Intrusion Detection System from Machine Learning Perspective

P. Roshni Mol
Ph.D. Research Scholar,
Department of Computer Science,
Sri Sarada College for Women (Autonomous),
Salem-16.

Dr. C. Immaculate Mary
H.O.D & Associate Professor,
Department of Computer Science,
Sri Sarada College for Women (Autonomous),
Salem-16.

Abstract: **The excessive growth of internet and the rise of cyber security threats have led to easily vulnerable network environment. Cyber security threat ranges from simple password hacking to advanced persistent threats. To provide a secured network environment, firewall, intrusion detection system, intrusion prevention system etc., can be utilized. Intrusion Detection System (IDS) is used to monitor packets, network traffic, malicious activity etc., IDS can be used to scan ports and it can be used to generate alerts when a suspicious activity occurs in the network. This paper deals with different types of intrusion detection system attacks, datasets and tools. An overview of machine learning algorithms was performed. Machine learning life cycle from data acquisition, exploratory data analysis, data preprocessing, model generation and visualization was discussed.**

Keywords: Intrusion, machine learning, survey, network traffic, cyber security.

I. INTRODUCTION

The evolution of internet has created an impact in global communication networks. Private organizations, enterprises, government agencies depend on E-commerce and digital transactions. Personal information, financial information and health care information are digitized. Cyber criminals, hackers, network intruders can attack the network to steal the confidential information. Cyber security techniques such as firewall, antivirus, intrusion detection system (IDS), intrusion prevention system (IPS), Security Information and Event Management (SIEM) are essential to secure the environment from both internal and external attacks.

Attacks which threatens Cyber security are malware, worm, Trojan, spyware, adware, ransomware, rootkit, backdoor, bot, exploit, botnet, scanning, sniffing, keylogger, spam, login attack, account takeover, phishing, social engineering, advance persistent threats, zero-day vulnerability etc., Intrusion is the source for many attacks. Some common intrusion attacks are Remote to User (R2U), probe, User to Root (U2R) and Denial of Service (DoS).

Intrusion Detection System (IDS) is used to detect the intruders and attacks in the network. Host based IDS, Network based IDS and Hybrid IDS are the different types of Intrusion Detection System. Host based IDS detects the attacks from a single system. Network based IDS is used to detect attacks such as botnet in the network. Hybrid IDS is the combination of both Host and Network IDS. Intrusion Detection System must produce high detection rate. It must reduce the false alarm rate. A good intrusion detection system must detect the known and unknown attacks. Security analyst can handle the detected attacks. Alerts based prevention system can be generated. Security analytics can be performed through machine learning.

II. LITERATURE REVIEW

A review of NSL-KDD dataset which is an intrusion dataset along with various machine learning algorithms was performed.

El Mostapha Chakir [1] proposed Information gain technique which selects the features of NSL-KDD dataset. Machine learning algorithm support vector machine (SVM) based on Radial-basis kernal function (RBF) was used. Particle swarm optimization (PSO) algorithm was used to optimize features selected by SVM

Hee-su Chae [2] proposed Correlation-based Feature Selection (CFS), Information Gain (IG) ,Gain Ratio(GR) were used for feature selection. Only 22 features were selected for further processing. Decision tree algorithm was implemented to classify the NSL-KDD dataset. 99.794% accuracy was attained through this model.

Jamal Hussain [3] proposed hybrid classification based on Support Vector Machine (SVM) and ArtificialNeural Network (ANN) in NSL-KDD dataset. Anomaly such as intrusion was detected using SVM. Misuse detection such as Denial of Service, probing, Remote to User, User to Root was detected using ANN.

Lifang Zi [4] proposed a adaptive clustering method to find Distributed Denial of Service Attacks (DDoS) in 2000 DARPA Intrusion Detection Scenario Specific Data Set. A modified Global K-means algorithm (MGKM) was used to cluster DDoS attacks.

Dr. Saurabh Mukherjee [5] proposed Correlation-based Feature Selection, Information Gain and Gain Ratio for feature selection. Feature Vitality Based Reduction Method was used to find reduced input features naive bayes classifer was to classify the intrusion attacks from NSL-KDD dataset.

From the literature review it is evident that the researchers had focused more on feature selection and reduction. Hybrid algorithms and optimization techniques had been adapted to create a optimized model.

III. MACHINE LEARNING

Machine learning is the core of artificial intelligence. Through machine learning algorithms complex problems of artificial intelligence can be solved. machine learning algorithms are categorized into three main groups. They are supervised, semi-supervised and unsupervised algorithms. Datasets with labels can be easily processed by supervised machine learning algorithms. Classification algorithms like logistic regression, decision tree, decision forests, support vector machine navie bayes, k-nearest neighbors, are some of the examples for supervised machine learning algorithm. Unsupervised algorithms works well with unlabeled datasets. Clustering algorithms like k-means, hierarchical clustering, density based spatial clustering of applications with noise (DBSCAN), are some of the examples for unsupervised algorithm. Semi-supervised algorithms can be applied for small amount of labeled and large amount of unlabeled datasets. Machine learning algorithms can be implemented through languages like Python or R. Fig.1 shows the generic structure of machine learning.

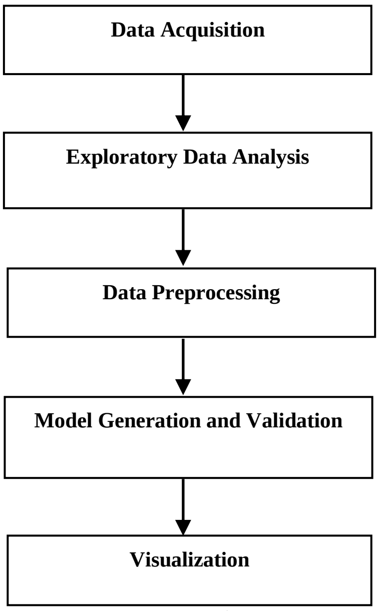

Fig.1 Generic structure of Machine Learning

IV. DATA ACQUISITION

Limited benchmark intrusion datasets are available. Some of them are DARPA1998, KDD99, NSL-KDD, UNSW-NB15, PREDICT, CAIDA, Internet Traffic Archive, DEFCON, ADFA Intrusion detection dataset, KYOTO, ISCX 2012, ICS Attack, IDS-2017, CTU-13. Creating new dataset is time consuming and it is a tedious process for the researchers.

Real time datasets can be acquired with the help of tools which are commercially available or open source. Tools that are commercially available are TippingPoint, Sourcefire, and Fortinet . Open source tools like Bro-IDS, Snort, Suricata, OSSEC, Samhain Labs, OpenDLP can be used for academic and research purposes. Simulation tools like ns3, netsim, EXata+Cyber package suite can be used to simulate the network attacks and it can be collected for further analysis.

V. EXPLORATORY DATA ANALYSIS

Exploratory Data analysis is an essential step for developing a machine learning model. EDA is prerequisite to figure out the dataset, to find summary of the data using statistics, to perform visualization for better interpretation, to find the outliers which are not necessary for the model, to find the correlation between the features and to provide hypothesis. Basic statistics functions such as Count, Mean, Standard Devaition, Minimum Value, Maximum Value, 25th Percentile, 50th Percentile (Median), 75th Percentile provides the summary of the dataset. Outliers and the correlation of the features can be found using visualizations such as Box plot, Histogram, Scatter plot etc.,

VI. DATA PREPROCESSING

Data preprocessing is the process of cleaning the dataset. Dataset with missing values, duplicate data, and mixed data will be inconsistent for processing. Several preprocessing techniques are available to handle this inconsistent data. Missing values can be NAN (Not A Number) values or blank values. Missing values can be dropped if it is not necessary for prediction. Dropping missing values may tend to bring changes in the accuracy of the model. Another approach of handling missing values is it can be imputed with the mean value in case of numerical attribute and with most frequent value in case of categorical attribute.

Duplicate or redundant data can be dropped since it increases the computation. Normalization can be done for the attributes so that it can be equal ranges. Visualization of data will be much better after normalization. Transformation is the process of converting the data in the required form. Categorical values are hard to analyse. Hence it could be transformed to numerical values by one hot encoding. Aggregation of the attributes is also a part of data preprocessing. Aggregating the attributes reduces the dimensionality of the dataset. Dimensionality of the dataset plays major role in analysis. Due to Curse of dimensionality storage space and processing time of dataset increases tragically. Visualization will also be difficult for high dimensional data. To overcome this, dimensionality reduction can be done. Dimensionality reduction algorithms like Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE) can be used to reduce the dimension of the dataset by feature selection.

VII. MODEL GENERATION AND VALIDATION

Machine learning algorithms can be categorized into supervised and unsupervised algorithms. If the attributes of the dataset has label, supervised algorithm can be applied, Unlabeled datasets can be processed using unsupervised algorithm. After the profound exploratory data analysis of the selected dataset, a suitable machine learning algorithm can be applied to generate the model.

Generally dataset is divided into three parts for processing. 60% of data is used for training, 20% is used for validation and remaining 20% of data is used for testing.

Cross validation [6] is the process of selecting the shuffled samples from dataset inorder to evaluate the performance of the applied machine learning model. K- Folds Cross Validation technique is widely used for cross validating the samples. The number of groups to be formed is decided by the parameter K. If the parameter K is assigned 4, then four groups will be created from the dataset. From each group certain amount of samples are trained and remaining will be tested. Each group will be assigned a score. Finally the average of all collected scores will be the metrics used to evaluate the performance of the model.

VIII. VISUALIZATION

"A Picture is worth a thousand words" by Frederick R. Barnard. Visualization is the important phase in machine learning life cycle. Datasets may be either large or small. To explore the unknown dataset, visualization is beneficial. Mere statistical methods and calculations will not provide the clear picture of the dataset. Correlation and better understanding of the attributes can be accomplished through visualization.

Numerous visualization libraries are available in python. Matplotlib, pandas visualization, seaborn, ggplot, plotly are some of them. Scatter plot, line chart, histogram, bar chart, box plot, heatmap, faceting, pairplot are some of the interactive plots available in python libraries.

IX. PERFORMANCE METRICS

Performance metrics[7] is used to identify the performance of the various machine learning models. Following terms and metrics are used to assess the performance of the model developed.

True Positive (TP) - True positive denotes the samples which are correctly classified as positive.

True Negative (TN) - True negative denotes the samples which are correctly classified as negative.

False Positive (FP) - False positive denotes the samples which are wrongly classified as positive

False Negative (FN) - False negative denotes the samples which are wrongly classified as negative

$$\text{Accuracy}= \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision}= \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall}= \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 Score} = \frac{2*precision*recall}{precision+recall} \quad (4)$$

$$\text{Sensitivity}=\frac{TP}{TP+FN} \quad (5)$$

Sensitivity denotes the true positive rate also known as recall. Prediction of actual true positive samples correctly is sensitivity.

$$\text{Specificity}=\frac{TN}{TN+FP} \quad (6)$$

Specificity denotes the false Positive rate which predicts the actual false positive samples correctly.

Mean Absolute Error (MAE) [8] is the average of all absolute errors. The difference between the actual values and the predicted values give absolute error.

$$\text{MAE}= \text{actual data} – \text{predicted data} \quad (7)$$

Mean Squared Error (MSE) is the average of the square of the difference between actual values and predicted values.

$$\text{MSE}= \frac{1}{N}\sum_{i=1}^{n}(actual\ data - predicted\ data)^2 \quad (8)$$

X. OPTIMIZATION

Once the model is developed, the performance is tested. To increase the performance of the model with less execution time and storage, optimization of the model is essential. Optimization algorithm such as Stochastic Gradient Descent can be adapted to optimize the performance of the model.

XI.CONCLUSION AND FUTURE DIRECTION

To predict and overcome the attacks there is necessity of good intrusion detection system which produces high detection rate and less false alarm rate. Machine Learning approach to the intrusion dection system plays a vital role for future secured network. Further intrusion detection system can be extended to deep learning.

REFERENCES

[1] El mostapha chakir, mohamed moughit, Youness idrissi khamlichi, "An effective intrusion detection model based on svm with feature selection and parameters optimization", Journal of Theoretical and Applied Information Technology 30th June 2018. Vol.96. No 12

[2] Hee-su Chae, Byung-oh Jo , Sang-Hyun Choi , Twae-kyung Park, "Feature Selection for Intrusion Detection using NSL-KDD", Recent Advances in Computer Science, ISBN: 978-960-474-354-4

[3] Jamal Hussain, Samuel Lalmuanawma, Lalrinfela Chhakchhuak, "A two-stage hybrid classification technique for network intrusion detection system", International Journal of Computational Intelligence Systems, Vol. 9, No. 5 (2016) 863-875

[4] Lifang Zi , John Yearwood , Xin-Wen Wu, "Adaptive Clustering with Feature Ranking for DDoS Attacks Detection", 2010 Fourth International Conference on Network and System Security

[5] Dr. Saurabh Mukherjeea, Neelam Sharma, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction", Elsevier Ltd, Procedia Technology 4 (2012) 119 – 128

[6] https://machinelearningmastery.com/k-fold-cross-validation/

[7] https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

[8] https://www.studytonight.com/post/what-is-mean-squared-error-mean-absolute-error-root-mean-squared-error-and-r-squared

AUTHORS PROFILE

P. Roshni Mol is a Ph.D Research Scholar in Department of Computer Science, Sri Sarada College for Women(Autonomous), Salem, Tamilnadu, India. She has completed her Post graduate course and M.Phil at Sri Sarada College for Women(Autonomous), Salem. Her Area of interest includes Image processing, Network security and Cryptography.

Dr. C. Immaculate Mary has completed her M.C.A in St.Joseph's College , Trichy and did her M.Phil and Ph.D in Mother Teresa Womens University, Kodaikanal. She is working as Associate Professor and Head in the Department of Computer Science, Sri Sarada College for Women(Autonomous) , Salem, Tamilnadu, India. She has twenty nine years of experience in the field of Computer Science. She has been specialized in this area of Data Mining and Image Processing. She has attended many National, International conferences and Workshop and published research papers in several International Journals.