ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022

Analysis of Attributed Networks for Anomaly Detection Self-Supervised Learning in Contrast

Mr.Dilipkumar E
Department of MCA
Dhanalakshmi Srinivasan College of Engineering and Technology

Ms.Thilagavathy D
Department of MCA
Dhanalakshmi Srinivasan College of Engineering and Technology

Abstract: The state of the cyberspace portends uncertainty for the future Internet and its accelerated number of users. New paradigms add more concerns with big data collected through device sensors divulging large amounts of information, which can be used for targeted attacks. In this paper, we model cyber-attack prediction as a classification problem, Networking sectors have to predict the type of Network attack from given dataset using machine learning techniques. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments etc. A comparative study between machine learning algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the type cyber Attacks.

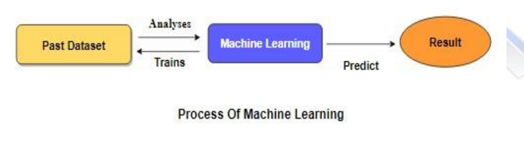
I. INTRODUCTION

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels.

At a high level, these different algorithms can be classified into two groups based on the way they "learn" about data to make predictions: supervised and unsupervised learning. Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories.

Classification predictive modeling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation.

This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.



Process of Machine Learning

Supervised Machine Learning is the majority of practical machine learning uses supervised learning. Supervised learning is where have input variables (X) and an output variable (y) and use an algorithm to learn the mapping function from the input to the output is y = f(X). The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (y) for that data.

DATASET

This Dataset contains 3000 records of features. It is classified into 4 classes.

- DOS Attack
- R2L Attack
- U2R Attack
- Probe Attack

EXISTING SYSTEM

- ➢ They proposed first to create a contrastive self-supervised learning to the anomaly detection problem of attributed networks.
- ➢ Their model captures the relationship between each node and its neighbouring structure and uses an anomaly-related objective to train the contrastive learning model.
- ➢ The training phase and the inference phase. In the training phase, the contrastive learning model is trained with sampled instance pairs in an unsupervised fashion.
- ➢ After that the anomaly score for each node is obtained in the inference phase.

DISADVANTAGES OF EXISTING SYSTEM

- The performance is not good and its get complicated for other networks.
- The performance metrics like recall F1 score and comparison of machine learning algorithm is not done.

PROPOSED SYSTEM

The proposed model is to build a machine learning model for anomaly detection. Anomaly detection is an important technique for recognizing fraud activities, suspicious activities, network intrusion, and other abnormal events that may have great significance but are difficult to detect.

Then the visualisation of the data is done to insights of the data .The model is build based on the previous dataset where the algorithm learn data and get trained different algorithms are used for better comparisons. The performance metrics are calculated and compared.

ADVANTAGES OF PROPOSED SYSTEM

- The anomaly detection can be automated process using the machine learning.
- Performance metric are compared in order to get better model.

MODULE 1

VARIABLE IDENTIFICATION PROCESS/DATA VALIDATION PROCESS

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

DATA VALIDATION/ CLEANING/ PREPARING PROCESS

To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models.

ADVANTAGES OF TRAIN/TEST SPLIT

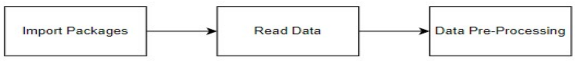
- This runs K times faster than Leave One Out cross-validation because K-fold cross-validation repeats the train/test split K-times.
- Simpler to examine the detailed results of the testing process.
- Advantages of cross-validation:
- More accurate estimate of out-of-sample accuracy.
- More “efficient” use of data as every observation is used for both training and testing.

DATA PRE-PROCESSING

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. applied model in Machine Learning method of the data has to be in a proper manner.

Some specified Machine Learning model needs information in a specified format Random Forest algorithm does not support null values.

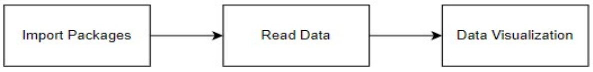
MODULE DIAGRAM



Data pre-processing diagram

GIVEN INPUT EXPECTED OUTPUT
input : data
output : removing noisy data

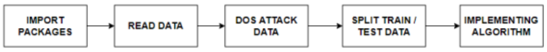
MODULE DIAGRAM



Data visualization diagram
GIVEN INPUT EXPECTED OUTPUT
input : data
output : visualized data
MODULE 2

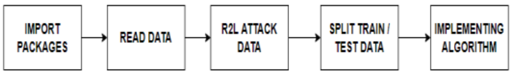
A DoS or DDoS attack is analogous to a group of people crowding the entry door of a shop, making it hard for legitimate customers to enter, disrupting trade.

MODULE DIAGRAM



GIVEN INPUT EXPECTED OUTPUT
input : data
output : getting accuracy
MODULE 3

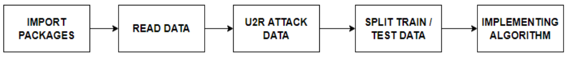
Now-a-days, it is very important to maintain a high level security to ensure safe and trusted communication of information between various organizations.
MODULE DIAGRAM



R2L attack process
GIVEN INPUT EXPECTED OUTPUT
input : data
output : getting accuracy

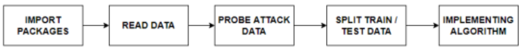
MODULE 4

Remote to local attack (r2l) has been widely known to be launched by an attacker to gain unauthorized access to a victim machine in the entire network.
MODULE DIAGRAM:

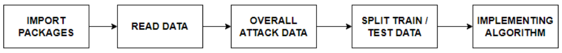


U2R attack process
GIVEN INPUT EXPECTED OUTPUT
input : data
output : getting accuracy
MODULE 5

Probing attacks are an invasive method for bypassing security measures by observing the physical silicon implementation of a chip. As an invasive attack, one directly accesses the internal wires and connections of a targeted device and extracts sensitive information.
MODULE DIAGRAM:



Probe attack process
GIVEN INPUT EXPECTED OUTPUT
input : data
output : getting accuracy
MODULE 6
Complex attacks can be divided into exploration and exploitation phases. First, attackers will attempt to collect information on the organization they intend to attackA phishing attack is usually carried out by sending an email purporting to come from a trusted source and tricking its receiver to click on a URL that results in installing malware on the user's system with a spam email), as well as by the characteristics of URLs included in the message.
MODULE DIAGRAM:

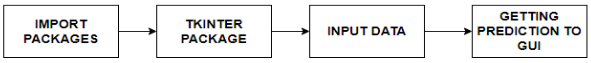


Overall attack process
GIVEN INPUT EXPECTED OUTPUT
input : data
output : getting accuracy
MODULE 7

The graphical user interface (GUI) is a form of user interface that allows users to interact with electronic devices through graphical icons and audio indicator such as primary notation, instead of text-based user interfaces, typed command labels or text navigation.

Graphical user interface (GUI) wrappers find a way around the command-line interface versions (CLI) of (typically) Linux and Unix-like software applications and their text-based user interfaces or typed command labels.

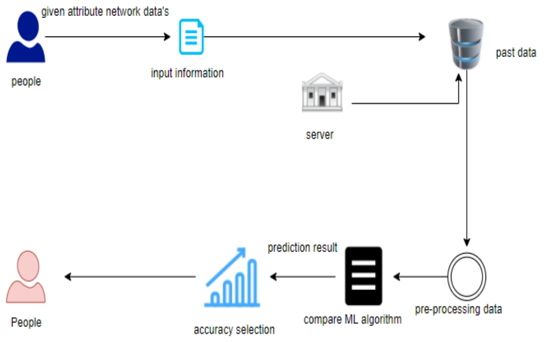
MODULE DIAGRAM:



Prediction to GUI
GIVEN INPUT EXPECTED OUTPUT
input : data values

output : predicting output

SYSTEM ARCHITECTURE



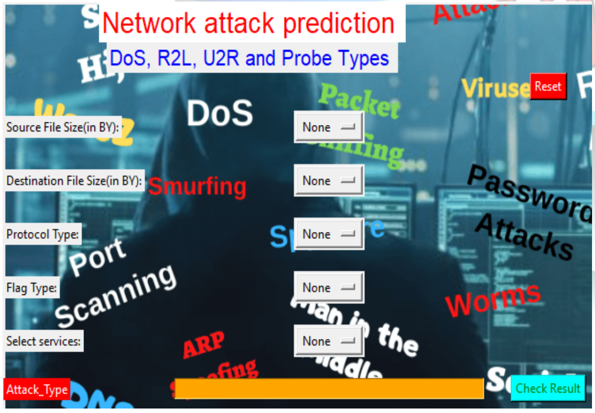
Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs.

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields.

Acceptance testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

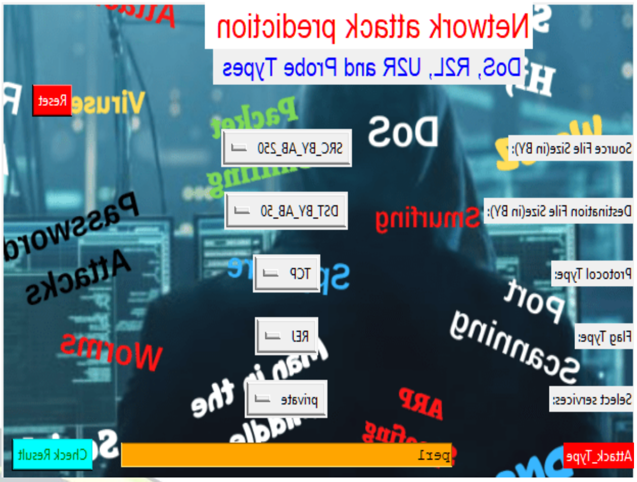
SAMPLE SCREENSHOTS



Network attack prediction
DoS, R2L, U2R and Probe Types

Source File Size(in BY): None
Destination File Size(in BY): None
Protocol Type: None
Flag Type: None
Select services: None
Attack_Type

Input screen



Network attack prediction
DoS, R2L, U2R and Probe Types

Source File Size(in BY): SRC_BY_AB_250
Destination File Size(in BY): DST_BY_AB_50
Protocol Type: TCP
Flag Type: REJ
Select services: private
Attack_Type: perl

Output screen

CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out by comparing each algorithm with type of all network attacks for future prediction results by finding best connections. This brings some of the following insights about diagnose the network attack of each new connection.

REFERENCES

[1] Z. Liu, C. Chen, X. Yang, J. Zhou, X. Li, and L. Song, “Heterogeneous graph neural networks for malicious account detection,” in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 2077–2085.

[2] L. Tang and H. Liu, “Relational learning via latent social dimensions,” in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2009, pp. 817–826.

[3] Y. Zhang *et al.*, “Your style your identity: Leveraging writing and photography styles for drug trafficker identifiidentification in darknet markets over attributed heterogeneous information network,” in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 3448–3454.

[4] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, “Graph convolutional neural networks for Web-scale recommender systems,” in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 974–983.

[5] W. Fan *et al.*, “Graph neural networks for social recommendation,” in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 417–426.

[6] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Proc. Int. Conf. Learn. Represent.*, 2017,