Machine Learning Techniques for Detecting Network Anomalies

M. Ravi Author[1], V. Raja Shekhar Author[2], K. Ruthvik Author[3], and T. Abhigna Author[4]
[1,2,3,4] *Member, Dept ofInformation Technology JB Institute of Engineering and Technology*

Abstract: **As technology progresses, the number of internet users grows every day. Mobile technology has made communication more convenient, yet it is still insecure. Similarly, a number of internet-connected and data-sharing devices have been developed. These are Internet of Things (IoT) devices. The Internet of Things (IoT) is a rapidly growing industry with numerous uses. The number of people using IoT devices is expanding, as is the number of new goods. Users appreciate the functionality of IoT devices, but many are utterly oblivious of the security concerns that lurk beneath. As a result, improving data transfer and user privacy security is critical. Machine learning techniques are becoming more essential in identifying network breaches (or assaults), allowing network administrators to take preventative steps. To improve the performance of an Intrusion Detection System, we recommend using machine learning techniques.**

***Keywords*— Intrusion Detection System, Machine Learning, security, Anomaly detection, Misuse detection, Classifiers, EDA.**

1.INTRODUCTION

A network anomaly is a sudden and brief departure from the network's regular operation. Some anomalies are purposefully generated by malevolent intruders, such as a denial-of-service attack in an IP network, while others are completely unintentional, such as an overpass collapsing in a busy road network. A prompt response, such as dispatching an ambulance after a car accident or sounding an alert if a surveillance network detects an intruder, requires quick detection. Data is collected at a high pace by network monitoring equipment. As a result, creating an effective anomaly detection system necessitates extracting important data from a large volume of noisy, high-dimensional data. The emergence of harmful software (malware) creates a significant issue for intrusion detection system designers (IDS). Malicious attacks have evolved, and the most difficult difficulty is identifying unknown and obfuscated malware, since malware creators employ various evasion tactics for information concealment in order to avoid detection by an IDS. Furthermore, security concerns such as zero-day attacks meant to target internet users have increased. Therefore, computer security has become essential as the use of information technology has become part of our daily lives. As a result, various countries such as Australia and the US have been significantly impacted by the zero-day attacks. According to the 2017 Symantec Internet Security Threat Report, more than three billion zero-day attacks were reported in 2016, and the volume and intensity of the zero-day attacks were substantially greater than previously (Symantec, 2017). As highlighted in the Data Breach Statistics in 2017, approximately nine billion data records were lost or stolen by hackers since 2013. A Symantec report found that the number of security breach incidents is on the rise. In the past, cyber criminals primarily focused on bank customers, robbing bank accounts or stealing credit cards (Symantec, 2017). However, the new generation of malware has become more ambitious and is targeting the banks themselves, sometimes trying to take millions of dollars in on eattack (Symantec, 2017). For that reason, the detection of zero-day attacks has become the highest priority.

Machine learning has been utilized to improve intrusion detection during the previous few decades, and there is currently a need for an up-to-date, comprehensive taxonomy and assessment of this recent work. There are numerous studies that use the KDD-Cup 99 or DARPA 1999 datasets to validate the development of IDSs, but no clear response to the question of which data mining approaches are more effective. Second, although being a crucial component for the success of 'on-line' IDSs, the time spent developing IDS is not taken into account in the evaluation of some IDS strategies.

In this research, we suggest a few machine learning algorithms for detecting heuristic and anomaly-based assaults, which are expected to perform well when used with Network Intrusion Detection Systems.

2. ARCHITECTURE

Architecture diagram explains the design of the project. It acts as a Blue Print for the project. It gives a brief idea of the project overview.

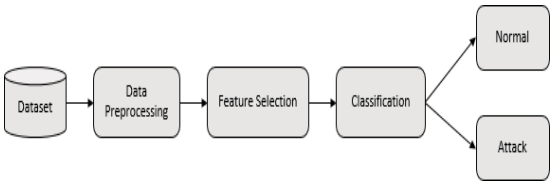

Fig1:Architecture of Anomaly Detection System

Based on the defined goal (s) we have used three of the supervised modeling techniques. Such as

- Decision Trees
- Random Forest
- Support Vector Machine

To develop a model, a Machine Learning algorithm is trained using a training data set. The ML algorithm makes a prediction based on the model when new input data is introduced.

The accuracy of the forecast is assessed, and if it is acceptable, the Machine Learning method is used. If the accuracy isn't good enough, the Machine Learning algorithm is retrained with a new batch of training data.

Building a Predictive model that can be used to discover a solution to a Problem Statement is part of the Machine Learning process. Assume you've been given an issue to address using Machine Learning to better understand the process. The below steps are followed in a Machine Learning process:

Step1: Define the goal of the problem statement. We must now establish precisely what needs to be projected. The purpose of this scenario is to evaluate the possibility of rain using weather parameters. It's also a good idea to make mental notes on what kind of data you'll need to solve the problem and how you'll get there at this point.

Step2: Data Collection

- What type of data is required to tackle this problem?
- Is the data readily available?
- How do I obtain the information?

After you've figured out what kind of data you'll need, you'll need to figure out how to get it. Data can be collected manually or by web scraping. If you're a newbie trying to learn Machine Learning, though, you won't have to worry about acquiring data. There are plenty of data resources on the internet; simply download the data set and get started.

Returning to the issue at hand, the data required for weather forecasting comprises factors such as humidity, temperature, pressure, location, whether you reside in a hill station, and so on. Such information must be gathered and preserved in order to be analyzed.

Step3: Data Preparation

Almost never is the information you acquire in the appropriate format. Missing values, redundant variables, duplicate values, and other errors will be found across the data set. It's critical to eliminate such inconsistencies because they can lead to inaccurate calculations and predictions. As a result, you search the data set for discrepancies at this point and correct them immediately.

Step4: Exploratory Data Analysis

The initial stage of Machine Learning is Exploratory Data Analysis, or EDA. The purpose of data exploration is to discover patterns and trends in the data. At this point, all of the valuable insights have been gleaned, and the correlations between the variables have been identified.

For example, we discovered that many characteristics had a strong correlation between some variables in the dataset utilized in this work, KDDcup99, and that this linked data can be deleted for better data processing. Such connections must be understood and mapped at this point.

Step5:Building a Machine Learning Model

The Machine Learning Model is built using all of the insights and patterns discovered during Data Exploration. The data set is always separated into two parts, training data and testing data, at this stage. The model will be built and analyzed using the training

data. The model's logic is based on the Machine Learning Algorithm that is currently in use.
The type of problem you're trying to answer, the data set, and the problem's complexity all influence whatever algorithm you use. In the next sections, we'll go over the various types of problems that Machine Learning can address.

Step6:Model Evaluation& Optimization
Once the model has been developed using the training data set, it's time to put it to the test. The testing data set is used to determine the model's effectiveness and ability to accurately anticipate outcomes. After the accuracy has been calculated, any additional model enhancements can be made. You can use techniques like parameter tweaking and cross-validation to improve the model's performance.

Step7: Predictions
Once the model is evaluated and improved, it is finally used to make predictions. The final output can be a Categorical variable (eg. True or False) or it can be a Continuous Quantity (eg. the predicted value of a stock).
In our case, for predicting the network anomaly packet, the output will be a categorical variable (Normal or Attack).

3. ALGORITHMS

Random Forest:
As the name implies, a random forest is made up of a huge number of individual decision trees that work together as an ensemble. Each tree in the random forest produces a class prediction, and the class with the most votes becomes the prediction of our model.
Random Forest Model Visualization Predicting the Future. Any of the individual constituent models will outperform a large number of reasonably uncorrelated models (trees) working as a committee.

Decision Tree:
A decision tree is a supervised learning strategy with a pre-defined target variable that is widely used in classification problems. Both categorical and continuous input and output data are supported by this tree.
The sample (population) is divided into two or more sub-populations, with the most significant splitter or differentiator selected by the input variables. The ultimate goal is to create a prediction model that can take data from a sample (the branches) and estimate the sample's target value accurately (the leaves).

Support Vector Machine:
A support vector machine is a collection of supervised learning algorithms that use hyperplane graphing to evaluate new, unlabeled data. In the most common scenario, incoming data is separated into two groups. SVM models read each data point as a p-dimensional vector, and the computer attempts to create a linear classifier by fitting the data point into a hyperplane (p-1 dimension).
Each input is turned into a point in n-dimensional space (n being the number of features), with each feature's value defined as the value of a single hyperplane coordinate. To classify the two classes, the hyperplane that most clearly separates them is used.

4. MODULES

A. Data Collection:
The dataset used is KDDcup99 developed by DARPA consisting of nearly 5lakhs of of records with almost 41 network features of packet. It consists of features like duration, service, protocol type, flag, logged in, source and many more.

B. Data Preprocessing:
To begin, we must first preprocess the data by deleting all null values and extraneous columns.
Certain aspects of the generated heat map demonstrate a correlation between them. As part of the data cleaning process, this highly connected data is removed. All categorical attributes, such as protocol type, service, and flag, are mapped to number values, resulting in a dataset with all feature values in numerical form.

C. Prepare Data for Training and Testing:
The right response, also known as a target or target attribute, must be included in the training data. The learning algorithm searches the training data for patterns that connect the input data attributes to the goal (the result you want to predict), and then generates an ML model that captures these patterns.

i. Trainingdatasetconsists70%or80%data.
ii. Testing dataset consist of30%or20%data.

D. Model Evaluation:

After the model is trained it is tested against with the other half the data which is taken to evaluate the model behaviour. The three main metrics used to evaluate a classification model are accuracy, precision, and recall.

5. CONCLUSION

Machine learning has introduced new methods for intrusion detection systems, with researchers and academics developing intrusion detection system models that use a wide range of classifications. This report looked at three machine learning classifiers that are commonly used in intrusion detection systems. The Support Vector Machine model outperformed the other models (Random Forest, Decision Tree) used in the papers studied in terms of predicted accuracy and detection rate.

6. FUTURE ENHANCEMENT

Future research will necessitate more thought in order to improve the performance of intrusion detection systems. Hybrid and ensemble machine learning classification techniques should be used more frequently because single machine learning classifiers perform better when combined in a specific way. Some classifiers outperform others on specific datasets. More models must be developed in the future to perform well on multiple datasets. When the best fit model is used in conjunction with the Intrusion Detection System, it improves the Intrusion Detection System's performance.

RE1FERENCES

[1] W. –C. Lin, Shih-Wen K. Chih-Fong T. (2015). Intrusion detection system based on combining cluster centers and nearest neighbors. Knowledge-Based Systems 78 (pp.13-21). Elsevier.

[2] A.S.A. Aziz. (2016). Comparison of classification techniques applied for network intrusion detection and classification. Journal of AppliedLogic24.Elsevier,109-118.

[3] Nabila Farnaaz and M.A Jabbar. (2016). Random Forest Modeling for Network Intrusion Detection System. International Multi-conference on information processing (IMCIP) 12 (pp.213-217). Elsevier.

[4] Kayvan A. Saadiah Y. Amirali R. and Hazyanti S. (2016). Anomaly Detection Based on Profile Signature in Network using Machine Learning Techniques. IEEETENSYMP.(pp.71-76).IEEE.

[5] Bobba Brao and Kailasam Swathi. (2017). Fast KNN Classifiers forNetwork Intrusion Detection System. Indian Journal of Science and Technology.10(14).Researchgate.(1-10).

[6] Chie-Hong L. Yann-Yean S. Yu-Chun Lin and Shie-Jue L.(2017).Machine Learning Based Network Intrusion Detection. 2nd IEEE International Conference on Computational Intelligence and Applications. (pp.79-83). IEEE.

[7] Deyban P. Miguel A. A, David P. A, and Eugenio S. (2017). Intrusion detection in computer networks using hybrid machine learning techniques. XLIII Latin American Computer Conference (CLEI). (pp.1-10).IEEE

[8] Ponthapalli R. et al. (2020). Implementation of Machine Learning Algorithms for Detection of Network Intrusion. International Journal of Computer Science Trends and Technology (IJCST).(163-169).

[9] Rajagopal S., Poornima P. K. and Katiganere S. H. (2020). A Stacking Ensemble for Network Intrusion Detection using Heterogeneous Datasets. Journal of Security and Communication Networks. Hindawi. (1-9).

[10]]FarahN.H etal.(2015). Application of Machine Learning Approaches in Intrusions Detection Systems. International Journal of Advanced Research in Artificial Intelligence. IJARAI.(9-18).