Smart Approach For Attack Identification Using Machine Learning Classification Techniques

Marmita Solanki[1], Prof. Nimesh V. Patel[2]
[1]Dept of computer engineering
[2]Assistant Professor, Dept of computer engineering
[1, 2] LDRP institute of technology and research, KSV University - GANDHINAGAR

Abstract- *To examine malicious activity that happens in a framework, intrusion detection system is utilized. Execution of an Intrusion Detection is for the most part relies upon exactness. Precision is improved for decreasing bogus alerts and to build identification rate. Many AI procedures like SVM (Support Vector Machine) and Naïve Bayes are applied. A fundamental impediment of Intrusion Attack Detection Systems (IADSs), in spite of their location technique, is the tremendous number of alarms they produce consistently that can easily debilitate security administrators. These procedures are notable for tackling order issues. The precision is estimated subsequent to utilizing various sorts of methods. The planned XGBoost–DNN model utilizes XGBoost technique for feature choice followed by deep neural network (DNN) for classification of network intrusion. The XGBoost–DNN model has 3 steps: normalization, feature choice, and classification.*

I. INTRODUCTION

Intrusion Detection is used to recognize abnormal behavior takes place in a network or system. Hence intrusion is one of the major issues in network security. Various techniques of intrusion detection are performed to get highest accuracy. Detection rate and false alarm rate plays an important role for the analysation of accuracy. Algorithms like SVM (Support Vector Machine) and Naïve Bayes are applied and classification can be addressed by these algorithms. Several researchers proposed network intrusion detection systems (NIDS) to protect cloud environment from cyber attacks. Recently, machine learning techniques for intrusion detection have proven their efficiency.

Data Mining

Data mining denotes to extracting or "mining" information from massive amounts of knowledge. several people provide data processing as another word for an additional wide used word, information Discovery from information, or KDD.

Knowledge discovery as procedures illustrated in Figure 1.1 and involves of an iterative order of the following steps:

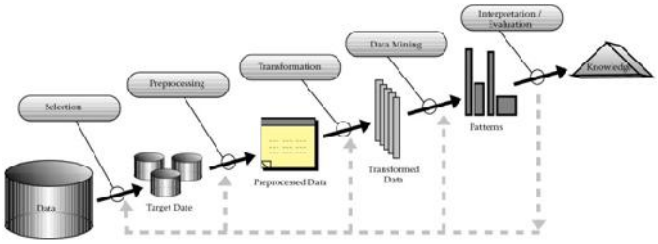


Figure 1.1 KDD process in Data Mining[3]

1. Data cleaning to eliminate noise and unreliable data.
2. Data integration where many data sources may be pooled.
3. Data selection where data applicable to the analysis task are reclaimed from the database.
4. Data transformation where data are transformed or fused into forms suitable for mining by accomplishing summary or aggregation operations.
5. Data mining an indispensable procedure where intelligent approaches are applied in order to extract data patterns.
6. Pattern evaluation to find the truly interesting patterns signifying knowledge grounded on some interestingness measures.
7. Knowledge presentation where visualization and knowledge representation methods are used to present the mined knowledge to the customer.

Steps 1 to 4 are different forms of data preprocessing, where the data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base [1].

Machine Learning

Machine learning is a subfield of artificial intelligence (AI) [2]. The objective of machine learning commonly is to realize the structure of data and fit that data into models that can be understood and exploited by people.

Machine learning systems in its place let computers to train on data inputs and practice statistical analysis in order to output values that fall inside a particular range. As of this, machine learning enables computers in building models from sample data in order to mechanize decision-making procedures centered on data inputs. Two of the most broadly accepted machine learning approaches is **supervised learning** which trains algorithms based on example input and output data that is labeled by humans, and **unsupervised learning** which provides the algorithm with no labeled data in order to allow it to find structure within its input data.

Supervised Learning

In supervised learning, the pc is delivered with sample inputs that are characterized with their most well- liked outputs. The determination of this method is for the algorithmic rule to be capable to "learn" by equaling its real output with the "taught" outputs to get errors, and rework the model consequently. supervised learning thus uses patterns to guess label values on extra unlabelled knowledge. a standard use case of supervised learning is to utilize past knowledge to predict statistically probable future happenings.

Unsupervised Learning

In unsupervised learning, knowledge is untagged, therefore the learning rule is left to find commonalities among its input file. As untagged knowledge square measure additional ample than labeled knowledge, machine learning ways that assist unattended learning square measure in the main valuable. the target of unattended learning is also as direct as crucial hidden patterns within a dataset, however it's going to even have a objective of feature learning, that lets the machine machine to automatically learn the representations that square measure needed to reason data.

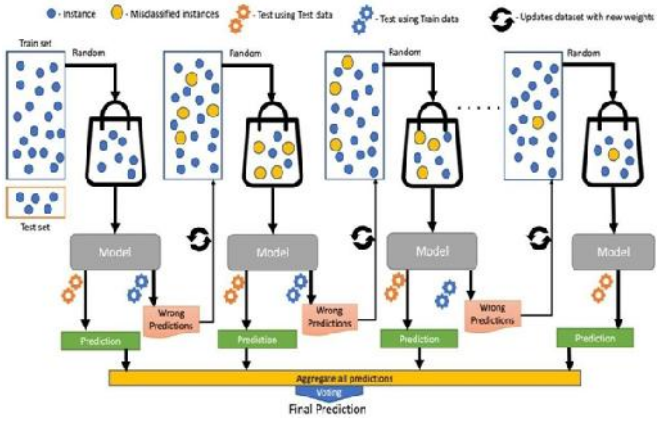
Learning approaches will be classified into linear and nonlinear approaches. Linear approaches area unit easier, whereas nonlinear approaches area unit additional versatile in behavior. For supervised learning, the strategies will be to boot classified as classification- or regression-based strategies. Classification-based strategies try and classify the information by separate and categorical labels, whereas regression-based strategies work the information to a nonstop perform and therefore work with continuous labels for the information [4].
For unsupervised learning, the strategies area unit primarily categorised as clump technique, that cluster the information into clusters supported underlying similarities [4].

Problem Statement

Now a days due to excessive volume of data, false alarm report intrusion to network gets increased and detection accuracy gets reduced. This is one of the major issues when the system encounters unknown attacks. Due to large volumes of IADS false alarms, it is a quite tough task for the security officers to investigate manually which are the real suspicious alarms and thereafter take proper action against them. Even sometimes, some real suspicious alarms are ignored mistakenly by the security officer due to large volumes of false alarms and thereby mistakenly interpret a real alarm to be a false alarm.

II. PROPOSED METHOD

Proposed method is optimized XGAdaptive Boost classifier which is helps us to find the unwanted and wanted mails from the data. XGAdaptive Boost work as the decision tree but in decision tree trees are not connected to each other where as in XGAdaptive Boost trees are inter connected to each other. AS shown in the figure Firstly the data divide into two parts training and testing. From the training set randomly data picked for the input and goes into the model for the training. The model gives the output in two parts which is prediction and wrong prediction. The wrong prediction data again take as the input for the next round and again perform all steps and gives the output. This process is continuing up to the result of wrong prediction is not going to come 0. After these entire rounds we find the actual result of our data.



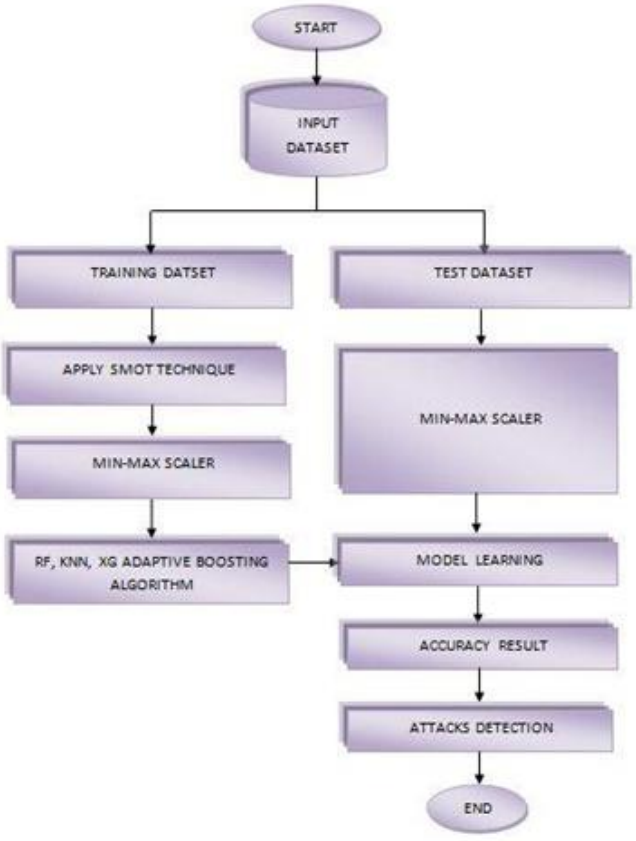
Proposed Algorithm:

- Take Input from Dataset.
- Data-preprocessing from Dataset.
- Divide Training (80%) and Testing (20%) data from Dataset.
- Use SMOT technique on Dataset.
- MIN-MAX Normalization on Dataset.

- Classification Model using XG ADAPTIVE BOOSTING Algorithm .
- Model Learning.
- Intrusion Detection Result.

In my proposed algorithm first take a input from the dataset then applying the preprocessing on the data. After preprocessing the data is divided into two part training and testing then applying the smote class to balance the imbalanced data. Then apply min max normalization to complete the data. After normalization applying the different techniques to train the data. The train data train the 2nd part of testing data and do the model evaluation and gives the accurate output.

Proposed architecture



Advantages of the method:

- Implements parallel processing .
- Allow users to define custom optimization objectives and evaluation criteria.
- Allows user to run a cross-validation at each iteration of the boosting process and thus it is easy to get the exact optimum number of boosting iterations in a single run.
- Voting approach

III. RESULTS

Accuracy - Accuracy is that the most intuitive performance live and it's merely a magnitude relation of properly foreseen observation to the entire observations. One might imagine that, if we've got high accuracy then our model is best.Yes, accuracy could be a nice live however only if you've got regular datasets wherever values of false positive and false negatives ar virtually same. Therefore, you've got to appear at alternative parameters to guage the performance of your model. For our model, we've got got zero.803 which suggests our model is approx. eightieth correct.

Accuracy = TP+TN/TP+FP+FN+TN

Precision - Precision is that the quantitative relation of properly foretold positive observations to the overall foretold positive observations. The question that this metric answer is of all passengers that tagged as survived, what percentage really survived? High preciseness relates to the low false positive rate. we've got 0.788 preciseness that is pretty sensible.

Precision = TP/TP+FP

Recall (Sensitivity) - Recall is that the magnitude relation of properly foretold positive observations to the all observations in actual category - affirmative. The question recall answers is: Of all the passengers that actually survived, what percentage did we tend to label? We have got recall of 0.631 which is good for this model as it’s above 0.5.

Recall = TP/TP+FN

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall. In our case, F1 score is 0.701.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

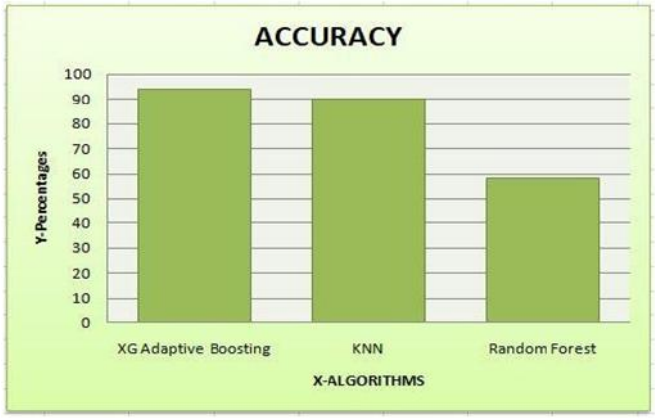
Dataset Distribution %	XG Adaptive Boosting %	Random Forest %	KNN %
50-50	94	60	92
60-40	94	61	91
70-30	94	71	90
80-20	94	58	90

TABLE-1 . ACCURACY TABLE

Dataset Distribution %	XG Adaptive Boosting %			Random Forest %			KNN %		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
50-50	81	83	82	52	58	47	84	80	81
60-40	81	89	84	54	65	48	79	84	81
70-30	73	75	74	57	56	55	67	69	68
80-20	74	75	74	50	48	44	70	69	69

TABLE-2. PRECISON-RECALL-F1 SCORE TABLE

Comparison chart for the method:



IV. CONCLUSION

Extensive analysis goes on within the field of laptop intrusion detection and a number of other IADSs area unit already developed. however their performance is poor by manufacturing false positives at higher rate. Researchers projected many intrusion detection approaches and every detection approach is appropriate just for police work a selected style of attack(s). due to restricted attack coverage of every approach, there's associate degree imperative have to be compelled to arrive of a generic detection approach that handles most sorts of attacks. For that it's needed to know and analyze the techniques that area unit already investigated by many researchers. Keeping that in sight here, we've got created an effort to review the renowned intrusion detection approaches. Comparison of varied approaches is formed to indicate the strength and weakness of those approaches. we have a tendency to hope this study are going to be helpful for analysisers to hold forward research on system security for styles of IADS that not solely can have known strengths however conjointly overcome the drawbacks.

REFERENCES

[1] Venkatraman, S., B. Surendiran, and P. Arun Raj Kumar. "Spam e-mail classification for the internet of things environment using semantic similarity approach." The Journal of Supercomputing 76.2 (2020): 756-776.

[2] Cormack, Gordon V. "Email spam filtering: A systematic review." (2008).

[3] A. Qaroush, I. M.khater, M. Washaha, “Identifying spam e-mail based-on statistical header featuresand sender behavior,” Proceedings of the CUBE International Information Technology Conference,

[4] Ferrara, E.: The history of digital spam. arXiv preprint arXiv :1908.06173 (2019)

[5] Adriana-Christina Enache, Victor Valeriu Patriciu, “Intrusions Detection Based on Support Vector Machine Optimized with Swarm Intelligence”, 9th IEEE international symposium on Applied Computational Intelligence and Informatics”, P.P. 978-1-4799-4694-5/14, May 2014

[6] Rousseeuw PJ, Hubert M(2018) Anomaly detection by robust statistics. Wiley Interdiscip Rev Data Min Knowl Disc 8, Springer 2020

[7] Hudan Studiawan, Chritian Payne, Ferdous Sohel; “Graph Clustering And Anomaly Detection of Access Control log for Forensic Purposes”, ELSEVIER, Digital Investigation

[8] https://www.google.com/search?sxsrf=ALeKk00CM8QbbB0n-GyJAqtEGIHmUQvx2w%3A1614848728025&ei=2KJAYO-FAY2Y4-EP_am6kAw&q=what+is+google+colab&oq=what+is+&gs_lcp=Cgdnd3Mtd2l6EAMYAzIECCMQJzIECCMQJzIECCMQJzIECAAQQzIECAAQQzIECAAQQzIECAAQQzIHCAAQsQMQQzIECAAQQzIECAAQQzoHCAAQsAMQQzoFCAAQsQM6CggAELEDEIMBEEM6AggAOgcIIxDqAhAnOggIABCxAxCDAVDjNViDV2CmZWgCcAJ4BIABAyAKIAfUUkgEIMC4xMS4yLjGYAQCgAQGqAQdnd3 Mtd2l6sAEKyAEKwAEB&sclient=gws-wiz

[9] https://www.google.com/search?q=knn+algorithm&oq=knn&aqs=chrome.0.69i59j69i57j0i67l2j0i20i

263j0i67l2j69i61.2347j0j7&sourceid=chrome&ie= UTF-8
[10] The economic of Spam, 2012. Available from: http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.26.3.87.
[11] Cormack, Gordon V. "Email spam filtering: A systematic review." (2008).
[12] https://www.tutorialspoint.com/google_colab/what_is_google_colab.htm
[13] https://www.geeksforgeeks.org/machine-learning/
[14] Chauhan, Himadri, Vipin Kumar, Sumit Pundir, and Emmanuel S. Pilli. "Comparative Analysis and Research Issues in Classification Techniques for Intrusion Detection." In Intelligent Computing, Networking, and Informatics, pp. 675-685. Springer India, 2014.
[15] Chandolikar, N. S., and V. D. Nandavadekar. "Comparative Analysis of Two Algorithms for Intrusion Attack Classification Using KDD CUP Dataset,"."International Journal of Computer Science and Engineering (IJCSE) Vol 1 (2012): 81-88.