Network Intrusion Detection Approach using Machine Learning Based on Decision Tree Algorithm

Elmadena M. Hassan*
Computer Science Department, FMCS, University of Gazira, Wad madani, Sudan, madinamohammed@uofg.edu.sd
Mohammed A. Saleh
Computer Department, College of Science and Arts in Ar Rass Qassim University, Kingdom of Saudi Arabia, m.saleh@qu.edu.sa
Awadallah M. Ahmed
Computer Science Department, FMCS, University of Gazira , Wad madani, Sudan, awadallahd@uofg.edu.sd

Abstract

Computer security, as well cyber security, is safeguarding information systems from stealing, destruction, and misusing computer hardware, software, data, and the delivered services. In general, machine learning is the area of studying, which grants a computer system to grasp, although not explicitly programmed. Often, anomaly-based Intrusion Detection Systems (IDS) experiences high false alarms rates (FAR), and since many different mechanisms are used by the researchers to protect the system from high false alarms and least detection rates, the challenge is to reduce high false alarms and achieve high detection rate is remain, and therefore; a new approach need to be applied. The objective of this study is to specify a network traffic technique to distinguish the normal from abnormal attacks, and also, to use specific algorithm to reduce the high false alarms rate (FAR). The dataset used in this study is NSL-KDD, where the data are divided into two parts (60%) for training and (40%) for testing. The results show that the decision tree (DT) algorithm achieved high detection rate (DR) and low false alarms rate (FAR) in comparison with other machine learning algorithms. This study achieved rate of detection for random tree about (99.7%) and for J48 about (99.8%), but for naïve Bayes about (86.8%). Also, the rate of false alarm for random tree about (0.2%) and for J48 about (0.3%), but for naïve Bayes about (6%), and hence the researchers concluded that the decision tree algorithm accomplishes high detection rate (DR), and low false alarms rate (FAR) compared to other algorithms of machine learning.

Keywords:
Decision Tree Algorithm; NSL-KDD dateset; Anomaly Detection

1. Introduction

As the size and class of electronic network attacks increase, it becomes tough to discover intrusion into the network of interests. Advanced persistent threats (APTs) became a greater threat to all companies and nation states. Cybercriminals are still using more refined technologies to illicitly access systems, and corporations and

staff are increasingly to adopting new, more refined technologies and networks in the workplace. All these elements considerably hinder the mission of defending the network and trends suggest that these problems are possible to be extended over time. Fighting these difficulties need totally different tools and strategies to discover and defend attacks.

Decision tree is a method from the sphere of data mining, can help in this mission. Decision trees give distinctive views into the matter of recognizing malignant activity and may help to create technology-specific techniques to prevent attacks [1].

Computer security, as well as cyber or IT security, is guarding information systems from stealing or destructing to the software, hardware, or information on them. In addition, it protecting from perturbation or misusing of the IT services they deliver. The type of attacks, such as buffer overflow, teardrop and ping of death [2].

The Intrusion detection system (IDS) is defined as a system for discovering intrusions which making an attempt to misuse the data or computing resources of a computer system. It has two detection models include signature based IDS and anomaly based systems [3].

The target of this research is to decrease the high rate of false alarms, and also to achieve high detection rate. With the aim of achieving this goal, the next points are set:

• Use specific algorithm to reduce the high false alarms.

• Specify an efficient technique to distinguish the normal from abnormal attacks.

The reset of this paper is organized as follows: section 2 represents the previous related works, section 3 denotes the proposed approach, an explanation for the dataset, and the detection metrics that will be used, section 4 presents the experimental setup, section 5 shows the results and discussions, and finally section 6 is the conclusion.

2. Previous Related Works

In [4] a study applied decision tree to explain how the decision tree algorithms are used to reveal attacks in result of students. It aims to make models for decision tree (DT), in order to detect abnormal circumstances mechanically in student results tests. The results of this research achieved reasonable performance in the evaluation, namely in accuracy, sensitivity and specificity.

In [5] a study anticipated novel machine learning algorithms for declining false positives alarms in intrusion detection system (IDS). This work aims to state necessary entering attributes for structuring consort intrusion detection system (IDS), which is cost-effective. The results of this research show that the intended approach reduces the quantity and ratio of false positives, and poise detection rates (DR) for several types of network intrusions.

In [6] a study developed intrusion detection system (IDS) to detect attacks in computer networks. This work aims to analyze KDD99 test dataset by utilizing

particular algorithms of machine learning like J48, random tree, Bayes net, random forest to identify these algorithms precision throughout categorizing attacks to wide-ranging classes. The results of this work explained that the random tree and random forest algorithms are the utmost effective in acting the classification on test dataset. As well, it depicted that CFS method for decreasing the detection time and increases the accuracy rate.

In [7] a study introduced the involvement of the four attributes categories in terms of evaluating detection rate (DR) and false alarm rates (FAR) metrics. This research studied NSL-KDD dataset in perspective of four attributes classes rather than behavior of specific attribute. The research showed that it can assist dataset appropriateness, in which greater detection rate (DR) is gained with a lowest false alarm rates (FAR). The shortcoming of this study it does not suitable for online intrusion detection.

In [8] a new study uses different classification algorithms to discover the anomalies in the network traffic patterns. This work aims to utilize and analyze the NSL-KDD dataset to study the capacity of various classification algorithms in realizing the abnormalities in patterns of network traffic.

In [9] study explained effective approach for detection and classification of lung cancer-related CT scan images into benign and malicious category. This work aims to apply image processing and machine learning approaches for detection and classification of lung cancer. The results of this work found that accuracy of MLP classifier is higher with value of 88.55% in comparing with the other classifiers.

In [10] study provided confidentiality to the user while utilizing cloud-based web services. This work aims to suggest DNA-based encryption/decryption technique to embed the confidentiality in the communication between client and the cloud service provider. The results of this work observed that DNA-based encryption/decryption technique gives the option to selectively encrypt only the confidential information rather that the whole information as in the case of SSL.

3. Proposed Approach based on Decision Tree Algorithm

The proposed approach based on Decision Tree algorithm follows the next pseudo code:

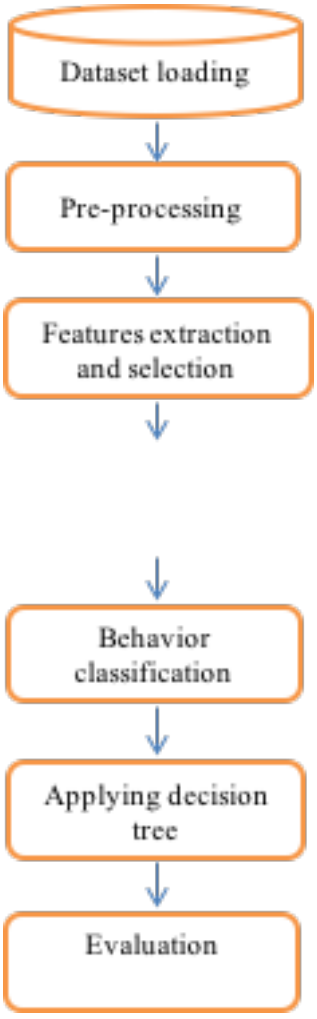
Step 1: Load the dataset into Weka.
Step 2: Apply pre-processing on the dataset.
Step 3: Divide the dataset into two labels called basic and traffic respectively.
Step 4: Calculate detection metrics (DR, FAR)
Step 5: Classify the behavior if it is normal or anomaly.
Step 6: Apply the decision tree algorithm.
Step 7: Evaluate detection metrics.

As well, it involves the following subsequent phases, as shown in Fig.1:

1. Dataset loading
2. Pre-processing

3. Features extraction and selection
4. Calculating detection metrics
5. Classification of behavior
6. Applying decision tree
7. Evaluation

Fig.1. The Proposed Approach based on Decision Tree Algorithm



The researchers use the decision tree algorithm for attack detection and reduce false alarms rate and to apply decision tree algorithm they following many steps include dataset loading in this phase we load the dataset in Weka workbench and the dataset will utilized in this paper is NSL-KDD dataset, pre-processing this phase is no need for pre-processing to be applied on the data used in this paper, in the feature extraction and selection phase they divide the both KDDTest+.arff and the KDDTrain+.arff into two labels called basic and traffic respectively. In calculating detection metrics phase, it calculates accuracy, Detection Rate (DR), False Alarm Rate (FAR), precision, and F-score. Indeed, these metrics are attained from the four principal components of a classification algorithm results that are presented in confusion matrix form, which demonstrates the actual instance classes versus predicted classes[11]. As presented in the table 1 below, these components are true negative (TN) component, false negative (FN) component, false positive (FP) component, and true positive (TP) component. In the classification of behavior phase, the researchers classify the behavior if it is normal or anomaly, and to perform this phase they apply the decision tree algorithms (J48, Random tree). In the reducing false alarms phase, the researchers apply the decision tree algorithm for reduce the false alarms rate and compare it with naïve Bayes to observe the effect of algorithms in reducing false alarms rate. The last phase is evaluating detection metrics, and in this phase the researchers evaluate the system depend on the rate of false alarms and detection for the decision tree.

To evaluate the gained results, the researchers need to calculate the accuracy, DR, FAR, precision and F-score. All these metrics are derived from the four basic result elements of any classification algo-

rithm presented in the form of confusion matrix, which illustrates as in table 1.

Table 1. Confusion matrix for IDS

Confusion Matrix		Predicted Instances	
		Normal	Anomalous
Actual Instances	Normal	TN	FP
	Anomalous	FN	TP

2.1 NSL – KDD Dataset

NSL-KDD is a dataset offered to settle several of the ingrained matters of KDD99 dataset. While this version of KDD dataset experiences amount of the problems, mentioned by McHugh, that is not perfect illustrative of surviving actual networks due to the lack of universal datasets for network IDS, it widely used proficiently as benchmark dataset. Furthermore, the size of records of the NSL-KDD for train dataset and test dataset is suitable. Therefore, it is probable to apply tests for the whole dataset. The existing attack types of NSL-KDD dataset are: DOS, probing, U2R and R2L [12]. The NSL-KDD dataset holds the succeeding benefits compared to authentic KDD dataset:

- There are no repeated records in the train dataset, and hence; the classifier will not be prejudiced to many of repeated records.
- There is not any repetitive record in the planned test sets; so, learners' performance is not aligned with strategies that have higher detection rates in repeated records.
- The number of chosen records from every problem level collection is reversely related to the amount of records in the original KDD dataset.
- The numbers of records in the train dataset and test dataset are satisfactory, which keeps it reasonable to perform the experiences over the whole records, not to arbitrarily pick slight share.

NSL-KDD training dataset involves 4,900,000 single association vectors, each has 41 features, and classified as; either normal, or an attack. Table 2 shows the class of all 42 attributes of NSL-KDD dataset.

Table 2. Classes of KDD dataset attributes

No.	Label	Attribute Name	No.	Label	Attribute Name
1	B	duration	22	T	is_gust_login
2	B	protocol_type	23	T	Count
3	B	services	24	T	srv_count
4	B	flag	25	T	serror_rate
5	B	src_bytes	26	T	srv_serror_rate
6	B	dst_bytes	27	T	rerror_rate
7	B	land	28	T	srv_rerror_rate
8	B	wrong_fragmentt	29	T	same_srv_rate
9	B	urgent	30	T	diff_srv_rate
10	B	hot	31	T	srv_diff_host_rate

No.	Label	Attribute Name	No.	Label	Attribute Name
11	B	Num_faile_logins	32	T	dst_host_count
12	B	logged_in	33	T	dst_host_srv_count
13	B	num_compromised	34	T	dst_host_same_srv_rate
14	B	root_shell	35	T	dst_host_diff_srv_rate
15	B	su_attempted	36	T	dst_host_same_src_port_rate
16	B	num_root	37	T	dst_host_srv_diff_host_rate
17	B	num_file_creations	38	T	dst_host_serror_rate
18	B	num_shells	39	T	dst_host_srv_serror_rate
19	B	num_access_files	40	T	dst_host_rerror_rate
20	B	num_outbound_cmds	41	T	dst_host_srv_rerror_rate
21	B	is_hot_login	42	-	Class

2.2 Detection Metrics

This research utilized accuracy, detection rate (DR), precision, false alarms rate (FAR), F-measure, and recall as main detection metrics, which are based on calculating true positive (TP) rate value, false positive (FP) rate value, true negative (TN) rate value, and false negative (FN) rate value. True positive (TP) rate value denotes instance that is truly an attack, and classified correctly as an attack. FP represents those instances which are actually normal but classified as an attack. False positive (FP) rate value expresses instance that is in reality an attack, but classified wrongly as a normal. Where, true negative (TN) rate value symbolizes instance that is a normal, and classified as a normal, as well [13]. Accuracy represents how many instances were correctly classified, while precision represents out of all the instances classified as attacks, and how many were actually an attack. Recall represents how many attacks were correctly classified, percentage of attacks caught [14].

False alarm rate (FAR) is defined as the rate at which normal instances are wrongly classified as abnormally. Detection rate (DR) defines the proportion of accurately forecasted attacks to the whole number of real attacks. F-measure is defined as the coordinated mean of detection rate (DR) and the precision value.

4. Experimental Setup

This paper purposes to study and explain the function of 41 attributes of NSL-KDD dataset, and to utilize decision tree (DT) algorithm on detection rate (DR) and false

alarm rate (FAR) for Intrusion Detection System (IDS) with relevance to two attributes class labels as in table 3.

Table 2. Classes of KDD dataset attributes

Attribute class	Abbreviation	Attributes
Basic	B	1 – 21
Traffic	T	22 – 41

The machine used in this research is HP laptop with 32 bit windows 7
operating system, corei3 and 4GB RAM. Furthermore, Weka 3.6.9 environment to implement the algorithms.

The NSL-KDD dataset was chosen for this experimental study whose attributes are tagged in two categories as present in table 3, where the data are divided into two parts (60%) for training and (40%) for testing. And this dataset we can find in (http://github.com/defcom17/NSL-KDD). This dataset holds a number of preparations, thus that its records is categorized into binary categories, such as normal/ abnormal, or in one of five categories like normal, user to root (U2R) attack, denial of service (DOS) attack, probe attack, and , remote to local (R2L) attack. This study deals with the binary classification dataset detailed in Table 4.

Table 4. Instances of NSL-KDD dataset

	Normal Class Instances	Anomalous Class Instances	Total
KDDTrain+	67343	58630	125973
KDDTest+	9711	12833	22544

5. Results and Discussions

The results are presented as of confusion matrix of the three arrangements of datasets in Table 5 for naïve Bayes, Table 6 for random tree and Table 7 for J48.

Table 5. Result set for naïve Bayes algorithm

No	Attribute class combination	Naïve Bayes			
		TN	FN	FP	TP
1	BT	25362	3090	1626	20311
2	B	26675	9544	313	13857
3	T	24286	3267	2702	20134

Table 6. Result set for Random tree algorithm

No.	Attribute class combination	Random tree			
		TN	FN	FP	TP
1	BT	26931	67	57	23334
2	B	26766	119	222	23282
3	T	26735	299	253	23102

Table 7. Result set for j48 algorithm

No	Attribute class combination	J48			
		TN	FN	FP	TP
1	BT	26895	50	93	23351
2	B	26757	124	231	23277
3	T	26728	202	260	23199

The outline of results for DR is presented in Table 8 and for FAR is presented in Table 9. The essential measures applied in this paper are detection rate (DR) and false alarm rate (FAR). The classification results in detection rate (DR) and false alarm rate (FAR) form for the three cases of attribute classes mixtures are shown for Random tree, J48 and Naïve Bayes classifiers.

Table 8. Detection rate (DR) for random tree, j48 and naïve Bayes algorithms

No.	Attribute class combination	Detection Rate (%)		
		Random tree	Naïve Bayes	J48
1	BT	99.7	86.8	99.8

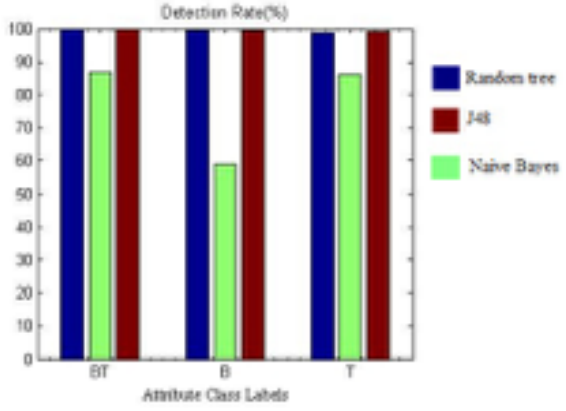
No.	Attribute class combination	Detection Rate (%)		
		Random tree	Naïve Bayes	J48
2	B	99.5	59.2	99.5
3	T	98.7	86	99.1

Table 9. False alarms rate (FAR) for random tree, j48 and naïve Bayes algorithms

No.	Attribute class combination	False Alarms Rate (%)		
		Random tree	Naïve Bayes	J48
1	BT	0.2	6	0.3
2	B	0.8	1.2	0.9
3	T	0 .9	10	1

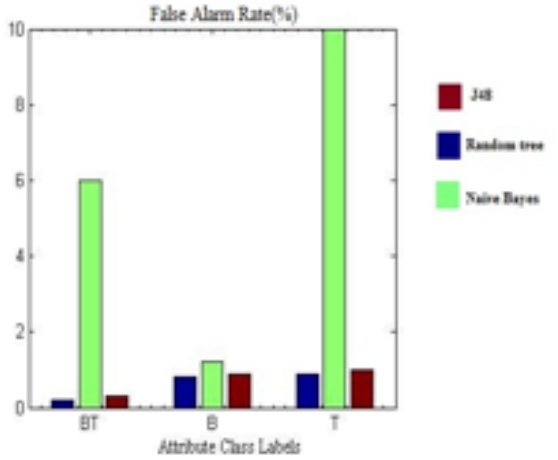
Fig. 2 and Fig. 3 present plots for the three classification algorithms with respect to one and two labeled attribute combinations respectively.

Fig. 2. Detection rate (DR) distributions for Random Tree, J48 and Naïve Bayes Algorithms



Taken into account the analysis of DR, Fig. 2 plot depict the DR for basic, traffic and BT attribute label combination. The random tree and J48 represent high detection rate for basic, traffic and BT labels but the naïve Bayes has low detection rate for three labels. Hence it can be concluded from Fig. 2 that the random tree and J48 have important contribution to achieving high detection rate (DR), whereas the naïve Bayes achieving low detection rate (DR).

Fig. 3. False alarms rate (FAR) for Random Tree, J48 and Naïve Bayes Algorithms



Taken into account the analysis of FAR, Fig. 3 plot depicts the FAR for basic, traffic and BT attribute labels combination. The random tree and J48 represents better FAR for basic, traffic and BT labels than Naïve Bayes. Hence it can be concluded from Fig. 3 that the random tree and J48 have significant contribution towards the reduction of the FAR whereas the naïve Bayes increase the FAR.

6.Conclusion

The fast advancement of data mining algorithms and strategies has led to machine learning shaping a distinct field of technology. It can be seen as a subclass of the artificial intelligence field, where the most important plans are the ability of a system to learn from its own activities.

Decision tree (DT) be attached to a class or category of supervised learning algorithms. In contrary to disparate supervised learning algorithms, decision tree (DT)

algorithm is applied to resolving classification and regression problems, too. NSL-KDD dataset is one of the commonly exercised datasets for performance testing of Intrusion Detection System (IDS). In this paper, researchers used decision tree (DT) algorithm to achieve high detection rate (DR) and low false alarms rate (FAR). They applied decision tree algorithm and naïve Bayes on Weka. The decision tree achieved better FAR for basic, traffic and BT labels than Naïve Bayes. Also, the decision tree gained high detection rate for basic, traffic and BT labels but the naïve Bayes has low detection rate. Hence, they conclude with that the decision tree (DT) algorithm achieved high detection rate (DR) and low false alarms rate (FAR), which is better than other algorithms of machine learning.

In the future work researchers recommend applying other machine learning algorithm for five classes (normal, U2R, R2U, DOS and probe) to reduce the false alarms rate.

References

[1] Jeff Markey and Antonios Atlasis, "Using decision tree analysis for intrusion detection," SANS Institute InfoSec Reading Room, 2011.

[2] Christina Mei-Fang Lee, An evaluation of machine learning techniques in intrusion detection., 2007.

[3] Ja Jabez and B Muthukumar, "Intrusion detection system (IDS): anomaly detection using outlier detection approach," Procedia Computer Science, vol. 48, pp. 338--346, 2015.

[4] Hamza O Salami, Ruqayyah S Ibrahim, and Mohammed O Yahaya, "Detecting Anomalies in Students' Results Using Decision Trees," International Journal of Modern Education and Computer Science, vol. 8, no. 7, p. 31, 2016.

[5] Dewan Md Farid and Mohammad Zahidur Rahman, "Attribute weighting with adaptive NBTree for reducing false positives in intrusion detection," arXiv preprint arXiv:1005.0919, 2010.

[6] Chibuzor John Ugochukwu and EO Bennett, "An Intrusion Detection System Using Machine Learning Algorithm," International Journal of Computer Science and Mathematical Theory, vol. 4, no. 1, pp. 2545--5699, 2018.

[7] Preeti Aggarwal and Sudhir Kumar Sharma, "Analysis of KDD dataset attributes-class wise for intrusion detection," Procedia Computer Science, vol. 57, pp. 842--851, 2015.

[8] L Dhanabal and SP Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 6, pp. 446--452, 2015.

[9] Gur Amrit Pal Singh and P. K. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans,", Neural Computing and Applications, 2018.

[10] Gunjan Gugnani S. P. Ghrera P. K. Gupta Reza Malekian B. T. J. Maharaj, "Implementing DNA Encryption Technique in Web Services to Embed Confidentiality in Cloud," Proceedings of the Second International Conference on Computer and Communication Technologies pp 407-415, 2015.

[11] Gupta, PK, Tyagi, Vipin, Singh, S.K., "Predictive Computing and Information Security," Springer Singapore, 2017.

[12] Rahul P. Tolankar, Vaibhav P. Sawalkar, and Niraj N. Kasliwal, "Review on IDS in Cloud Enviroment By Using FC - ANN," in National Conference on Innovative Trends in Science and Engineering, vol. 4, 2016, pp. 382 - 386.

[13] Wen Yu, Haibo He, and Nian Zhang, Advances in Neural Networks-ISNN 2009: 6th International Symposium on Neural Networks, ISNN 2009 Wuhan, China, May 26-29, 2009 Proceedings.: Springer, 2009, vol. 5552.

[14] Anubhavnidhi Abhashkumar and Roney Michael, "implementation an intrusion detection system using a decision tree".