Mid-Term Report

CS5350 Machine Learning Exploratory Project

TYLER JONES AND VIJAY BAJRACHARYA

CONTENTS

1 INTRODUCTION

Traditional cybersecurity stacks consist of multiple layers of protection. Firewalls, endpoint protection, secure gateways, and authentication protocols are a few of the many tools used to protect a company's network. Despite these efforts, the ever-evolving cybersecurity landscape continues to plague even the largest of companies in the world. To make monitoring and detection tasks easier, many businesses have started adopting intelligent and autonomous solutions into their security stacks. For our project, we want to perform a comparative analysis of various machine learning algorithms on public cybersecurity data. Moreoever, we want to assess their viability as an alternative approach to cyber-threat management.

2 MID-TERM PROGRESS

2.1 Research

Heavily inspired by the paper on cybersecurity knowledge bases by Li et al (2020), our exploratory project is an appendage to a larger capstone project involving the autonomous detection and contextualization of network threats using knowledge graphs and neural networks. To support the results that we will generate using the neural network, we wanted to first form a baseline for comparison using traditional machine learning algorithms. Creating this baseline is the main goal for this project.
In order to detect network intrusions, we had to familiarize ourselves with the most common types of network attacks. Since an entire network architecture has too many components to cover in one project, we decided to focus solely on network packet data for this exploratory project. For packet data, we searched for pre-existing public datasets that had specific attack types associated with network traffic. This is where we discovered the UNSW_NB15 dataset, the details for which were published in five different papers that

are listed in the references section of this report. Following the collection of data, we created a list of all the machine learning algorithms that we would be using for our project. So far, our list of algorithms contains decision tree, random forest, support vector machine, naive bayes, and k-nearest neighbors.

2.2 Data Collection and Storage

All of the algorithms that will be used in this project will utilize the USNW_NB15 dataset. The dataset consists of raw network data packets that were generated in the Cyber Range Lab of the Australian Centre for Cyber Security based on real and synthesized attack patterns. It contains 100 gigabytes of network traffic but we will only be using a small subset of this data that was pre-sampled into testing and training data. We will be training our models on 82,332 labelled examples with 49 different features such as protocol, service, time to live, source to destination bytes etc. Moreover, the dataset is divided into 10 different attack families - normal, DoS, fuzzer, backdoor, exploit, renaissance, analysis, generic, shellcode and worms.

The training and testing sets are stored in the form of csv files. Due to the large size of these files, we had to compress the data using gzip. Every time an analysis is run, our program will decompress and extract the csv files into a local folder and use the relative paths to train the model and calculate classification errors.

2.3 Data Preprocessing

The packet data contains a mixture of both numerical and categorical data. Our first goal was to unify the dataset by mapping each categorical feature to a binary vector. To achieve this, we created a pandas dataframe from our csv files and split the data into feature vectors and ground truth labels. Then, we selected all feature vectors that were categorical in nature and used the sk-learn one-hot encoder to provide a numerical representation for categorical features. The converted features were then inserted back into the training set.

2.4 Results

So far, we have been able to use a decision tree classifier to train a model on our data and predict labels for the testing set. With a maximum tree depth set to 20, we were able to generate the following results:

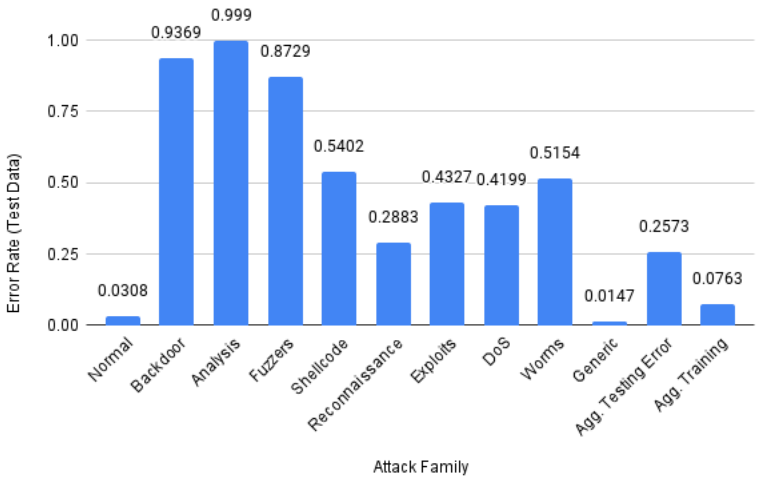

Figure 1: Comparison of classification error rates for each attack family

As we can see, while the aggregate testing error is 25% and the training error is 0.07%, the classifier performs very poorly on specific attack types. This might be a result of several factors including number of training examples, diversity of training examples, number of labels, the decision tree algorithm or the selected hyper-parameters.

3 PLANS FOR THE FUTURE

The next step in our project is to add more algorithms specifically, random forest, support vector machine, naive bayes, and k-nearest neighbors. We also want to optimize existing algorithms such as decision trees by tuning hyper parameters like tree depth. Furthermore, we are considering using other ensemble techniques such as adaboost in addition to random forests. Once we have a suite of algorithms, we want to generate detailed reports for each one and compare the accuracy and running times of these algorithms. Moreover, to support our findings, we are also going to produce graphs and visualizations for all of the classification results.

From the results and visualizations, we can determine whether or not an intelligent solution for threat detection on a network is a viable approach for companies to adopt. These results will also act as a comparative baseline for the neural network that we will be building over the course of the senior capstone project.

REFERENCES

1. Li, Kun, Huachun Zhou, Zhe Tu, and Bohao Feng. "CSKB: A Cyber Security Knowledge Base Based on Knowledge Graph." In International Conference on Security and Privacy in Digital Economy, pp. 100-113. Springer, Singapore, 2020.

2. Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." In 2015 military communications and information systems conference (MilCIS), pp. 1-6. IEEE, 2015.

3. Moustafa, Nour, and Jill Slay. "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set." Information Security Journal: A Global Perspective 25, no. 1-3 (2016): 18-31.

4. Moustafa, Nour, Jill Slay, and Gideon Creech. "Novel geometric area analysis technique for anomaly detection using trapezoidal area estimation on large-scale networks." IEEE Transactions on Big Data 5, no. 4 (2017): 481-494.

5. Moustafa, Nour, Gideon Creech, and Jill Slay. "Big data analytics for intrusion detection system: Statistical decision-making using finite dirichlet mixture models." In Data analytics and decision support for cybersecurity, pp. 127-156. Springer, Cham, 2017.

6. Sarhan, Mohanad, Siamak Layeghy, Nour Moustafa, and Marius Portmann. "Netflow datasets for machine learning-based network intrusion detection systems." arXiv preprint arXiv:2011.09144 (2020).