A Comparative Approach for Host Based Intrusion Detection Using Naiyve Bayes and KNN Algorithm

Pushpendra Chaturvedi

Lecturer, SOS in Computer Science and Application, Jiwaji University, Gwalior, Madhya Pradesh, India

Correspondence should be addressed to Pushpendra Chaturvedi; pushpendra19810219@gmail.com

Received 6 March 2024; Revised 19 March 2024; Accepted 28 March 2024

ABSTRACT- Despite the existence of various types of network intrusion detection system, growth of attacks at host level has increased in the present time. Therefore, there is a huge potential of research in this field and which motivates this research work. This paper analyses the pattern of four classes of attacks used to deploy host-based intrusion. KNN and Naïve-Bayes algorithms are employed and compared in this research work to determine the presence of intrusion using standard measures of performance.

KEYWORDS*:* Intrusion detection, K-NN, Naïve - Bayes

I. INTRODUCTION

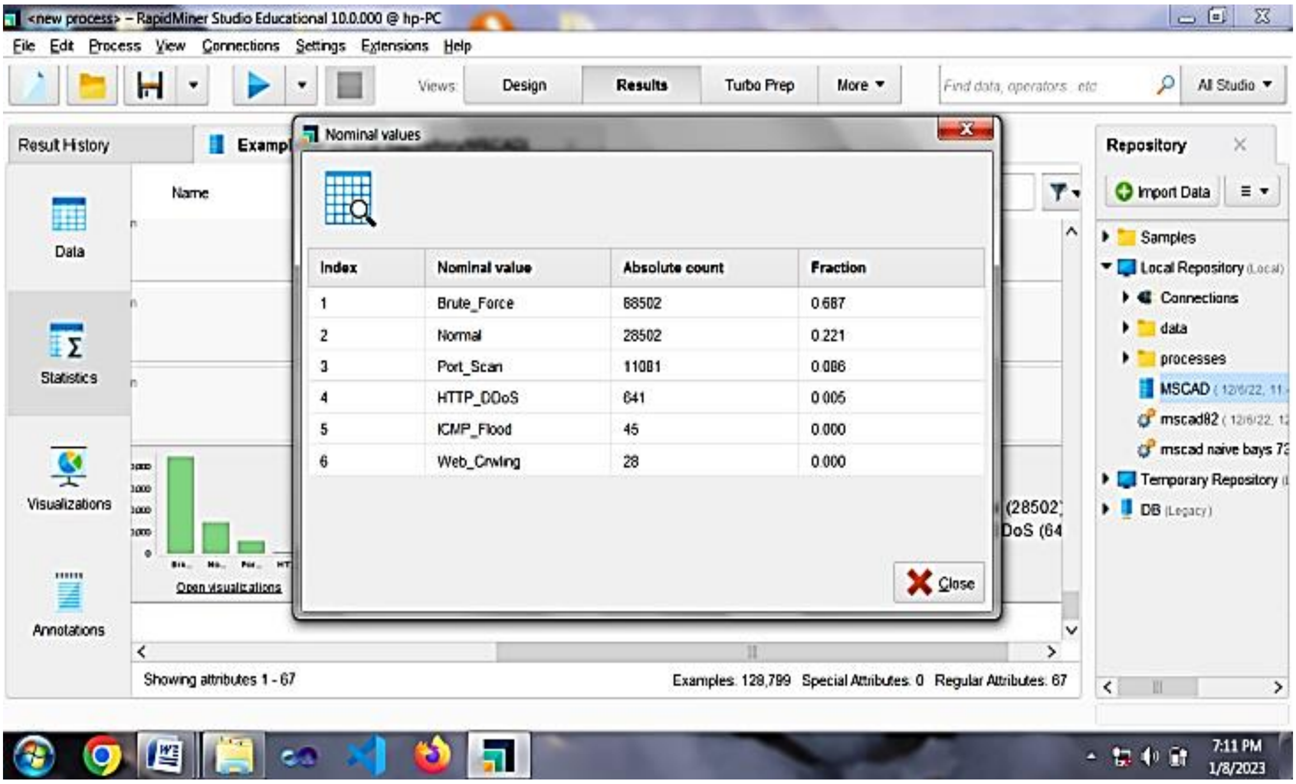
Advancements in the technology and increase in the online activities raises an alarm about security: Intrusion based attacks have been a concerning and challenging field of research over the past years and became potential threat to security in the present era of time. Statistical report and previous research work have identified the fact that there is rapid increase in the rate of intrusion-based attacks. Intrusion can be deployed at network or host level by the attackers. Intruders always keep an eye over the network to monitor network traffic to understand the network traffic of targeted network in network-based intrusion. In host-based attacks, intruders target the host and deploy malicious activities to gain unauthorized access of system. This paper studies five classes of attacks which are practiced by intruders to execute intrusion at host level. Brute- Force is one such class of attack employed by the attacker to crack passwords, login credentials and encryption keys. The other four classes of attacks such as HTTP_DDoS, ICMP_Flood, Port_Scan, and Web_Crwling provides foundation to deploy Host based intrusion using Brute- Force attack. HTTP_DDoS urate type of attack in which the targeted server is saturated with http requests and become unable to respond to legitimate requests. This results into volumetric distributed denial of service attacks. ICMP – flood also falls under the denial-of-service attacks category pings or echo request are used by the intruder to determine state of the host and network connected to it. Intruders employs port scan to determine vulnerable host in the network and evaluate the security levels such as firewalls. Attackers uses data collected from crawler's bots to accomplish intrusion at host level. Various researchers and security experts have contributed in the field of intrusion detection in the past years. The next section of this paper summarizes some of the previous findings of the researchers in the field of host-based intrusion detection.

II. RELATED WORK

Intrusion detection system can be broadly divided into two broad categories network based and host-based intrusion detection system. Network based intrusion detection system monitors entire network whereas host-based intrusion detection system emphasized over individual hosts.[1] Intruders tries to access network in order to gain access of targeted host therefore a lot of research work is done in this field and several researchers has adapted hybrid approach for host-based intrusion detection which comprises network based and host-based intrusion detection system. Authors have proposed various types of methodologies in the past to detect network-based intrusion. Author in [2] determine the type of attack in a network using C4.5 decision tree algorithm over KDD data set. However, it observed in many cases that C4.5 suffers from the problem of over fitting and split attributes which has high impact on the performance with respect to accuracy and prediction. In another study [3] authors have proposed a model for using data mining techniques by integrating support vector machines (SVM), decision trees, and Naïve Bayes. Their findings emphasize over reducing false positive which is a concerning issue in case of intrusion detection system. Authors in [4] proposed use of genetic algorithm and data mining-based Intrusion Detection System. Genetic algorithm performed well in intrusion detection system only for known attacks but not suitable for detecting unknown future intrusions. Another Host-based intrusion detection was proposed by the author in [5] using dynamic and static behavioral models. In another study [6] authors have proposes a methodology to improve the intrusion detection accuracy of anomaly-based intrusion detection systems by employing various machine learning algorithms for classification of normal and attack types. They used the ADFA Linux Dataset which consists of system call traces for attacks on the latest operating systems to check the effectiveness of the proposed intrusion detection models. Authors also developed models and perform simulations for host-based intrusion detection systems based on machine learning algorithms to detect and classify anomalies using the Arena simulation tool. Authors in [7] used host-based intrusion detection methods and data mining techniques to detect the misbehavior and unknown attacks based on the system calls. They used to send mail processor's system dataset for identifying the abnormal behavior of system calls. Authors in their proposed system collected the system calls sequences from send mail process. Further they employed classification rule mining techniques for detecting the intrusion in system calls. Their proposed approach gave good performance and reduces the time complexity as well as false alarm rates. A novel reinforcement learning approach is proposed by the author in [8] for host-based intrusion detection using sequences of system calls. A Markov reward process model is introduce by the authors in their study for modeling the behaviors of system call sequences and the intrusion detection problem is converted to predict the value functions of the Markov reward process. Authors in [9] proposed security model which combines network based and host based with efficient data mining approach to determine any type of intrusion coming from public network or occurring in computer system. It important to study the performance of existing algorithm because hybrid approaches have shown better performance in the field of intrusion detection. Therefore, the next section of this paper studies and analyzes the performance of two prominent multiclass classification algorithm for host-based intrusion detection.

III. ANALYSIS AND RESULTS

This research work is carried out using a dataset containing 128799 instances and 67 regular attributes. There are five classes of attacks namely Brute_Force, HTTP_DDoS, ICMP_Flood, Port_Scan, and Web_Crwling along with the normal instances.This dataset contains 88502 instance of Brute_Force ,11081 instances of Port_Scan , 641 instance HTTP_DDoS , 45 instances of ICMP_Flood and 28 instances of Web_Crwling as depicted in figure1 . However, instances of Brute- force attack is found greater in comparison of other four individual classes of attacks as depicted in the figure 1 and 2.



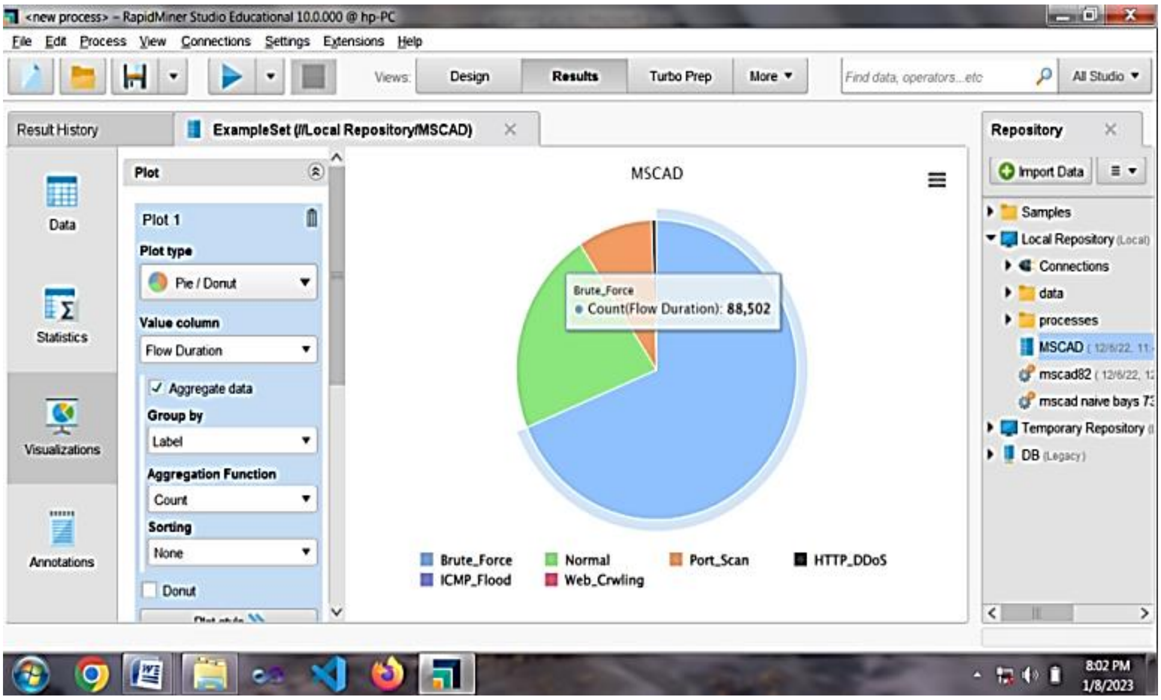
Nominal values

Index	Nominal value	Absolute count	Fraction
1	Brute_Force	88502	0.687
2	Normal	28502	0.221
3	Port_Scan	11081	0.086
4	HTTP_DDoS	641	0.005
5	ICMP_Flood	45	0.000
6	Web_Crwling	28	0.000

Showing attributes 1 - 67

Examples: 128,799 Special Attributes: 0 Regular Attributes: 67

Figure 1: Screenshot Data Set in Rapid Miner



Plot 1

Plot type: Pie / Donut

Value column: Flow Duration

☑ Aggregate data

Group by: Label

Aggregation Function: Count

Sorting: None

□ Donut

Figure 2: Screenshot of number of instances of individual attack classes in Rapid Miner

K – Nearest neighbor (K-NN) and Naïve- Bayes are the two eminent algorithms in the field of intrusion detection. Due to predictive ability of K-NN can be used to classify malicious and non-malicious attack patterns. This algorithm can be employed for classification and regression-based problems. K-NN can handle multi class cases as well as deal with non –linear data without making underlying assumptions. Naïve Bayes is another algorithm which can be used for multiclass prediction-based problems and can work with continuous and discrete data. This algorithm is also scalable, despite of several disadvantages it suffers from the limitation that it assumes that all features are mutually independent. This assumptions would lead to degradation in performance of Naïve – Bayes algorithm and finding predictor features that are completely independent of each other is practically impossible. In this paper, performance of Naïve- Bayes and K-NN algorithm is analyzed since both the algorithms possess multi class prediction ability. This paper compares the performance of K- NN and Naïve -Bayes algorithm based on accuracy, precision and recall. Analysis and implantation of algorithm is done using rapid Miner. Performance of K-NN is evaluated in rapid miner over the dataset to detect the five classes of attacks. Accuracy, Precision and recall is used as standard measures to assess the performance of algorithms in detecting host-based intrusion. Table 1 and Table 2 shows the performance of K- NN and Naïve Bayes algorithm in predicting the five classes of attacks using precision and recall. As it can be seen from the respective tables with respect to recall K-NN comes up with 99.97% in case of Brute_Force , 98.29% in case of port- scan recall values which are higher in comparison of Naiva – Bayes . Naïve – Bayes shows 99.88% and 10.56% recall values in predicting Brute_Force and port- scan respectively. However, in case of HTTP_DDoS and Web_Crwling both the algorithms perform equally with the recall value 93.75% and 12.50% respectively. In case of ICMP_Flood Naïve Bayes algorithm perform better with recall value 92.86% in comparison of K-NN with 71.43%. As far as precision is concerned K-NN scored 99.82, 94.74, 90.91 98.61 precision in predicting Brute_ Force, HTTP_DDoS, ICMP_Flood, Port_Scan and Web_Crwling respectively whereas Naïve – Bayes scored 92.53, 9.97, 0.91 74.36 and 0.05 precision values which are relatively lower in comparison of K-NN. Table 3 and figure 3 illustrates the overall accuracy of the stated algorithms. K-NN perform with the accuracy of (99.6 %) which is higher than the overall accuracy of Naïve – Bayes algorithm (80.4).

Table 1: Performance of Naïve – Bayes Algorithm using Precision and recall

Attacks classes	Precision	Recall
Brute_Force	92.53	99.88%
HTTP_DDoS	9.97	93.75%
ICMP_Flood	0.91	92.86%
Port_Scan	74.36	10.56%
Web_Crwling	0.05	12.50%

Table 2: Performance of K-NN Algorithm using Precision and recall

Attacks classes	Precision	Recall
Brute_Force	99.82	99.97%
HTTP_DDoS	94.74	93.75
ICMP_Flood	90.91	71.43
Port_Scan	98.61	98.29
Web_Crwling	33.33	12.50

Table 3: Comparison of K-NN Algorithm and Naïve – Bayes algorithm based on accuracy

Algorithm	Accuracy
Naïve–Bayes algorithm	80.4
K-NN Algorithm	99.6

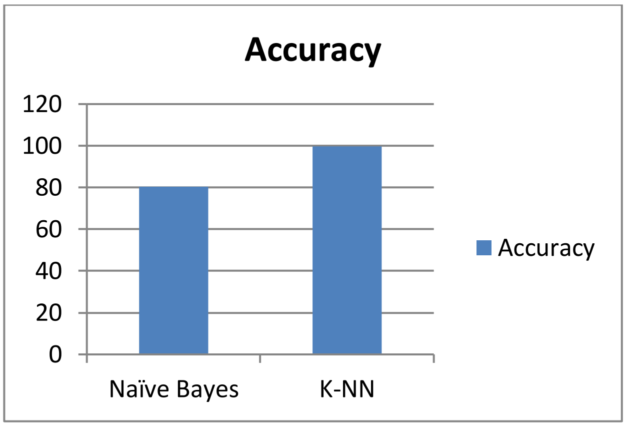


Figure 3: Chart showing overall accuracy of Naïve bayes and K-NN algorithms.

IV. CONCLUSION

This paper studies and analyses the performance of two algorithms Naïve – Bayes and K- NN algorithm for host-based intrusion detection. Performance of K-NN appears more promising during analysis in comparison of Naïve Bayes algorithm. K-NN outperform with an accuracy of 99.6%, however Naïve – Bayes only scale up to 80.4% of an accuracy.to detect host-based intrusion.

REFERENCES

[1] Glenn M. Fung and O. L. Mangasarian, "Multicategory Proximal Support Vector Machine Classifiers", Springer Science and Business Media, Machine Learning, 59, 77–97, 2005.

[2] R. Sahani, C. Rout, J. Chandrakanta Badajena, A. K. Jena, and H. Das, "Classification of intrusion detection using data mining techniques," in Progress in Computing, Analytics and Networking: Proceedings of ICCAN 2017, pp. 753-764, Springer Singapore, 2018.

[3] K. Goeschel, "Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis," in SoutheastCon 2016, pp. 1-6, IEEE, March 2016.

[4] V. K. Kshirsagar, S. M. Tidke, and S. Vishnu, "Intrusion detection system using genetic algorithm and data mining: An overview," International Journal of Computer Science and Informatics ISSN (PRINT), vol. 2231, no. 5292, 2012.

[5] D. Y. Yeung and Y. Ding, "Host-based intrusion detection using dynamic and static behavioral models," Pattern recognition, vol. 36, no. 1, pp. 229-243, 2003.

[6] Y. Shin and K. Kim, "Comparison of anomaly detection accuracy of host-based intrusion detection systems based on different machine learning algorithms," International Journal of Advanced Computer Science and Applications, vol. 11, no. 2, 2020.

[7] P. Ramprakash, M. Sakthivadivel, N. Krishnaraj, and J. Ramprasath, "Host-based intrusion detection system using sequence of system calls," International Journal of Engineering and Management Research (IJEMR), vol. 4, no. 2, pp. 241-247, 2014.

[8] X. Xu and T. Xie, "A reinforcement learning approach for host-based intrusion detection using sequences of system calls," in Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I, pp. 995-1003, Springer Berlin Heidelberg, 2005.

[9] S. K. Singh, N. Chaurasia, and P. Sharma, "Concept and proposed architecture of Hybrid Intrusion Detection System using data mining," International Journal of Engineering and Advanced Technology (IJEAT), vol. 2, pp. 274-276, 2013.

[10] M. Almseidin, J. Al-Sawwa, and M. Alkasassbeh, "Generating a Benchmark Cyber Multi-Step Attacks Dataset for Intrusion Detection," pp. 3679 – 3694, 1 Jan. 2022.

ABOUT THE AUTHOR

Pushpendra Chaturvedi is a graduate from Jiwaji University, Gwalior, Madhya Pradesh, India and Master of Computer Application (MCA) from Rajiv Gandhi Proudyogiki Vishwavidyalaya (R.G.P.V), Bhopal, Madhya Pradesh, India. He has a More than 15 years' experience in academics.