INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-2102)
Available online at: https://www.ijariit.com

Network intrusion detection system

Kj Bhoomika
kjbhoomika.is16@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Impana N.
impanan.is16@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Priya D.
priyad@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

ABSTRACT

With fast development of PC utilization and PC network, the security of the PC framework has turned out to be very significant. Businesses are looking for consistently new sorts of assaults. As the danger turns into a genuine chapter year by year, interruption discovery advancements are essential for organization and PC security. An assortment of interruption recognition approaches be available to determine this severe condition; in any case, the fundamental problem is execution. It is imperative to increment the discovery rates and lessen bogus alert rates in the space of interruption recognition. To recognize the interruption, different methodologies have been created and proposed over the most recent decade. This paper effectively compares the accuracy of different classification algorithms, like algorithms like Support Vector Machine (SVM), Naive Bayes, KNN, Decision Tree[15]. This study aims to perform a comparative analysis of these different machine learning algorithms on datasets available to predict which model best suits the intrusion detection.

Keywords— *Intrusion Detection, Classification, Alert rates, Interruption, Accuracy, SVM, KNN, Decision Tree, Naïve Bayes.*

1. INTRODUCTION

The organization basically based absolute interruption recognition frameworks (NIDS) are insightfully dispensed contraptions that latently investigate site guests passing through the devices they might be mounted. NIDS can be equipment or programming based on absolute contraptions associated with heaps of local area mediums, which incorporates Ethernet, FDDI, and others, depending on the producer. NIDS routinely have local area interfaces. One is for wanton being mindful of local area discussions, while the option is for following and revealing.

Despite the fact that there is an assortment of NIDS merchants, all frameworks seem to work in one of two different ways: signature-based or abnormality-based. Both are apparatuses for recognizing generous and pernicious traffic. Fast organization information blockage, tuning issues, encryption, and mark creation slack time are for the most part possible issues with NIDS.

1.1 Study on related work

Amarudin, and Ridi Ferdiana in 2020[2] did a systematic review of IDS. The paper had a research mainly to analyse and detect the different research trends of techniques, methods and the datasets that are being used on the IDS.
It also had researches done on how the accuracies could be increased in different systems.

The authors[3] did a research on the IDS types, mainly focusing on the machine learning based intusion systems. Identifying known and unkown attacks was a major concern in some IDS methodologies. The most known machine learning IDS systems are provided in this paper and also the needs for an accurate detection is discussed.

Manohar H. Bhuyan, H.J. Kashyap[6] researched on distributed denial of service attack. The paper presented a survey of the attacks, detection methods and the tools that are used in the wired networks. It also enhances the issues in real life, research challenges and also the solutions in particular areas.

S. Ganapathy, P. Yogesh, and A. Kannan[19] proposed a paper on an intelligent multi level classification technique. The algorithm that is used is a combination of a tree classifier that makes use of labeled training data and Multiclass SVM algorithms. It is used to improve the detection rates and the alarm rates that are false.

2. OVERVIEW OF THE TECHNIQUES

K-Means Clustering

K-means is quite possibly, the most essential and generally used algorithms in machine learning. It is one of the unsupervised stratergy. K-means clustering will fing groups in the obtained result, with variable K which denotes the number of groups. Based on the datasets characteristics, data points are assigned by the algorithm to one of the K group.

Bayesian Network

Bayesian network models are a kind of probabilistic graphical model[1]. By plotting the conditions on the edge of a coordinated chart, it intends to exploit restrictive reliance. A fact used in this model states that all the nodes which are not

bound by an edge are independent to create a directed acyclic graph.

Random Forest Classifier

The Random Forest algorithm can combine many algorithms together for classification purposes hence it is called an ensemble classifier. Many decision tress are built on a random subset of the data. It adds up each tree's total votes to determine the test's class or gives weight to each tree's contribution

Support Vector Machine

A separating hyper-plane defines the Support Vector Machine as a discriminative classifier[14]. All in all, given named preparing information (supervised learning), the calculation yields an ideal hyper-plane that classifies new models.

Decision Trees

Decision tree is a decision-making tool that uses a tree-like graph or model to represent decisions and their potential results, such as chance event outcomes, resource costs, and more. It's one method of displaying an algorithm with just conditional power[5].

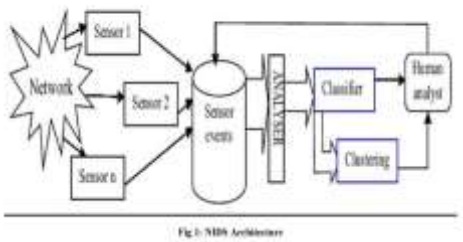
Every association is named one or the other ordinary or an assault, with precisely one explicit assault type. Every association record comprises around a hundred bytes. The assaults are categorized into four principle gatherings:

- DOS: disavowal of-administration
- R2L: unapproved access from a distant machine
- U2R:unapproved admittance to neighbourhood root advantages
- examining: reconnaissance and another testing. Each gathering has separate assaults, and there is a sum of 21 sorts of assaults.

3. METHODOLOGY

Solo learning calculations can "learn" the normal example of the organization's regular example and flag inconsistencies with no designated data[5]. It can recognize new sorts of interruptions, however, bogus positive admonitions are normal. Subsequently, just a single unaided calculation K-implies bunching is utilized. A labelled dataset can be utilized and a managed AI model to show the contrast between a normal and an assault parcel in the organization to limit bogus positives.

The regulated model is able to do deftly taking care of perceived assaults just as perceiving variations of those assaults. Standard administered calculations are utilized. The major critical and varying cooperation of starting with AI models is getting the data which is reliable. KDD Cup 1999 is used to build judicious models that fit for perceiving interferences or attacks, and critical affiliations.



3.1.1 Dataset description: The proposed model is evaluated using the KDD99 data set[6]. The number of total instances used are around 494,021. The dataset contains about fourty-one attributes, which includes 'the class', that indicates if an instance given of an attack is a normal instance or not.

3.1.2 Data preprocessing: Most of the datasets in large-scale generally contain noisy, unwanted and other types of the data which generally show difficult obstacles in data modelling and knowledge detection. In general, IDC algorithms work with 1 or more types of raw data input, such as Support Vector algorithm which only works with data based on numerical values. Henceforth, it gets ready information and change all information of the dataset to numerical information.

3.1.3 Standardization: One of the stable machine learning key technique to obtain a reliable result is the standardization. Some feature values can range from very tiny to one of the most largest value. Hence, the process that is analyzed can explode the scale. The standardized characteristics is processed by scaling to a unit variance in the Spark Millib of Spark-Chi SVM model.

3.1.4 Dataset Visualization

1. There are 2 weeks of attackless occurences and 5 weeks of attack occurrences, which makes it suitable for anomaly detection.
2. There are five major groupings: These are the DOS (Denial of Service), Probe, R2L (Root Local), U2R (User 2 Root) and Normal[6].
3. The dataset contains a total of thirty eight attacks in which twenty four types are in training and fourteen attacks in testing. The attacks which are new(14) are intented to assess the ability of IDS to infer to mysterious strike or attack.
4. It is difficult to find 14 new attacks for a ML algorithm IDS.
5. To attack cases, the dataset is highly skewed. Around 80% of flow is the attack traffic [12]. Ordinarily, a typical organization contains roughly 99.99% percent of regular instances. This theory is broken by KDD99.
6. Rare attacks in the KDD99 dataset are the U2R and R2L attacks.
7. The amount of identical records in the training and testing datasets one-sideness result for often attacks of DOS and normal instances.
8. For most machine learning algorithms, KDD99 dataset is one of the massive dataset, but most studies only use a small portion[7].

	Training Size	(%)	Test Size	(%)
Normal	972781	19.85	60593	19.48
DOS	3883390	79.27	231455	74.41
Probe	41102	00.83	4166	01.33
U2R	52	00.001	245	00.07
R2L	1106	00.02	14570	04.68
Total	4898431	100	311029	100

Table 1: Dataset Visualization

Heatmap to visualize the features is represented in Fig2.

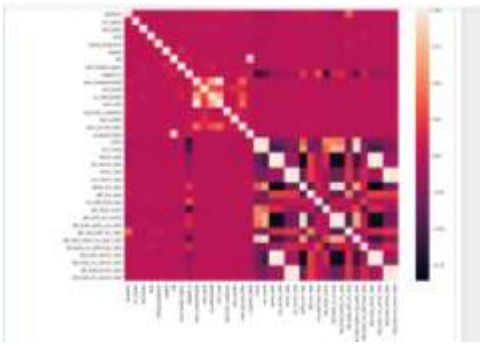


Fig2. Heatmap

4. COMPREHENSIVE ANALYSIS

Comprehensive analysis of different algorithms are presented here, such as knn classification, support vector machine based, k-means based, decision tree based for intrusion detection. As we have discussed earlier, major classification is done based on intrusion detection with respect to the detection rate, time and false alarm rate achieved by the different methods.

Random forest class classifier is utilized to choose important highlights among every one of the highlights set accessible. A structured presentation comprising of highlights and their significance esteem is plotted, the highlights with greatest significance are chosen. Table 2 gives the analysis of the comparision prediction done

Table 2. Comparing prediction of different models

Model: Knn Model	Anamoly: 8931	Normal: 13615
Model: Byesian Model	Anamoly: 17215	Normal: 27875
Model: Logistic Regression	Anamoly: 26275	Normal: 41359
Model: Decision Tree	Anamoly: 38576	Normal: 51602
Model: SVM	Anamoly: 47505	Normal: 65217

Mathematical representation of the accuracy calculation:
The accuracy or also known as classification rate is the one that measures how precise the IDS detects regular and abnormal traffic. It is the percentage of all the accurately predicted cases to all other instances.

Accuracy = (true positive + true negative)/(true positive + true negative + false positive + false negative)

5. CONCLUSION

This paper gives a broad audit of the organization interruption location instruments dependent on the ML strategies to furnish the new specialists with the refreshed information, ongoing patterns, and progress of the field. The proposed calculation will accomplish a recognition exactness of over 95%, with a bogus positive pace of under 1% and a bogus negative pace of around 1%. In comparison, the conventional SVM-based plan will accomplish a discovery precision of approximately 90%, with a fake positive rate of 5% and a bogus negative of 10%. One of the significant matters is the component determination on which the most bit of the discovery exactness depends. More examination can be accomplished by joining few highlights to decrease the time it takes to identify an oddity in network traffic.

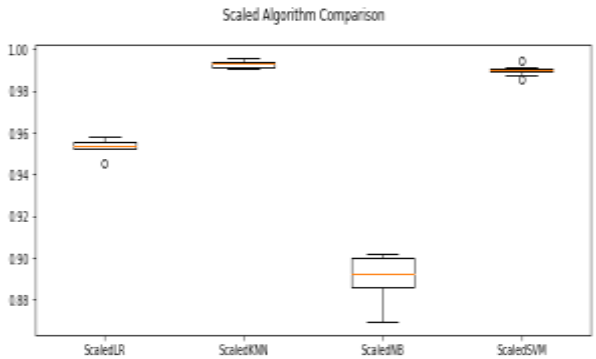


Fig 2. Model Evaluation

6. REFERENCES

[1] M. Elbasiony, Reda, etal, "A hybrid network intrusion detection framework based on random forests and weighted k-means," Ain Shams Engineering Journal, Vol.4, No.4, pp.753-762,2013.

[2] S.A.Joshi and Varsha S.Pimprale, "Network Intrusion Detection System (NIDS) based on data mining," International Journal of Engineering Science and Innovative Technology (IJESIT),Vol.2,No.1,pp.95-98,2013.

[3] Sannasi Ganapathy, etal., "Intelligent feature selection and classification techniques for intrusion detection in networks: a survey," EURASIP Journal on Wireless Communications Networking,Vol.1,pp.271,2013.

[4] Louvieris, Panos, Natalie Clewley and Xiaohui Liu, "Effects-based feature identification for network intrusion detection," Neurocomputing, Vol. 121, pp. 265-273, 2013.

[5] Jaehak Yu,etal., "An in-depth analysis on traffic flooding attacks detection and system using data mining techniques," Journal of Systems Architecture, Vol.59, No.10, pp.1005-1012,2013.

[6] Monowar H. Bhuyan, et al., "Detecting distributed denial of service attacks: methods, tools and future directions," The Computer Journal, Vol.57, No.4, pp.537-556, 2013.

[7] Iftikhar Ahmadetal., "Enhancing SVM performance in intrusion detection using optimal feature subset selection based on genetic principal components," Neural computing and applications, Vol.24, No.7-8, pp.1671-1682, 2014.

[8] Wenying Feng, etal., "Mining network data for intrusion detection through combining SVMs with ant colony networks," Future Generation Computer Systems, Vol.37, pp.127-140, 2014.

[9] G. Meera Gandhi, "Machine learning approach for attack prediction and classification using supervised learning algorithms", International Journal of Computer Science & Communication, Vol. 1, No. 2, pp. 247-250. July-December 2010.

[10] K. AbdJalil, and S.Mara, "Comparison of machine learning algorithms performance in detecting network intrusion", In Proceedings of Networking and Information Technology (ICNIT), pp. 221 – 226, Manila 2010.

[11] S. Mukkamala, A.H.Sung, and A. Abraham, "Intrusion detection using an ensemble of intelligent paradigms" J. Netw. Comput. Appl., vol. 28, no. 2, pp. 167–182, 2005.

[12] M.Govindarajan and R.Chandrasekaran, "Intrusion Detection using an Ensemble of Classification Method s," In Proceedings of World Congr. Eng. Computer. Sci., vol. I, no. October, 2012.

[13] Karan Bajaj, Amit Arora, "Improving the Intrusion Detection using Discriminative Machine Learning Approach and Improve the Time Complexity by Data Mining Feature Selection Methods", International Journal of Computer Applications (09 75 – 8887) Volume 76–No.1, August 2013.

[14] W.K.Lee, and S.J.Stolfo, "A data mining framework for building intrusion detection model," Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA: IEEE Computer Society Press, pp. 120-132, 1999.

[15] Marjan Bahrololum, Elham Salahi, Mahmoud Khaleghi,"An Improved Intrusion Detection Technique based on two Strategies Using Decision Tree and Neural Network," Journal of Convergence Information Technology,Vol.4, No.4, December 2009.

[16] M.Bahrololum,E.Salahi and M.Khaleghi, "Anomaly intrusion detection design Using Hybrid of Unsupervised and supervised neural Network," International Journal of Computer Networks & Communications (IJCNC), Vol.1, No.2, July 2009.

[17] Latifur Khan, Mamoun Awad, Bhavani Thuraisingham,"A new intrusion detection system using support vector machines and hierarchical clustering," Journal of VLDB Journal, vol.16, pp.507-521, 2007.

[18] Iftikhar Ahmad,Azween Abdullah,Abdullah Alghamdi,Muhammad Hussain,"Optimized intrusion detection mechanism using soft computing techniques,"Telecommun System,2011.

[19] S.Ganapathy, P.Yogesh, and A.Kannan," An Intelligent Intrusion Detection System for Mobile Ad-Hoc Networks Using Classification Techniques," Advances in Power Electronics and Instrumentation Engineering, Communications in Computer and Information Science Vol.148, pp 117-122,2011.

[20] Yurcik W, "Contrlling intrusion detection systems by generating false positives:squealing proof-of-concept", in Proceeding of the IEEE local Computer Network Conference, 2002. pp.93--101.

[21] Yu Guan, Nabil Belacel and Ali A. Ghorbani,"Y-Means: A Clustering Method for Intrusion Detection," Canadian Conference on Electrical and Computer Engineering, vol.2, pp. 1083- 1086, 2003.

