Summary 21

Shaun Pritchard

Florida Atlantic University

CAP 6778

December 2, 2021

M. Khoshgoftaar

Detecting SSH and FTP Brute Force Attacks in Big Data

In this study, a simpler approach is presented for detecting brute-force attacks on the CSE-CIC-IDS2018 Big Data dataset that provides stronger classification results. Researchers state they have implemented a contribution to demonstrate that it is possible to train and test simple Decision Tree models with two independent variables to classify CSE-CIC-IDS2018 data with better results than previously reported research, where more complex Deep Learning models were used. Based on the results of this study, Decision Tree models trained on data with two independent variables performed similarly to Decision Tree models trained on data with a greater number of independent variables.

There are several significant questions about this data that need to be answered in this study. In essence, they sought to learn whether decision tree classification could classify CSE-CIC-IDS2018 FTP-Brute Force, SSH-Brute Force, and Combined-Brute Force attacks data effectively. In addition, they wanted to know the smallest number of features in the dataset they could use to achieve consistently strong classification performance.

CSE-CIC-IDS2018 was used to derive three datasets for SSH, FTP, and combined attack types. To rank features of a dataset, the researchers used three filter-based feature ranking techniques and four supervised learning-based feature ranking techniques. Put a maximum of 20 attributes into each ranking by truncating the 7 feature rankings.

Thereafter, they implemented select features shared by n, out of the 7 rankings, where n varies from 4 to 7. By applying the approach to three datasets for the case in which a feature needs to appear in six out of seven rankings, we get datasets with four features, which they refer to as six-agree datasets.

In the study, 7 out of 7 rankers rated the BWD_Packets feature among the top 20 most important features for combined and SSH data. As a result of these experiments, simple models have an AUC and AUPRC score greater than 0.99, and they can detect brute force attacks in CSE-CIC-IDS2018. The Decision Tree, trained and tested on only two features, the BWD_Packets and the minimum segment size forward of the CICCSE-IDS2018 Big Data, represents demonstrably and effectively the detection of SSH-Brute Force attacks and FTP-Brute Force attacks. It is interesting to note that there is only one published study related to this research.

Detecting Information Theft Attacks in the Bot-IoT Dataset

I found this study to be very interesting since it was based on the Internet of Things (IoT) devices dataset Bot-IoT, which was designed to train machine learning classifiers on network intrusion detection in IoT networks for security risk assessments. Researchers developed a predictive model for detecting information theft attacks using Bot-IoT. To detect information theft traffic, they implemented an innovative approach that uses eight classifiers and two performance metrics.

This research implemented ensemble learning classifiers that utilized four ensembles (CatBoost, LightGBM, XGBoost, and Random Forest) and four non-ensemble learners (Decision Tree, Logistic Regression, Naive Bayes, and Multilayer Perceptron). Area Under the Receiver Operating Characteristic Curve (AUC) and Area Under the Precision-Recall Curve (AUPRC) were used to evaluate the classifiers. To evaluate the most suitable classification method, they used cross-validation to train and test Bot-IoT instances based on information theft traffic.

In this study, the goal was to build a model capable of detecting information theft and data exfiltration attacks with high accuracy. Four hundred seventy-seven instances of normal behavior and 79 instances of information theft were evaluated. The Information Theft subcategory of Information Theft refers to the interception of data sent from input devices, and Data Exfiltration refers to the unauthorized transfer of data from a computer. There is a slight class imbalance in the subgroup of 556 instances when it comes to Normal and Information Theft traffic.

According to their results, the eight classifiers trained and tested via cross-validation are generally considered reliable. The ensemble classifiers, like CatBoost, LightGBM, and XGBoost, are the most effective models for detecting IoT information theft traffic based on our AUC and AUPRC results.