMiner. The work also indicated that the actively learned model had better classification accuracy. The classifiers were tested based on Detection rate and Accuracy. Results also show that for a given attack category, certain algorithms shows superior detection performance compared to others.

At the same time, the factor such as ever growing amount of data for classification and constraints on response time, have made DM tasks a challenging job in the IDSs domain. So, to override the constraints on size of data to be classified and computational performance, the choices of cloud computing platforms for Intrusion detection is available. When making scientific predictions, Machine Learning has unique ability to evaluate large number of variables than a human possibly could do. Again ML is a time consuming task, so Cloud computing paradigm proved to be an important alternatives to speed-up ML tasks. Combining the advantages like handling large volume of data, speed of execution, scalability and use of exciting new technologies like Azure ML studio help to prevent critical security issues. In future, we propose Classification Frameworks for Network Intrusion Prediction in real Cloud.

This work will definitely provide an important reference for the future research.

REFERENCES

[1] Hua TANG, Zhuolin CAO "Machine Learning-based Intrusion Detection Algorithms" Journal of Computational Information Systems5:6(2009) 1825-1831 Available at http://www.JofCI.org

[2] YU-XIN MENG "The Practice on Using Machine Learning For Network Anomaly Intrusion Detection" 2011 IEEE

[3] Chi Cheng, Wee Peng Tay and Guang-Bin Huang "Extreme Learning Machines for Intrusion Detection" - WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia

[4] Naeem Seliya , Taghi M. Khoshgoftaar "Active Learning with Neural Networks for Intrusion Detection" IEEE IRI 2010, August 4-6, 2010, Las Vegas, Nevada, USA 978-1-4244-8099-9/10/2010 IEEE

[5] Kamarularifin Abd Jalill, Mohamad Noorman Masrek "Comparison of Machine Learning Algorithms Performance in Detecting Network Intrusion" 201O International Conference on Networking and Information Technology 978-1-4244-7578-0, 2010 IEEE

[6] Jingbo Yuan , Haixiao Li, Shunli Ding , Limin Cao "Intrusion Detection Model based on Improved Support Vector Machine" Third International Symposium on Intelligent Information Technology and Security Informatics 978-0-7695-4020-7/10, 2010 IEEE

[7] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey "Intrusion Detection Using Data Mining Techniques" IEEE 2010.

[8] Anand Motwani, Vaibhav Patel, Anita Yadav, "Optimal Sampling for Class Balancing with Machine Learning Technique for Intrusion Detection System", International Journal of Electrical, Electronics and Computer Engineering 4(2): 47-51(2015)

[9] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani "A Detailed Analysis of the KDD CUP 99 Data Set" Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009).

[10] Stuti Nathaniel, Anand Motwani, Arpit saxena, "Cloud based Predictive Model for Detection of 'Chronic Kidney Disease' Risk", International Journal of Computer Sciences and Engineering, Vol.6, Issue.4, pp.185-188, 2018

[11] http://www.rapidminer.com/