~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Artificial Intelligence for Cybersecurity Applications

CSI5388

Theme Selection

Project: Deliverable #1

Group Members
Alexandra Sklokin - 300010511
Pranav Pawar - 300333613
Anisha Swapnil Deachake - 300369581
Aryan Gulati - 300351365

Professor
Paula Branco

September 18, 2023

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

1 - Theme Selection

Our group has chosen to investigate the topic of feature selection for identifying cybersecurity attacks (Topic #3). In this project we would like to propose a new feature selection method, and compare it to classical feature selection algorithms commonly used for Machine Learning (ML). We will either use the domain of Distributed Denial of Services (DDoS) attack detection or malware attack detection, since many datasets are available within these domains [1,2,3,4,5,6]. We will prefer datasets with many features, so that we can adequately investigate the effects of the feature selection techniques (and so that each method does not select the same subset of features).

During the project description and literary review stages, we will investigate papers about ML for DDOS/malware attack detection. We will also investigate some classical and novel feature selection techniques, so that we can get inspiration for our own new method. At the moment, we have developed a few possibilities, but would like to do some more investigation. Some of our suggestions are:

(1) Intersection of baseline feature sets.
ex. Baselines: M1 = {A,B,C} M2 = {B,C,D} M3 = {A,B,E}
OurMethod = {B}
(2) Weighted average of baseline feature sets.
ex. OurMethod = {2/9A, 3/9B, 2/9C, 1/9D, 1/9E}
(3) Only features which occur in more than one baseline feature sets.
ex. OurMethod = {A,B,C}

During the implementation stage, we will begin by preprocessing the datasets (data cleaning, normalization, balancing, etc.). First, we will use the datasets with all features to train a baseline Neural Network (NN) model, which will likely overfit the data. Then we will use classical feature selection methods and retrain the same NN model, likely improving the generalization of the model. Some classical feature selection techniques are:

(1) ANOVA
(2) Pearson's Correlation Coefficient/ Chi-squared / Mutual Information
(2) SelectKBest
(3) XGBoost
(4) Forward and Backward Feature Selection

Finally, we will implement our novel feature selection method and compare it to the previous methods. To compare the methods, we will use performance evaluation methods, such as graphing the epochs versus the training loss/Mean-Squared-Error (MSE), validation loss/MSE, and training time. We will also use metrics such as the testing accuracy, precision, recall, and F1-Score to compare models. We hope that our novel method will outform the baseline feature selection techniques.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

2 - References

[1] https://www.kaggle.com/datasets/devendra416/ddos-datasets
[2] https://www.kaggle.com/datasets/yashwanthkumbam/apaddos-dataset
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

[3] https://www.unb.ca/cic/datasets/ddos-2019.html
[4] https://www.unb.ca/cic/datasets/index.html
[5] https://www.kaggle.com/code/maidaly/malware-detection-with-machine-learning
[6] https://github.com/topics/malware-dataset

~~~~~~
~~~~~~