International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Network Intrusion Detection and Measuring the Data Set Performance by Machine Learning Technique (MLT)

A.Abirami, R.M. Bhavadharini, N.B.Prakash, G.R.Hemalakshmi

Abstract*: *Intrusion Detection System (IDS) is the most mainstream approach to protect a computer network from different malicious activities to identify an intrusion. There have been a lot of attempts towards more exceptional performance specifically in IDSs which depends on Data Mining (DM) and Machine Learning Techniques (MLT). Though there is a destructive issue in that available assessment, DataSet (DS), called KDD DS, can't reflect current network circumstances and the most recent attack situations. As far as we could know, there is no possible assessment DS. We present a novel evaluation DS in this paper, called Kyoto, based on the 5 years of actual traffic information, which derived from different sorts of honey pots. This Kyoto DS is utilized for testing and assessing distinctive MLT has examined in this work. The attention was on unprocessed measurements True +ve (TrPo), False +ve (FaPo), True – ve (TrNa), and False – ve (FaNa) to assess execution and to improve the identification rate of IDS.

Keywords* : *Machine learning, Intrusion Detection System, Network Security, Malicious User, Attacks

I. INTRODUCTION

Building a dependable system is a very troublesome task considering all unique sorts of potential attacks. These days, computer network and their administrations generally utilized in industry, business, and all fields of life. Security workforce and everybody who has importance of duty regarding giving assurance to network and its users have severe concern about intruder attacks. An IDS is a device or programming application that examines a system or framework for malicious action or strategy violations. The more significant part of the existing IDSs experiences the unknown effects of the IDS process that corrupts the demonstration of IDSs. Productive aspect choice computation makes the characterization procedure utilized in identification increasingly reliable. MLT is getting prevalent lately. MLT is the utilization of Artificial Intelligence (AI) that gives frameworks the capacity for robotic study and improves as knowledge without being unequivocally programmed. We utilize 7 strategies in this work for an outstanding DS and MLT accessed regarding accuracy, recall, and precision. The seven procedures embraced are K-Means, K-Nearest Neighbors (K-NN), Fuzzy C-Means (FC-M), and Support Vector Machine (SVM), Naïve-Bayes (NB), Radial Basis Function (RBF) and Ensemble Algorithm (EA) containing the previously mentioned 6 Algorithm (ALGO) s. The ROC metric is discovered and progressively practical for the consequences grade of the ALGOs.

By using 3 factors, this ALGOs measured

(i) Precision
(ii) Recall
(iii) Accuracy

The structure of this paper is as follows. Section 2 synthesizes the basics of Network-based IDSs and research conducted in this domain. Section 3 describes the data set used for the experiments — section 4 based on the performance evaluation of machine learning classifiers. Several tests performed for evaluation purposes explained. Finally, in section 5 conclusion and future work of this research are outlined.

The organization of this paper is as per the following. Section II combines the fundamentals of Network-based IDSs and research directed in this domain. Section III depicts the DS utilized for the tests. Section IV depends on the performance assessment of MLT. A few examinations performed for assessment reasons for existing and are additionally clarified. The conclusion and future work of Section V in this research are summarized.

II. RELATED WORK

AuAhmed *et al.* 2016 led a review on network Anomaly Detection Techniques (ADT). The researchers arranged the ADT in the following: (I) Statistical, (II) Classification, (III) Clustering and (IV) Information Theory. The fundamental role of this work is to feature the difficulties and issues of utilizing the generally utilized DS for ADT traffic.

Ensemble Learning Technique (ELT) is a supervised learning ALGO which uses various single models to make a hybrid model accomplish better execution contrasted with that of utilizing an individual model. Xing Fan *et al.* 2017 examined a novel group technique to improve the presentation of paired order. The researchers classified ELT are: (1) approach focussed on getting new trial DS from the underlying trial DS to prepare various diverse single models. (2)

Manuscript published on November 30, 2019.
* Correspondence Author
A.Abirami*, Department of Computer Science and Engineering, Easwari Engineering College, Chennai, India.
R.M.Bhavadharini, Department of Computer Science and Engineering, Easwari Engineering College, Chennai, India.
N.B.Prakash, Department of Electrical and Electronics Engineering, National Engineeering College, Kovilpattii, India.
G.R.Hemalakshmi, Department of Computer Science and Engineering, National Engineeering College, Kovilpattii, India.

Discovering methods to mix individual models. Hu *et al.* 2012 propounded to take care of high false +ve issues in arranging IDS by utilizing Clustering Algorithm (CA). They endeavored two CA specifically K-Means and Fuzzy C Means. DARPA 2000 and LLDOS DS is used for the approval of the ALGO.

A hybrid approach exhibited by Novakov *et al.* 2013 for Network Anomaly Detection (NAD) dependent on numerical and spectral investigation strategies. They study the feasibility of PCA and Wavelet ALGOs in identifying NAD from a marked DS known as Kyoto2006+ giving a valuable benchmark to future analysts. The exploratory outcomes show an improved exactness of the hybrid method in examining with the dual ALGOs independently.

NAD generally utilized by the methods of the Clustering Algorithm (CA). 5 distinctive CA: K-Means, Improved K-Means, K-Medoids, EM clustering and separation based exception IDS utilized. Syarif et al. presented the utilization of CA to identify NAD. The researchers used NSL-KDD DS for the assessment of CA on NAD.

Yao et al. 2012 presented a NAD, which is utilizing an MLT, Random Forest (RF), in mix with entropy metric. RF is most appropriate for the investigation of unpredictable and conveyed information structure implanted in moderate DS. The author asserted that the new system is proficient in NAD.

Ren *et al.* 2009 proposed FC-M CA for IDS. The ALGO applied on 6 DS of KDD Cup DS with 6000 samples each. The detection rate changes somewhere in the range of 55.33% and 91.15%, while the false +ve rate runs somewhere in the range of 0.212% and 3.21%.

thor (s) can send paper in the given email address of the journal. There are two email address. It is compulsory to send paper in both email address.

III. PROPOSED WORK

Dissimilar to a few networks analysis, that exclusively gives DS transmission control through protocol information, each perception in KDD DS contains abnormality names and portrayals.

In this investigation, DS generally used. While it's usually used in formula investigation, there are two significant disadvantages:

- The data set is very 10 years later and along these lines doesn't reflect current attackers, and
- Although there are 41 alternatives portraying a stream, the basic descriptors, similar to supply port, are excluded.

To avoid the problem, Song et al. 2011 made another DS called Kyoto. This new DS contains 24 highlights - 14 highlights are equivalent to in the first DS, and there are 10 additional highlights recorded beneath.

Step 1: IDS Detection: It reflects if IDS set off a caution for the association; '0' signifies any alarms were not activated, and an Arabic Numeral Methods (ANM) the various types of alerts. Enclosure demonstrates the quantity of a similar signal.

Step 2: Malware Detection: It demonstrates if malware, otherwise called vindictive programming, was seen in the association; '0' signifies no malware was watched, and a string shows the relating malware found at the association. Enclosure demonstrates the quantity of similar malware.

Step 3: Ashula Detection: It implies if shellcodes and endeavor codes were utilized in the association, '0' signifies neither shell code nor adventure code was watched, nor an ANM the various types of the shell/endeavors codes. Enclosure demonstrates the quantity of a similar shellcode or endeavors code.

Step 4: Label: It shows whether the session was attacked or not; '1' signifies the session was ordinary, '- 1' signifies realized attack was distinguished in the session, and '- 2' means the implied attack seen in the session.

Step 5: Source IP Address (SIPA): In this session, SIPA utilized.

Step 6: Source Port Number (SPN): SPN demonstrated.

Step 7: Destination IP Address: IP additionally sterilized.

Step 8: Destination Port Number (DPN): DPN investigated this portion.

Step 9: Start Interval: It demonstrated when the session has begun.

Step 10: Time: It shows to what extent the session was being set up.

IV. RESULTS AND DISCUSSION

In the previous segment references, [16] MLT utilized on the named traffic information. The following parameters used to assess MLT (referred in figure 1).

Table 1: Measures of Misperception Matrix

Behavior	Malicious Behaviour	Normal Behaviour
Malicious Attack	True +ve	False -ve
Normal Behaviour	False +ve	True -ve

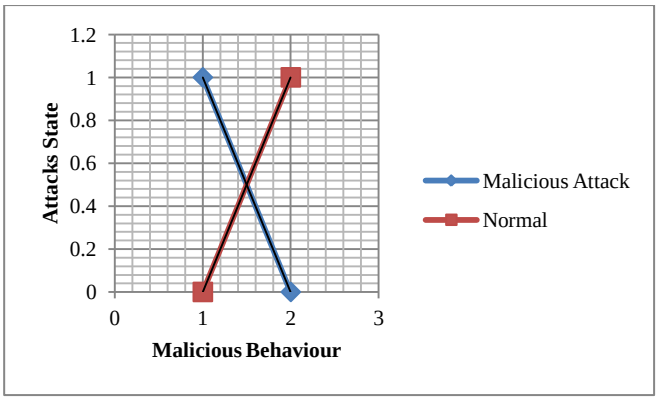

Figure 1. State of Attack Vs. Malicious Action

A. Data Value Forecast

a) TrPo (True +ve): this characterized as the quantity of abnormal Traffic Flow Anticipated (TFA) is anomalous.

b) FaPo (False +ve): this is characterized as the quantity of TFA as anomalous however Normal

c) TrNa (True - ve): this is described as the quantity of TFA as would be expected and ordinary

d) FaNa (False - ve): this is characterized as the quantity of TFA as peculiar however typical.

The objective is to increase TrPo and TrNa while reducing FaPo and FaNa used by the following metrics.

- Precision= TrPo/(TrPo+FaPo)
- Recall = TrPo/(TrPo+FaNa)
- Accuracy = (TrPo+TrNa)/(TP+TrNa+FaPo+FaNa)

The Precision, Recall and Accuracy of all the MLT are measured. Improved value of these metrics gives good taxonomy outcomes.

Table 2: Taxonomy Results

ALGO	Precision	Recall	Accuracy
KM	0.751	0.235	0.856
K-NN	0.9165	0.968	0.9854
FC-M	0.65	0.255	0.736
N-B	0.8167	0.8167	0.8672
SVM	0.885	0.7333	0.996
RBF	0.892	0.893	0.9123
EA	0.7846	0.895	0.9657

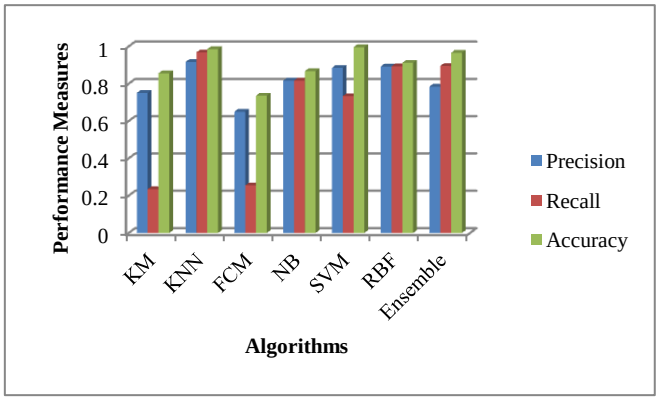

Figure 2. Performance analysis of different Machine Learning ALGO

Outcome acquired for K-NN from table 2 is far superior to that of K-Means and FC-M ALGOs. We can get excellent results by expanding neighborhood value to 17. K-Means and FC-M have two clusters with arbitrarily picked starting centroids. The two clusters indicate abnormality and regular traffic.

For KNN, the precision value is 0.915, and accuracy is pointed as 0.9754 inferring pretty much every abnormality recognized as inconsistent. Both precision and accuracy estimation of K-Means and FC-M are not good, Recall value is 25.56% for these two ALGOs, which demonstrate high evaluations of false -ve. The outcome acquired for N-B ALGOs on the test DS is very high similar to KNN.

The 3 metrics have above 90.19% result. SVM additionally improved than K-Means and Fuzzy C-Means. The Radial Basis Function (RBF) generally has excellent outcomes for all the 3 parameters. This neural system is reasonably like K-Nearest Neighbour model. To improve RBF system we have utilized 17 neurons in the covered layer.

In most cases all MLT have significant outcomes for precision, recall, and accuracy aside from K-Means and FC-M classification method. Though the findings of K-Means and FC-M can be improved. The quantity of clusters by utilizing ALGO for selecting first centroids has increased.

The Ensemble Algorithm (EA) joins the previously mentioned 6 ALGOs. Utilizing EA, we additionally get high qualities for precision, recall, and accuracy, however marginally less worth contrasted and RBF calculation and demonstrated in figure 2.

It is now and then hard for evaluating at this MLT utilizing multi implementation dimensions; in this paper using another measurement called ROC (Receiver Operating Curve). ROC is a plot among TPR and FPR at different threshold settings. ROC value examined to the region under ROC, which is somewhere in the range of 0 and 1. The higher the evaluation of ROC worth betters the MLT. The ROC estimation of each MLT is given in table 3 underneath:

Table 3: ALGOs analysis of ROC Metrics

ALGO	ROC
RBF	0.9127
EA	0.9269
K-NN	0.9651
N-B	0.9341
SVM	0.8673
KM	0.56618
FC-M	0.4148

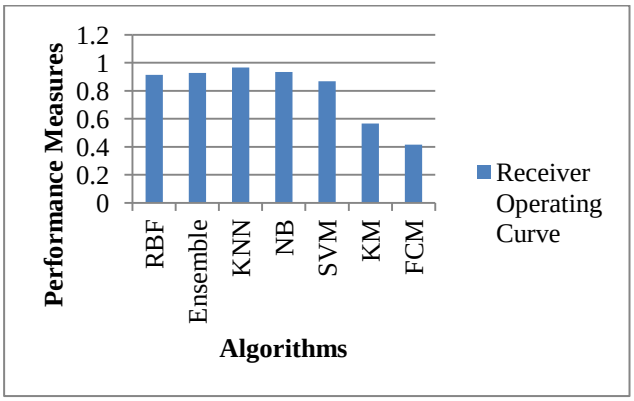

Figure 3. The performance measure of ROC Metrics

The investigation appeared in figure 3; the output of the RBF neural system is high, with an ROC estimation value of 0.9141. K-M and FC-M have the most minimal ROC estimation value of 0.56148. EA has very high outcome closer to RBF. There is a further extent of research here.

V. CONCLUSION

In this paper, a few investigations were performed and examined to assess the efficiency and the accompanying MLT performance: K-M, K-NN, FC-M, NB, SVM, RBF, and EA. Every one of the tests depended on the Kyoto DS. While the more significant part of the study work is utilized in KDD-Cup 1500 DS. Subsequently it doesn't fit with innovations since the DS is over 10 years of age.

In this work, we proposed to utilize Kyoto DS, which contains additional features. Our framework additionally uses data entropy as the traffic includes for IDS.

REFERENCES

1. Xing Fan, Chung-Horng Lung, Samuel Ajila, "An Adaptive Diversity-Based Ensemble Method for Binary Classification," Proc. of the 41st IEEE International Computer Software and Applications Conference (COMPSAC), July 2017
2. Liang Hu, Taihui Li, Nannan Xie, Jiejun Hu, "False Positive Elimination in Intrusion Detection Based on Clustering," Proc. of the 12th International Conference on Fuzzy Systems and Knowledge Discovery, 2012.
3. Stevan Novakov, Chung-Horng Lung, Ioannis Lambadaris, Nabil Seddigh, "Studies in applying PCA and wavelet algorithms for network traffic anomaly detection," Proc.of IEEE 14th International Conference on High-Performance Switching and Routing, 2013, pp. 185-190.
4. Dong Yao, Meijuan Yin, Junyong Luo, Silong Zhang, " Network Anomaly Detection using Random Forests and Entropy of Traffic," Fourth International Conference on Multimedia Information Networking and Security, 2012.
5. Wuling Ren, Jinzhu Cao, Xianjie Wu, "Application of Network Intrusion Detection Based on Fuzzy C-Means Clustering Algorithm," Proc. of the 3rd International Symposium onIntelligent Information Technology Application, 2009.
6. J. Song, H. Takakura, Y. Okabe, M. Eto, D. Inoue, and K. Nakao, "Statistical analysis of honeypot data andbuilding of Kyoto 2006+ data set for NIDS evaluation", Proc. of the 1st Workshop on Building Analysis DataSets and Gathering Experience Returns for Security,2011, pp. 29–36.
7. M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," Journal ofNetwork and Computer Applications, vol. 60, 2016.
8. K. Timm, "Strategies to reduce false positives and false negatives in NIDS," Tech. Rep., Access Date 10 Oct 2015.
9. Han Li, QiuXin Wu, "Research of Clustering Algorithm based on Information Entropy and Frequency Sensitive Discrepancy Metric in Anomaly Detection," Proc. of theInternational Conference on Information Science andCloud Computing Companion, 2013.
10. Stevan Novakov, Chung-Horng Lung, Ioannis Lambadaris, Nabil Seddigh, "Combining statistical and spectral analysis techniques in network traffic anomaly detection," Proc.of IEEE Conf. on Next Generation Networks andServices, 2012, pp. 94-101.
11. A. Lakhina, M. Crovella, and C. Diot, "Mining Anomalies Using Traffic Feature Distributions," Proc. of ACMSIGCOMM, 2005.
12. Iwan Syarif, Adam Prügel-Bennett, Gary B. Wills, "Unsupervised clustering approach for network anomaly detection," Networked Digital Technologies Communications inComputer and Information Science, vol. 293. BerlinHeidelberg: Springer, 2012, pp.135–45.
13. Z. Muda, W. Yassin, M.N. Sulaiman, and N.I. Udzir, "A K-Means and Naïve Bayes Learning Approach for Better Intrusion Detection," Information Technology Journal,2011.
14. Shenghui Wang, "Research of Intrusion Detection Based on an Improved K-means Algorithm," Proc. of the 2nd International Conference on Innovations in Bio-inspiredComputing and Applications, 2011.
15. S.A. Mulay, P. R. Devale, G.V. Garje, "Intrusion Detection System using Support Vector Machine and Decision Tree," International Journal of ComputerApplications, vol. 3, no. 3, 2010.
16. R. Kiruba Buri and T. Jayasankar, "Intelligence Intrusion Detection Using PSO with Decision Tree Algorithm for Adhoc Networks", Bioscience Biotechnology Research Communications, Special Issue Recent Trends in Computing and Communication Technology, Vol. 12, No.2, March (2019),pp.27-34

..