AN ANALYSIS ON NSL-KDD DATASET USING MACHINE LEARNING TECHNIQUES FOR INTRUSION DETECTION

Krishantha M Ranaweera[1] and Edirisingha Dilini Sakuntala[1]

[1]Department of ICT, Faculty of Technological Studies, University of Vavuniya

Abstract

With the growth of the usage of computers over the network, security vulnerabilities on all the computer systems seem very difficult and expensive. The Intrusion Detection System (IDS) generates huge numbers of false alerts. Therefore, it is necessary to assist in categorizing the degree of threat by using data mining techniques. We have used the NSL-KDD dataset for this research study. The response time was found to be high when the complexity of the dataset is high. Therefore, we have utilized Infogain feature selection algorithms. Four machine learning classification algorithms such as Sequential Minimal Optimization, Nave Bayes, J48, and Random Forest are utilized for this study. The Random Forest scores the best accuracy at 99.9%. However, J48 was chosen with an accuracy score of 99.8% with a minimum response time. .

Keywords: false positives, Intrusion detection, Machine learning, Data mining, Classification

1 INTRODUCTION

In this information-driven era, we have many applications with massive amounts of data that operate in real-time to achieve their objectives. As a result, storage capacity and processing time are constrained. With the growth of Internet usage and the invention of many task-specific applications, the information communication between channels have drastically increased. In the meantime, the number of users and the number of devices connected are accelerating year by year. The demand of applications for the new devices and gadgets are increasing, and they are being adopted in the market. A growing number of Machine to Machine (M2M) applications, such as smart meters, video surveillance, healthcare monitoring, transportation, and package or asset tracking, are contributing significantly to the proliferation of devices and connections [1]. By 2023, M2M connections will be half or 50 percent of the total devices and connections. In this juncture, the devices with different factors for developed capabilities and intelligence are introduced and adopted in the contemporary era.

With the growth of the Internet, and applications, malicious activities on user's information and privacy are on the rise as the degree of information technology and Internet usage rises. With these malicious activities and threats, it has become important to make our information systems secure, especially in the defense, commercial, banking and public sector. In addition, there are a slew of new cyber risks emerging all the time. Malicious acts on these data would be a tragedy in terms of securing the country's personal and secret information, which could lead to horrible disasters. As a result, a

variety of techniques have been implemented to combat malicious activity. The attacks happen as a result of aspects such as confidentiality, integrity, and availability failing.

Typically, intrusions are discovered after the fact, raising concerns about information security. The relevance of network security and user privacy is critical as applications are being used as a conduit to obtain user information. Identification of intrusions with the use of computer hardware is an economy challengeable task. Hence, the software-based solutions have been identified. This provides the rise to the cyber security, which involves protecting the system information by effectively preventing detecting and responding to attacks and used in different disciplines such as computer systems and networks within the sectors of communication, entertainment, transportation, shopping and medicine [2]. The risks with various intensities are used with embedded viruses to attack the computer systems and networks and can lead to different threats such as delete one's entire system, allow someone to penetrate someone else's system, alter files, attack other computers from someone's computer, allow someone to steal another person's credit card information and make unauthorized purchases. These cautious activities can be detected by gathering intrusion related knowledge occurring during system monitoring and then analyzing collected data as Intrusion Detection.

An Intrusion Detection System (IDS) is a cyber-security tool commonly deployed as part of a defense in depth cyber security strategy [3]. Intrusion Detection System (IDS) has been widely deployed to be a second line of defense for computer network systems along with other network security techniques such as firewall and access control. These tools are used to detect unauthorized use, misuse and abuse of computer systems by both system insiders and external intruders. Gathering and analyzing the intrusion related data manually would be a time-consuming task. Hence, the automated techniques can be used to boost the intrusion detection. Data Mining (DM) techniques are increasingly being used to identify attacks, anomalies or intrusions in a protected network environment [4]. On the other hand, machine learning techniques are used to automate the process of knowledge acquisition from past incidents of intrusion attacks. With the use of data mining and machine learning techniques of intrusion identification shows a maximum percentage of correct identification. However, with the rapid changes of attacks and hacking techniques, researchers need to find better techniques with highest accuracy. In order to detect intrusions on a network IDS use misuse -based or signature-based or anomaly-based detection techniques.

The study focuses on implementing an efficient model using classification algorithms to detect intrusions accurately. Four major machine learning classification algorithms are used to perform this study. The model's performance is measured and evaluated. The rest of the sections are organized as follows. Section 2 describes background related to our study including IDS, Machine learning techniques. Section 3 describes related works done with the same problem. Section 4 presents methodologies involving selecting the dataset, preprocessing, and feature selection and classification techniques. Section 5 describes the results and section 6 presents the conclusion related to this study.

2 Related Works

Because of the dramatic growth in attacks over the last three decades, there has been tremendous advancement in the field of intrusion detection systems. As a result, the researchers employ sophisticated algorithms and approaches to combat the network's most advanced attacks. Researchers develop their own datasets for their research studies. However, as the study grew in scope, so the number of datasets that were available on the internet were increased. Many modern scholars make extensive use of these databases in their studies. NSL-KDD and NSL KDD 99 are the most commonly utilized. Machine Learning algorithms are widely employed in cyber security to discover anomalies in the

network with high recognition rates, using those kind of datasets [5].

The researchers in, performed their study using the correlation-based feature selection in the feature engineering process and carry out their process with the four machine learning algorithms such as Bayes Net, Random Forest and Random Tree. However, in [6], T. Garg, S.S Khurana performs the compatible classification by utilizing NSL-KDD dataset with the machine learning algorithms. They used the Garret's Ranking Technique for the attribute selection and performance metrics were evaluated. Mrutyunjaya Panda Manas Ranjan Patra performed using three algorithms (ID3, J48, and Naïve Bayes) [7]. Its performance was based solely on classification accuracy of a very limited set of given instances. Examples of such methods include cross-validation tests, confidence tests. They use a cross-validation test for performance evaluation.

In [8], Pavel Nevlud, Miroslav Bures, Lukas Kapicak, Jaroslav Zdralek have presented detection of network anomalies by using machine learning systems. They tested some attacks in regular network traffic. As the first attack was used, ping sweeps to sub network targets to get information about active IP addresses. The port sweep was used as a second attack to scan open ports at the target victim computer.

In [9], M Panda, A Abraham, M Patra proposed an intelligent hybrid model using a combination of different classifiers to detect the attack on NSL KDD. The algorithms used are Support Vector Machine (SVM), Principal Component Analysis (PCA), and Random Forest and decision trees. He performed 10-fold cross validation on 25192 training instances. The result yielded Random Forest, when combined with other classifiers has a precision rate of 99.9%.

In [10], Sumouli Choudhury and Anirban Bhowal had put forth improved machine learning algorithms necessary for proper detection of network intrusion. They had also compared the performance of various classifiers in WEKA and concluded that Random Forest and Bayes Net are suitable for this purpose. The machine learning algorithms have also been compared and they performed machine learning algorithms can be deduced that Boosting is the best algorithm.

In [11], Deepthi Hassan Lakshminarayana reviewed over 56 papers on intrusion detection techniques from 2009 to 2019. He found that some of the detection techniques, such as machine learning, deep learning, and blockchain technology, play a vital role in constructing these lifesaving systems. In this work, He has considered three classifiers, namely Random Forest, Naïve Bayes, and decision trees.

In [12], the authors briefly review the WEKA workbench and the history of project, discuss new features in the recent 3.6 stable release, and highlight some of the many projects based on WEKA

3 METHODOLOGY

This section explains the experimental method carried out through the research study.

3.1 Dataset Selection

For the experiment we have selected a dataset known as NSL-KDD dataset [5]. This is a popular dataset for intrusion detection among researchers. NSL-KDD data set is the refined version of the KDDcup99 dataset. NSL-KDD dataset is a labeled dataset so we can experiment with the dataset by using supervised learning algorithms. Attacks they specify in the dataset can be Denial of Service Attack, User to Root Attack, and Remote to Local Attack, Probing Attack [13]. NSL-KDD dataset is a standard dataset used for the research on intrusion detection systems with 41 features.

3.2 Data Preprocessing and Feature Selection

Preprocessing the data is a very important step in preparing the data to be fed into the algorithm. The original dataset has 42 fields including 41 features and the class label, and some fields are categorical fields and others are numerical values. It is not possible to use these fields directly as the input to machine learning algorithms. Following steps are followed in the Preprocessing phase.

3.2.1 Discretization

Some feature values are of continuous data; therefore, the discretization process is followed to convert the continuous data into a finite set of intervals with minimum data loss.

3.2.2 Feature selection

Feature selection method used to reduce the dimension of the dataset and improve the accuracy of detection. Feature selection methods are mainly classified into two types, Wrapper and Filter methods. The study was done by utilizing Infogain feature selection methods. This selection method was used to rank attributes by their individual evaluations and a 10-fold cross validation method was used on dataset for testing and validation. Therefore, the dataset was tested for the performance evaluation with two datasets: • R- the Raw Dataset-with all the 41 features • ID- the Dataset after the feature selection process (Infogain).

3.3 Experiment with Classification Algorithms

As the next step we have identified four classification algorithms that are mostly used in intrusion detection. Classification algorithms in data mining are used in Intrusion Detection Systems to classify attacks or intrusions from ordinary things that happen in systems. It was decided to use the following classification algorithms for the evaluation [11], [9]. • Sequential Minimal Optimization (SMO) • Naïve Bayes • J48 • Random Forest

3.4 Experiment with Classification Algorithms

The performance of each classification algorithm was evaluated. The intrusion detection highly matters with the detection time. Therefore, the classification model was selected by considering the detection time as well as accuracy metric.

4 RESULTS AND DISCUSSIONS

It was noticed as the number of features were reduced as one third, i.e., 12 features have been selected after the feature selection process of Infogain feature selection method. The Figure below shows the list of features that were selected by the Infogain method. The Classification models were trained with the two types of datasets (R and ID datasets). We have considered algorithms such as the Random Forest Classifier, Naïve Bayes classifier, J48 classifier and Sequential Minimal Optimization classifier. Once the algorithms are applied on both datasets, the results were evaluated. The accuracy of the model is calculated with the correctly identified instances and the incorrectly identified instances. However, from the accuracy of the classification algorithms are more than 97%, it was found that Random Forest outperforms the other classification, when the whole dataset is used for evaluation.

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 0.816 +- 0.001      1   +- 0        5 src_bytes
 0.672 +- 0.001      2   +- 0        3 service
 0.632 +- 0.001      3   +- 0        6 dst_bytes
 0.519 +- 0.001      4   +- 0        4 flag
 0.518 +- 0.001      5   +- 0       30 diff_srv_rate
 0.51  +- 0.001      6   +- 0       29 same_srv_rate
 0.475 +- 0.001      7   +- 0       33 dst_host_srv_count
 0.437 +- 0.001      8   +- 0       34 dst_host_same_srv_rate
 0.41  +- 0.001      9   +- 0       35 dst_host_diff_srv_rate
 0.405 +- 0.001     10.4 +- 0.49    38 dst_host_serror_rate
 0.405 +- 0.001     10.6 +- 0.49    12 logged_in
 0.398 +- 0.001     12   +- 0       39 dst_host_srv_serror_rate
```

Figure 1. The selected attributes with Infogain feature selection method

The evaluation is checked by reducing the dimension of the dataset, by utilizing the Infogain feature selection method, i.e., with 12 features dataset, it was identified as the accuracy changes with the reduction of the complexity of the dataset.

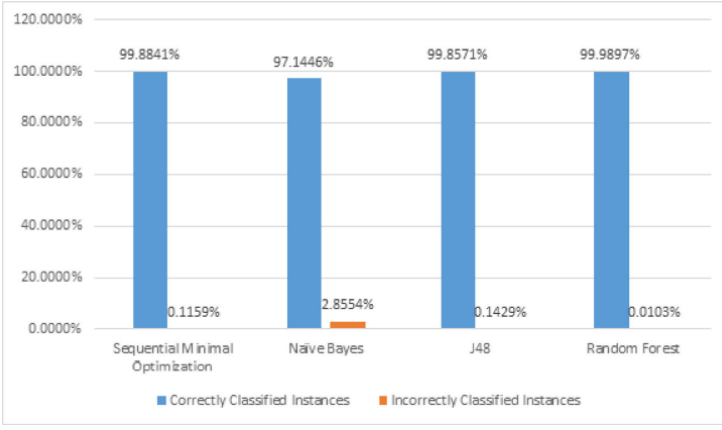

Figure 2. The accuracy score for the classification algorithms before pre-processing

The testing time is an important factor to consider in deciding the best model for detection of intrusions. This would help in identifying the model which can detect with minimal time. The Figure shows how the evaluation of the four classification algorithms vary their testing time with the two datasets (R and ID). Thereby it was clearly seen that the testing time reduces vastly with the feature reduction. Though the Random Forest outperforms in the detection with high constant accuracy of 99.977%, it shows the maximum testing time during the evaluation process. By assessing the testing time and accuracy, it was vividly highlighted that J48 has its minimum testing time for ID dataset as 0.05s, with the average accuracy of 99.846%. Therefore, by considering both the accuracy score and the testing time, J48 was chosen as the best model in detecting the intrusion.

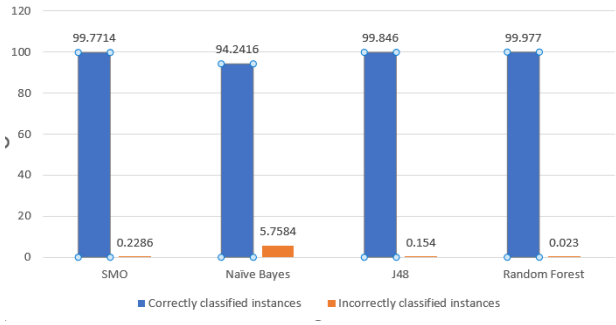

Figure 3. The accuracy score for the classification algorithms after pre-processing

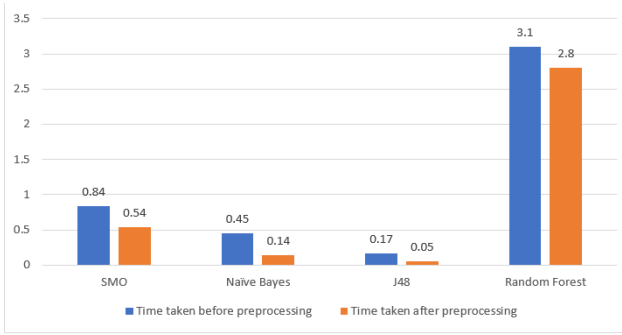

Figure 4. The time comparison between before and after pre-processing for the classification algorithms

5 Conclusion

Due to the enormous growth in cyber-attacks, there is a requirement of an effective intrusion detection system to protect the data and the network. This study is an effort toward identifying accurate machine learning algorithms for intrusion detection. We have used the NSL-KDD dataset for this research study. The testing time was found to be high due to its complexity of the dataset. Thereby, the Infogain feature selection algorithms was used to facilitate the selection of the feature with the highest rank. Classification algorithms Sequential Minimal Optimization (SMO), Naïve Bayes, J48, and Random Forest were used with the selected features and the whole dataset. The accuracy and the testing time vary between the algorithms. Finally, we have identified there are significant changes occurring when performing the classification with selected features. Though the Random Forest scores the best accuracy with 99.9%, J48 scores with 99.8% with less response time. The J48 was chosen as the best machine learning algorithm with high accuracy level of detection and minimum response time.

Acknowledgment

We would like to express my special thanks to my our supervisor, Ms. Rukshani Puvanendran, Lecturer, Department of ICT, University of Vavuniya, who gave us the golden opportunity to do this wonderful project on the topic “An Analysis on NSL-KDD Dataset using Machine Learning Techniques for Intrusion Detection”, which also helped us in doing a lot of research and we came to know

about so many new things. We were really thankful to them. Secondly, We would also like to thank my friends who helped us a lot in finishing this project within the limited time. It helped us increasing our knowledge and skills.

REFERENCES

[1] *Report, Cisco Annual Internet*. 2018-2023.

[2] Gina C Tjhai, Steven M Furnell, Maria Papadaki, et al. "A preliminary two-stage alarm correlation and filtering system using SOM neural network and K-means algorithm". In: *Computers & Security* 29.6 (2010), pp. 712–723.

[3] Ayman I Madbouly, Amr M Gody, and Tamer M Barakat. "Relevant feature selection model using data mining for intrusion detection system". In: *arXiv preprint arXiv:1403.7726* (2014).

[4] Wenke Lee, Salvatore J Stolfo, and Kui W Mok. "A data mining framework for building intrusion detection models". In: *Proceedings of the 1999 IEEE Symposium on Security and Privacy (Cat. No. 99CB36344)*. IEEE. 1999, pp. 120–132.

[5] *UNB, NSL-KDD Datasets Research Canadian Institute for Cybersecurity*. 2022.

[6] Tanya Garg and Surinder Singh Khurana. "Comparison of classification techniques for intrusion detection dataset using WEKA". In: *International conference on recent advances and innovations in engineering (ICRAIE-2014)*. IEEE. 2014, pp. 1–5.

[7] Mrutyunjaya Panda, Ajith Abraham, and Manas Ranjan Patra. "A hybrid intelligent approach for network intrusion detection". In: *Procedia Engineering* 30 (2012), pp. 1–9.

[8] Pavel Nevlud, Miroslav Bures, Lukas Kapicak, et al. "Anomaly-based network intrusion detection methods". In: *Advances in Electrical and Electronic Engineering* 11.6 (2013), pp. 468–474.

[9] Mrutyunjaya Panda and Manas Ranjan Patra. "A comparative study of data mining algorithms for network intrusion detection". In: *2008 First International Conference on Emerging Trends in Engineering and Technology*. IEEE. 2008, pp. 504–507.

[10] Sumouli Choudhury and Anirban Bhowal. "Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection". In: *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*. IEEE. 2015, pp. 89–95.

[11] Deepthi Hassan Lakshminarayana. *Intrusion detection using machine learning algorithms*. East Carolina University, 2019.

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, et al. "The WEKA data mining software: an update". In: *ACM SIGKDD explorations newsletter* 11.1 (2009), pp. 10–18.

[13] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, et al. "A detailed analysis of the KDD CUP 99 data set". In: *2009 IEEE symposium on computational intelligence for security and defense applications*. Ieee. 2009, pp. 1–6.