International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Prediction of Denial of Service Attack using Machine Learning Algorithms

Pl.Yazhini, Visalatchi

Abstract: ***DDoS attack is one of the significant security threats in today's Internet world. The main intention of the network thread is to make the resource unavailable such as flooding attacks. Here, Machine learning algorithms have been used for detecting DDoS attacks. Generally, the success of any algorithm has depended on the selection of appropriate data sets and the identification of attack parameters. The KDD-CUP dataset has been taken for a detail investigation of the DDoS attack. The K-nearest neighbor, ID3, Naive Bayes and C4.5 algorithms are compared in a single platform concluding with the positives with Naive Bayes. The main objective of the paper is to compare and predict the error rate, computation time, Accuracy of the algorithms using the Tanagra tool. Finally, these correlative algorithms have been compared and verified through experimental verification and graphical representation.***

Keyword: DDoS attack, classification algorithm, C4.5, ID3, Naïve Bayes, K-Nearest Neighbors.

I. INTRODUCTION

Websites may popular either it can be accessed frequently by a great number of users or it contains any useful information. Hence, it can be accessed by an enormous number of users it may sometimes lead to an overload of server, that's may lead to an attack. The massive growth in technologies leads to many hacking incidents. The different kinds of threats are emerging every day and their intention is to make resources unavailable to the user. The DoS attack can be measured using the amount of traffic they send to the host system per-second, For example, small attacks might be measured by few megabits(Mpbs), while in large attacks it might be terabit Per second(Tbps). Attackers use the botnet for large volume attacks to perform the attacks effectively. If it happens regularly server can't respond to the legitimate user. The expert compares and monitors the incoming packet traffic with traffic signatures; by using this method network administrators protect internet devices from attack.

There are numerous amounts of specialized tools, for a complete analysis of the aggregated KDD-CUP dataset. Here, the Tanagra tool or software is used to analyze the dataset. It helps in analyzing the cluster, Visualization of data, Regression analysis, decision trees, predictive analytics, text mining, etc. Over here, the classification algorithm is used to make decisions in the data analysis process. The comparison is done to measure the accuracy rate of every classification algorithm.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
Pl.Yazhini, M.Phil, Department of Computer Science, Dr.Umayal Ramanathan College for Women, Karaikudi
Visalatchi, Associate Professor, Department of Information Technology, Dr.Umayal Ramanathan College for Women, Karaikudi.

II. RELATED WORKS

In this section, we review the work which is related to the prediction, detection, mitigation, and accuracy of denial of service attack on applying popular machine learning algorithms.

In [5], the paper proposed Denial of service Attack Prediction using the Gradient Descent algorithm yields the result of accuracy 99.7% with a 3.3% error. To overcome the error rate use the Linear Regression algorithm in which the error can be reduced to 0.3%.

In [1], discussed A Classification Framework to Detect DoS Attacks used NSL-KDD dataset. The performance is compared with the 10 commonly used classification technique Naive Bayes (NB), Support Vector Machine (SVM), Multilayer Perceptron (MLP), K-Nearest Neighbor (KNN), Decision Tree (DT), Radial Basis Function (RBF), One Rule (OneR), PART, Bayesian Network, and Random Tree.

In [4], Tahir Alyas discussed Detection and mitigation of DDoS attacks in cloud computing using a machine learning algorithm. The two most effective algorithms Naïve Bayes and Random forest, concluding with that Naïve Bayes produce more positives.

In [7], focused on A study for the DDOS attack classification method. Here, three receptive classification algorithms can be compared to Naive Bayes, Decision tree, Artificial neural network from that Artificial neural network produces the best accuracy rate of 84.3% compared with the other two techniques. In [6], the paper proposed Research on multiple Machine learning for anomaly detection. In this approach NSL-KDD dataset is divided into two dataset its performance can be compared with the confession matrix.

III. CLASSIFICATION ALGORITHMS USED IN THE EXPERIMENT

The classification algorithm usually divides the data in the form of a subset of class. The main goal of the classification algorithm is to find the class the new data will fall. However, Classification can be performed on structured or unstructured data. The classification problems always have discrete value as its output.

A. K-Nearest Neighbors:

KNN algorithm or lazy learning algorithm is a supervised machine learning algorithm used to solve both classification and regression problems. In the KNN algorithm,

there is no need to train a model and that is an instance-based training model. Working procedure of KNN,

1. Load the data.
2. Initialize k to neighbors.
3. Calculate the distance.
4. Distance is calculated and adds them to an index.
5. Distance and indices are sorted from smallest to largest.
6. Pick the first K entries in the sorted list.
7. In a classification problem, return the mode in K labels.

B. Naive Bayes:

Naïve Bayes is particularly built for large datasets. Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Bayes theorem is mainly for calculating posterior probability.

Working procedure for Bayes theorem:

1. Dataset is converted into frequent tables.
2. Create a probabilistic table by the occurrence of data and finding the probability.
3. Use Bayes posterior probability equation to find the highest posterior probability.

C. ID3 algorithm:

ID3 (Iterative Dichotomiser 3), it generates the decision tree developed by Ross Quinlan. It divides the attribute into two groups one is the most dominant attribute and others to build a tree. After then it calculates entropy and Information gain for the attribute. Finally, the most dominant one is put as a root node in the decision tree. It continues until it made a decision for a branch. That's why it is called Iterative Dichotomiser.

D. C4.5 algorithm:

ID3 evolved version is C4.5 is used to generate the decision tree and it is developed by Ross Quinlan. It builds a decision tree from a set of sample data (training data) using the concept of information entropy. For each node of the tree attributes the training data is used. It works well on continuous and discrete data.

IV. TOOL DESCRIPTION

The data to be processed with Machine Learning Algorithms are ever-increasing in size. Mainly when need to process unstructured data. Tanagra is a software or tool for research purposes developed by Ricco Rakotomalala. It inherits several standard data mining algorithms such as Data visualization, statistics, Instance selection, Feature selection, Regression, clustering, Association, etc.

It is an "open source project" even researchers can add their own algorithm with software distribution license. Commonly other tools work in the workflow paradigm and its treatment represents in Tree diagram. Results are displayed in HTML format that makes it easy to export results in spreadsheet format. The primary aim of the Tanagra project is to provide a user-friendly environment for the researcher, student, and professional.

V. EXPERIMENT

The data used for the study is secondary data that was collected by third parties. Another alternative way to collect real traffic data of DDoS attacks by using simulation. Usually, simulation can be done between the peer-to-peer, one as a server system and another as the attacker. Here KDD-CUP dataset is used as a sample data (some of the attributes are taken for experimentation) for the detail investigation of DDoS attack. The dataset contains attributes such as duration, Protocol, services, flag, source_byte, and Destination_byte.

There are several correlative machine learning algorithms such as SVM, Decision tree, Artificial neural network, Genetic algorithm, K-means, Cluster analysis, ID3, C4.5, Naïve Bayes. Now, the K-nearest neighbor, ID3, Naïve Bayes and C4.5 algorithms are compared to measure the performance of each algorithm. The performance of the algorithm is estimated using the confusion matrix. The confusion matrix for the KNN, Naïve Bayes, and ID3.

A. Naïve Bayes:

Table3 contains the information of the Naïve Bayes algorithm displayed in the confusion matrix, computed on the learning sample (Classifier performance). The error rate is 0.0048 with computation time 78ms.

B. ID3:

Table4 contains the information of the ID3 algorithm displayed in the confusion matrix, computed on the learning sample (Classifier performance). The error rate is 0.0016 with computation time 234ms.

Table1: experimental result for K-NN

Error rate	0.1740
Values prediction	Confusion matrix

Value	Recall	1-Precision
normal	1.0000	0.1740
buffer_overflow	0.0000	1.0000
loadmodule	0.0000	1.0000
perl	0.0000	1.0000
neptune	0.0000	1.0000
smurf	0.0000	1.0000
guess_passwd	0.0000	1.0000
pod	0.0000	1.0000
teardrop	0.0000	1.0000

	normal	buffer_overflow	loadmodule	perl	neptune	smurf	guess_passwd	pod	teardrop	Sum
normal	18145	0	0	0	0	0	0	0	0	18145
buffer_overflow	2	0	0	0	0	0	0	0	0	2
loadmodule	1	0	0	0	0	0	0	0	0	1
perl	1	0	0	0	0	0	0	0	0	1
neptune	2	0	0	0	0	0	0	0	0	2
smurf	3695	0	0	0	0	0	0	0	0	3695
guess_passwd	1	0	0	0	0	0	0	0	0	1
pod	20	0	0	0	0	0	0	0	0	20
teardrop	99	0	0	0	0	0	0	0	0	99
Sum	21966	0	0	0	0	0	0	0	0	21966

Computation time : 109 ms.
Created at 1/25/2020 12:36:03 PM

Table2: experimental result for C4.5

Error rate	0.0004
Values prediction	Confusion matrix

Value	Recall	1-Precision
normal	1.0000	0.0005
buffer_overflow	0.0000	1.0000
loadmodule	0.0000	1.0000
perl	0.0000	1.0000
neptune	0.0000	1.0000
smurf	0.9995	0.0000
guess_passwd	0.0000	1.0000
pod	1.0000	0.0000
teardrop	1.0000	0.0000

	normal	buffer_overflow	loadmodule	perl	neptune	smurf	guess_passwd	pod	teardrop	Sum
normal	18145	0	0	0	0	0	0	0	0	18145
buffer_overflow	2	0	0	0	0	0	0	0	0	2
loadmodule	1	0	0	0	0	0	0	0	0	1
perl	1	0	0	0	0	0	0	0	0	1
neptune	2	0	0	0	0	0	0	0	0	2
smurf	2	0	0	0	0	3693	0	0	0	3695
guess_passwd	1	0	0	0	0	0	0	0	0	1
pod	0	0	0	0	0	0	0	20	0	20
teardrop	0	0	0	0	0	0	0	0	99	99
Sum	18154	0	0	0	0	3693	0	20	99	21966

Computation time : 468 ms.
Created at 1/25/2020 12:35:07 PM

Table3: experimental result for Naïve Bayes

Error rate	0.0048

Values prediction

Value	Recall	1-Precision
normal	0.9955	0.0003
buffer_overflow	0.0000	1.0000
loadmodule	0.0000	1.0000
perl	0.0000	1.0000
neptune	1.0000	0.0000
smurf	1.0000	0.0227
guess_passwd	0.0000	1.0000
pod	0.0000	1.0000
teardrop	1.0000	0.0000

Confusion matrix

	normal	buffer_overflow	loadmodule	perl	neptune	smurf	guess_passwd	pod	teardrop	Sum
normal	18064	15	0	0	0	66	0	0	0	18145
buffer_overflow	2	0	0	0	0	0	0	0	0	2
loadmodule	1	0	0	0	0	0	0	0	0	1
perl	1	0	0	0	0	0	0	0	0	1
neptune	0	0	0	0	2	0	0	0	0	2
smurf	0	0	0	0	0	3695	0	0	0	3695
guess_passwd	1	0	0	0	0	0	0	0	0	1
pod	0	0	0	0	0	20	0	0	0	20
teardrop	0	0	0	0	0	0	0	0	99	99
Sum	18069	15	0	0	2	3781	0	0	99	21966

Computation time : 78 ms.
Created at 1/25/2020 12:36:33 PM

Table4: Experimental result of ID3

Error rate	0.0016

Values prediction

Value	Recall	1-Precision
normal	1.0000	0.0009
buffer_overflow	0.0000	1.0000
loadmodule	0.0000	1.0000
perl	0.0000	1.0000
neptune	0.0000	1.0000
smurf	0.9976	0.0000
guess_passwd	0.0000	1.0000
pod	0.0000	1.0000
teardrop	1.0000	0.1681

Confusion matrix

	normal	buffer_overflow	loadmodule	perl	neptune	smurf	guess_passwd	pod	teardrop	Sum
normal	18145	0	0	0	0	0	0	0	0	18145
buffer_overflow	2	0	0	0	0	0	0	0	0	2
loadmodule	1	0	0	0	0	0	0	0	0	1
perl	1	0	0	0	0	0	0	0	0	1
neptune	2	0	0	0	0	0	0	0	0	2
smurf	9	0	0	0	0	3686	0	0	0	3695
guess_passwd	1	0	0	0	0	0	0	0	0	1
pod	0	0	0	0	0	0	0	0	20	20
teardrop	0	0	0	0	0	0	0	0	99	99
Sum	18161	0	0	0	0	3686	0	0	119	21966

Computation time: 234 ms.
Created at 1/25/2020 12:35:49 PM

C. Decision Tree (C4.5 algorithm)

count < 99.5000
src_bytes < 28.5000
src_bytes < 22.5000 then label = normal (98.62 % of 145 examples)
src_bytes >= 22.5000 then label = teardrop (100.00 % of 99 examples)
src_bytes >= 28.5000
dst_srv_count < 20.5000

src_bytes < 1478.0000 then label = normal (98.39 % of 186 examples)
src_bytes >= 1478.0000
protocol_type in [tcp] then label = normal (90.00 % of 20 examples)
protocol_type in [udp] then label = normal (0.00 % of 0 examples)
protocol_type in [icmp] then label = pod (100.00 % of 20 examples)
dst_srv_count >= 20.5000
dst_bytes < 8.5000
dst_srv_count < 175.0000 then label = normal (100.00 % of 191 examples)
dst_srv_count >= 175.0000 then label = smurf (100.00 % of 9 examples)
dst_bytes >= 8.5000 then label = normal (100.00 % of 17610 examples)
count >= 99.5000 then label = smurf (100.00 % of 3686 examples)

- *Duration:* variance of time period between two connections during particular time window.
- Protocol: Type of protocol namely TCP, UDP, ICMP, etc.
- Src_bytes: Total number of data bytes from source to destination.
- Destination_bytes: Total number of bytes from destination to source.
- Count: Number of the connection to the same host during a particular time window.

D. K-Nearest Neighbors:

Table1contains the information about the KNN algorithm displayed in the confusion matrix, computed on the learning sample (Classifier performance). The error rate is 0.1740 with computation time 109ms.

VI. RESULT AND DISCUSSION

Among the four comparable algorithms, Naive Bayes has the best accuracy rate compared with other algorithms. The accuracy rate of Naïve Bayes is 94.7%. Finally, we discuss why Naïve Bayes has the best performance compare with others because the K-NN algorithm is also referred to as a lazy learning algorithm it takes a large amount of time to compute the result even for a small dataset. Next, C4.5 basically works similarly to the ID3 but that improves from some of the features of ID3. C4.5 and ID3 take a huge amount of time to produce the result for large datasets. Hence, the naïve Bayes algorithm works very well for large datasets now we take 21,967 data for prediction. From that, we conclude Naïve Bayes works best for this dataset.

ALGORITHM USED	TIME TAKEN	ERROR RATE	ACCURACY
K-Nearest Neighbors	109ms	0.1740	84%
ID3	234ms	0.0016	92.2%
C4.5	468ms	0.0004	89%
Naïve Bayes	78ms	0.0048	94.7%

Table5: Accuracy rate for correlative classification algorithms.

As seen in figure1, from the testing scenario ID3 algorithm shows good detection accuracy. In future work, we try to detect the attack using simulation and to make tests effectively using a real dataset

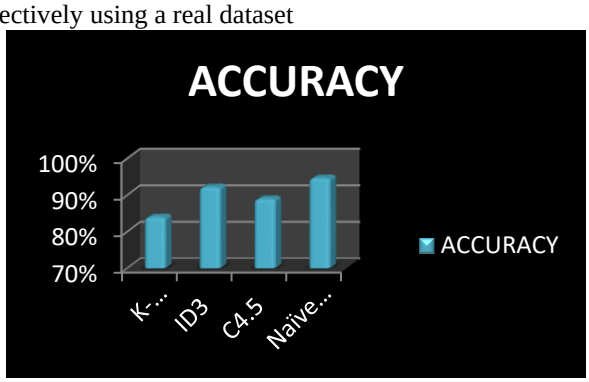

Figure1, Accuracy and error rate are compared.

VII. CONCLUSION AND FUTURE WORK

With their rapid advancement and accuracy, machine learning algorithms are proven to be more efficient and perfect for the data analysis process. From the result, we observed that the comparison of four algorithms ID3 is more efficient for data analysis. Naïve Bayes with error rate 0.0048, computation time 78ms, and Accuracy rate of 94.7%. In the subsequent work, we improve performance by applying the deep learning concept.

REFERENCE

1. S. Revathi. "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection" Vol. 2 Issue 12, December – 2013 IJERTIJERT ISSN: 2278-0181.
2. K.Kuppusamy and S.Malathi. "AN EFFECTIVE PREVENTION OF ATTACKS USING GI TIME FREQUENCY ALGORITHM UNDER DDOS" International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.6, November 2011.
3. KDD Cup Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.
4. Aroosh Amjad, Tahir Alyas*, Umer Farooq, Muhammad Arslan Tariq. "Detection and mitigation of DDoS attack in cloud computing using machine learning algorithm".

5. Gayathri Rajakumaran1 · Neelanarayanan Venkataraman1.” Raghava Rao Mukkamala2. “Denial of Service Attack Prediction Using Gradient Descent Algorithm”. Received: 10 June 2019 / Accepted: 4 October 2019 © Springer Nature Singapore Pte Ltd 2019.
6. Yuanyuan Sun 1, 2, 3a, Yongming Wang 1, 2b, Lili Guo 3c, Zhongsong Ma3, Shan Jin3 and Huiping Wang. “Researching on Multiple Machine Learning for Anomaly Detection”.
7. Ahmad Sanmorino. “A study for DDOS attack classification method” 1st International Conference on Advance and Scientific Innovation (ICASI).
8. Ketki Arora. “Impact Analysis of Recent DDoS Attacks” Ketki Arora et al. / International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397 Vol. 3 No. 2 Feb 2011.
9. J. O. Nehinbe.” A critical evaluation of datasets for investigating IDSs and IPSs Researches”, in IEEE International Conference on Cybernetic Intelligent Systems (CIS), IEEE, 2011, pp. 92–97. doi:10.1109/ CIS.2011.6169141.
10. M. Alkasassbeh, G. Al-Naymat, A. Hassanat, M. Almseidin, “Detecting Distributed Denial of Service Attacks Using Data Mining Techniques”, International Journal of Advanced Computer Science and Applications (IJACSA) 7 (1) (2016) 436–445.
11. R. Koch, M. Golling, G. D. Rodosek, Towards Comparability of Intrusion Detection Systems: New Data Sets, in: TERENA Networking Conference, Vol. 7, 2014.
12. M. H. Bhuyan, D. K. Bhattacharyya, J. K. Kalita, Towards Generating Real-life Datasets for Network Intrusion Detection, International Journal of Network Security (IJNS) 17 (6) (2015) 683–70.
13. J. J. Santanna, R. van Rijswijk-Deij, R. Hofstede, A. Sperotto, M. Wierbosch, L. Z. Granville, A. Pras, Booters – “An analysis of DDoS-as-a service attacks”, in: IFIP/IEEE International Symposium on Integrated Network Management (IM), 2015, pp. 243–251. doi:10.1109/INM. 2015.7140298.
14. J. Wang, I. C. Paschalidis, “Botnet Detection Based on Anomaly and Community Detection”, IEEE Transactions on Control of Network Systems 4 (2) (2017) 392–404. doi:10.1109/TCNS.2016.2532804.
15. M. Ahmad, S. Aftab, and S. S. Muhammad, ―” Machine Learning Techniques for Sentiment Analysis”: A Review,‖ Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, p. 27, 2017.
16. G. P. M. De Farias, A. L. I. De Oliveira, and G. G. Cabral, ―” Extreme learning machines for intrusion detection systems”,‖ Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7666 LNCS, no. PART 4, pp. 535–543, 2012.
17. M. Ahmad and S. Aftab, ―” Analyzing the Performance of SVM for Polarity Detection with Different Datasets”,‖ Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 10, pp. 29–36, 2017.
18. Badr HSSINA, Abdelkarim MERBOUHA, Hanane EZZIKOURI, Mohammed ERRITALI, ” A comparative study of decision tree ID3 and C4.5”, (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications.
19. Johan Baltié, DataMining : ID3 et C4.5, Promotion 2002, Spécialisation S.C.I.A. Ecole pour l’informatique et techniques avancees.

AUTHORS PROFILE

PL.YAZHINI includes Prediction of DDoS attack using Machine Learning alg**,** M.Phil Scholar, Department of Computer Science, Dr.Umayal Ramanathan College for Women, Karaikudi. Her research interest i orithm.

Dr.L.VISALATCHI, Associate Professor, Department of Information Technology, Dr.Umayal Ramanathan College for Women, Karaikudi. Area of research includes Data Mining, Computer Network.