Final Project- M1 - ST2ML2 - Group DAI - Machine Learning II

Machine Learning for DDoS attacks detection

The CICDDoS2019 dataset, created by the *Canadian Institute for Cybersecurity (CIC)*, is a collection of network traffic data that aims to facilitate research and analysis related to Distributed Denial of Service (DDoS) attacks. It provides a realistic representation of various types of DDoS attacks and can be used to develop and evaluate intrusion detection systems, anomaly detection algorithms, and other cybersecurity mechanisms.

The dataset consists of network traffic captured during controlled experiments conducted in a laboratory environment. It includes both normal network traffic (legitimate activity) and malicious traffic (DDoS attacks) to provide a comprehensive view of real-world scenarios. The dataset covers different types of DDoS attacks, such as UDP flood, ICMP flood, TCP SYN flood, and HTTP flood, among others.

Key features of the CICDDoS2019 dataset include:

Network Traffic Attributes: The dataset includes information about source and destination IP addresses, source and destination ports, protocol types, packet size, timestamp information, and other relevant network traffic attributes.

DDoS Attack Scenarios: It encompasses a variety of DDoS attack scenarios, each designed to simulate specific attack characteristics and patterns.

Traffic Flows: The dataset provides details about traffic flows, including flow duration, total packets, total bytes, and various flow statistics. These flow-level attributes enable the analysis of traffic patterns and the identification of anomalous behavior associated with DDoS attacks.

Labeling: Each network flow within the dataset is labeled as either benign (normal traffic) or malicious (DDoS attack). This labeling facilitates supervised learning and evaluation of intrusion detection algorithms.

The CICDDoS2019 dataset is a valuable resource for researchers, analysts, and cybersecurity professionals who are interested in studying DDoS attacks, developing detection methods, and enhancing network security.

To access detailed information, documentation, please visit the official website of the Canadian Institute for Cybersecurity (https://www.unb.ca/cic/datasets/ddos-2019.html). To download the dataset, visit the link (https://www.kaggle.com/datasets/dhoogla/cicddos2019).

Preprocessing Step:

Preprocessing is a crucial step that sets the foundation for accurate analysis and modeling. By applying these preprocessing techniques, you can enhance the quality and reliability of the CICDDoS2019 dataset for your research or analysis purposes. Here are some steps to consider for preprocessing the CICDDoS2019 dataset:

Data Cleaning: Check for and handle missing values, outliers, and inconsistent or erroneous data points.

Feature Selection or Feature dimensionality reduction: Analyze the dataset's features and select the most relevant ones for your analysis. Discard any irrelevant or redundant features to reduce dimensionality and computational complexity (PCA, KernePCA, t-SNE, …). Try to use linear and non-linear reduction techniques instead of using feature selection techniques.

Feature Scaling: Apply appropriate scaling techniques to normalize the numerical features within a specific range (standardization or normalization).

Encoding Categorical Features: If the dataset contains categorical features, encode them into numerical representations suitable for analysis. This can be done using techniques like one-hot encoding, label encoding, or ordinal encoding.

Handling Imbalanced Data (if applicable): If the dataset exhibits class imbalance (significant differences in the number of samples between classes), consider applying techniques such as oversampling, undersampling, or synthetic data generation to balance the classes.

Data Splitting: Split the dataset into training, validation, and testing sets. This ensures proper evaluation of your models and prevents overfitting. Consider using techniques like stratified sampling to maintain the class distribution in each subset.

As part of the preprocessing steps, *data visualization* is an essential component that can provide valuable insights into the CICDDoS2019 dataset. It helps in understanding the distribution, patterns, and relationships within the data (bar chart or pie chart to to visualize the distribution of the different classes (benign and malicious traffic) in the dataset, histograms or kernel density plots to visualize the distributions of numerical features in the dataset, scatter plots to visualize the relationship between pairs of numerical features, heatmap or correlation matrix to visualize the pairwise correlations between numerical features, and so on).

Supervised and Unsupervised Learning

To gain insights and explore different aspects of the CICDDoS2019 dataset, you can apply multiclassification and clustering algorithms using various attack types. Here's an approach to follow:

Identify attack types: Select **three or more** attack types from the CICDDoS2019 dataset that you want to focus on for multiclassification and clustering. Each attack type have a separate file in the dataset (https://www.kaggle.com/datasets/dhoogla/cicddos2019).

Merge the files: Merge the individual files corresponding to each attack type into a single dataset. This can be done using appropriate data manipulation techniques in Python, such as concatenating or merging DataFrames.

Perform multiclassification:

Choose multiclassification algorithms: Select multiclassification algorithms such as Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), or other suitable algorithms.

Prepare the dataset: Ensure that the merged dataset is preprocessed and ready for multiclassification.

Split the data: Split the dataset into training and testing sets using a stratified sampling technique to preserve the distribution of attack types in both sets.

Train and evaluate models: Train the selected multiclassification algorithms on the training set and evaluate their performance on the testing set. Use appropriate evaluation metrics to assess the accuracy and effectiveness of the models in classifying the attack types (F1 score, Confusion matrix, ROC curve).

Perform clustering:

Choose clustering algorithms: Select clustering algorithms such as K-Means, **DBSCAN**, Hierarchical Clustering, Gaussian Mixture Models (GMM), or other suitable algorithms.

Apply clustering algorithms: Apply the chosen clustering algorithms to the preprocessed dataset. Adjust the parameters of each algorithm as needed.

Evaluate clustering results: Evaluate the quality of the clustering results using appropriate metrics such as silhouette score, Davies-Bouldin index, or within-cluster sum of squares (elbow curve).

Deliverables:

1. A detailed notebook (with its export in HTML, allowing to see all the work provided without the need to execute the notebook).
2. A 10-minute presentation during the final session (24 May) describing your work, your approach, and your conclusions.
3. The task to be completed, in groups of **two**, as agreed upon during the last session.