Assignment 2

Alexander Hartl*
Maximilian Bachl*

1 INTRODUCTION

The detection of attacks in data networks is a fundamental task in data security. Due to the considerable amount of data which has to be analyzed the use of machine learning techniques for this purpose seems natural and is increasingly deployed.

For research the invention and assessment of techniques for network anomaly detection poses many challenges. Particularly, a considerable number of features can be extracted from network data which might be beneficial for anomaly detection. For the training of models usually datasets are used which have been generated artificially in a controlled test environment.

As a downside of this approach, it is unclear whether a machine learning model learn to classify based on characteristics that are inherent to the attacks which should be detected, or rather learns to classify based on patterns that were unintentionally created during dataset generation.

For a well-performing network anomaly detection technique it is therefore of utmost importance to study which patterns the technique looks at to distinguish attack traffic from normal traffic, and question if these explanations match with expert knowledge.

In this document, we redo parts of a recent paper which bases on the CIC-IDS-2017 dataset for evaluating the performance of several feature vectors and machine learning techniques for accurate anomaly detection. We use explainability methods for investigating if the decisions the anomaly detectors untertake are reasonable.

Furthermore, we add a backdoor to the trained model and show that attack detection can efficiently be bypassed if the attacker had the ability to modify training data. Finally, we apply the same explainability methods to the backdoored model and show how such attack attempts might be recognized before any harm is done.

2 MACHINE LEARNING APPROACHES FOR TRAFFIC CLASSIFICATION

2.1 Deep Learning

2.2 Random Forests

3 EXPLAINABILITY PLOTS

3.1 Partial Dependence Plots

Partial Dependence Plots (PDP) visualize dependence of a model's predictions by plotting the model's prediction for a modified dataset for which the feature's value has been fixed to a certain value and computing the an average over the dataset.

3.2 Individual Conditional Expectation

The averaging which is done for PD plots introduces the problem that the influence of a feature on individual samples is lost. In our case

3.3 Accumulated Local Effects

Due to feature dependence it is very likely that in the feature space areas exist which have a very low probability to occur. Since a model is trained with real, observed data, the training set therefore does not include samples for these areas, which causes the model's predictions to become indeterminate for these areas. This poses a problem when considering these predictions for computing PDPs.

In an attempt to overcome this problem, it is possible to only consider samples which are likely to occur for certain feature values, i.e. to consider the conditional distribution of remaining features, for computing explainability graphs. This is the concept for Accumulated Local Effects (ALE) plots.

When computing ALE plots, we experienced the problem of empty intervals. If there are intervals that do not contain any values, the usual definition which takes values between the interval's boundaries for estimating the conditional probability density for feature values.

For this reason, we modified this definition to instead use the closest 10 samples to the interval's center for estimating the distribution.

3.4 Interpretation

4 LOGISTIC REGRESSION AS SURROGATE MODEL

5 IMPLEMENTING A BACKDOOR

6 CONCLUSIONS

*Both authors contributed equally to this research.