CMPE 789 ML for Cyber Analytics

Towards building a Generic Benign Model across open-source datasets

by Jaseel Muhammad

Problem Statement

To understand the behavior of benign (normal traffic) data and classify them into different clusters based on their common properties. Also, to analyze and compare how the benign traffic is different across datasets. Finally, use this information to build a generic benign classifier using a distance threshold.

Datasets

- CIC-IDS 2017
  - 79 columns, 16 classes
  - Categorical Features: Destination Port, Protocol
- CIC-IDS 2018
  - 80 columns, 15 classes
  - Categorical Features: Destination Port, Protocol, Timestamp
- USB-IDS
  - 84 columns, 17 classes
  - Categorical Features: Flow ID, Source IP, Source Port, Destination IP, Destination Port, Protocol, Timestamp

	CICIDS2017	CICIDS2018	USBIDS
Benign Samples	2,273,097	13,484,754	305,922
Features	85	80	84
Year	2017	2018	2021

Experiment Setup

- Preprocessing
- Clustering without PCA
- Clustering with PCA
- Fine tune clustering hyperparameters
- Compare clusters between dataset
- Find distance threshold
- Classify new data point based on distance threshold

Preprocessing

- Drop categorical features
- Drop columns with only one value
- Replace NaN values with mean values of each column
- Winsorize the data
- Select only data with ‘Benign’ label for initial training

Why Clustering?

- Grouping similar classes together helps us to identify the different patterns in the data.
- Since the approach here is to make a model that does well on benign data, clustering helps in identifying the properties of a benign cluster very well.
- Using clustering, we don’t have to deal with a lot of features, as they get reduced to different cluster centroids and cluster labels

Clustering

- Agglomerative Hierarchical Clustering: This is a "bottom-up" approach in which each observation begins in its own cluster and pairs of clusters are merged as one moves up the hierarchy.
- K-Means clustering: It divides n observations into k clusters, with each observation belonging to the cluster with the closest mean, which is the cluster's centroid.
- DBScan clustering: It is a density-based clustering algorithm that works on the assumption that clusters are dense regions in space separated by regions of lower density.
- Silhouette Score: It is a metric for determining the goodness of a clustering technique. It has a value ranging from -1 to 1. If the score is 1, it means that the clusters are well separated and distinct from one another.

Euclidean and Pairwise distance

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

- Euclidean distance is basically the distance between two points on a cartesian plane.
- Pairwise distance is the distance between each and and every point in n-dimensional space

Classification

- The classifier was first trained on CIC-IDS 2017 with benign labels only to obtain the benign cluster centroids. This is my baseline.
- The malicious (new data) is added to the dataframe and the clustering is done, the centroids of the new cluster formed is obtained.
- The pairwise distance using the euclidean distance for each pair is calculated between the benign cluster centroids and the new data cluster centroid.
- Each class of the CIC-IDS 2017 is classified using the above method, and the median distance between the benign cluster centroid and the new cluster centroid is taken as the distance threshold between benign and malicious data.

Classification(contd.)

- The median distance is chosen because that is a more useful value than taking the mean value as it can be easily influenced by extreme values.
- Any new cluster that has a pairwise distance value greater than the threshold is classified as malicious (not benign).
- After finding the distance threshold, the classifier is then tested for accuracy with a subset of CIC-IDS 2017 dataset.
- It is also tested with CIC-IDS 2018 and the USB-IDS dataset

Classification Results

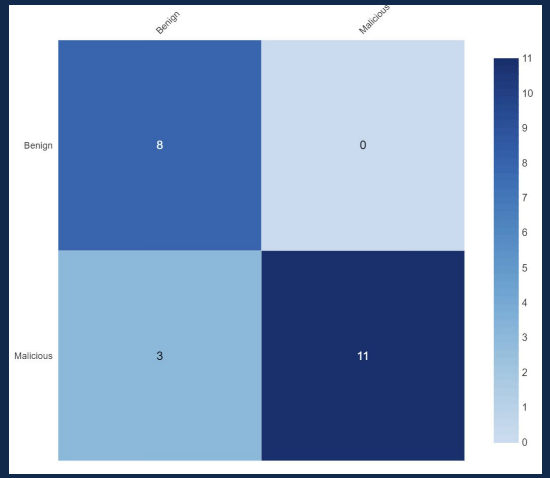
Trained on CIC-IDS 2017. Distance Threshold = 2.49

Dataset	Correct Predictions	Accuracy
CIC-IDS 2017	19/22	86.36 %
CIC-IDS 2018*	16/19	84.21 %
USB-IDS	18/35	51.42 %

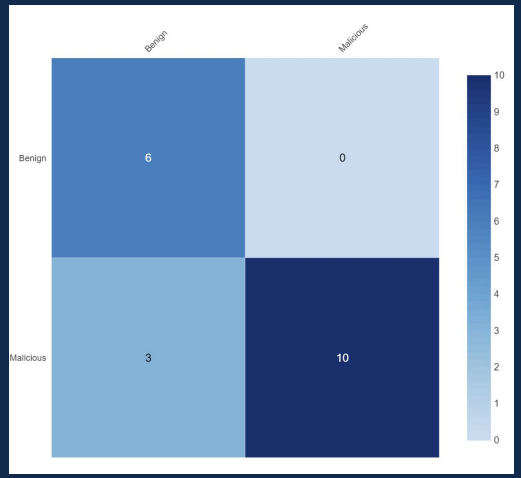
*Skipped Tuesday(02/20/18) due to very large size of csv file - 3.7 GB

Classification Results

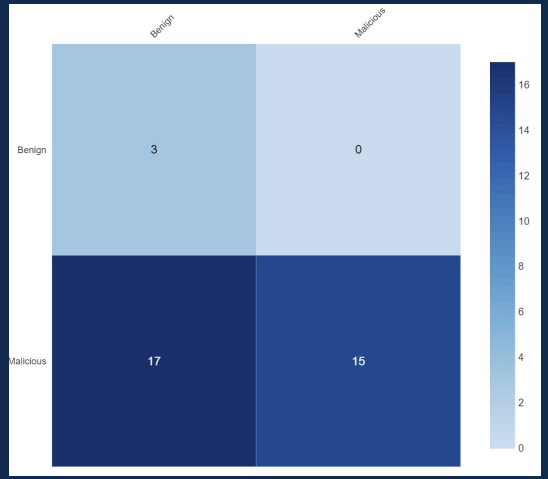
Trained on CIC-IDS 2017. Distance Threshold = 2.49



Tested on subset of CIC-IDS 2017



Tested on CIC-IDS 2018



Tested on USB-IDS

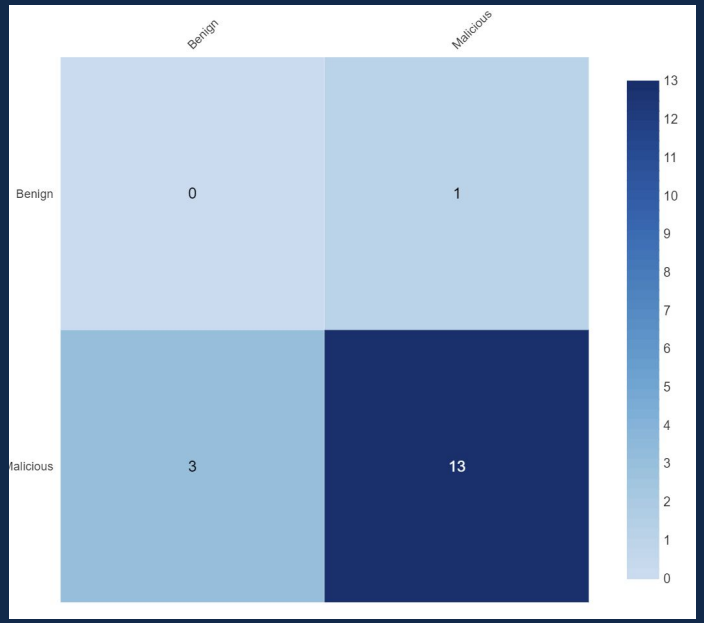
USB-IDS

Due to the low accuracy score on USBIDS when tested with CICIDS2017 data, I retrained it with part of USBIDS data and tested it with the remaining data in the dataset. This showed a considerable increase in correct predictions

Distance Threshold = 1.52

Dataset	Correct Predictions	Accuracy
USB-IDS	13/17	76.47 %

USB-IDS



Tested on subset of USB-IDS

Findings

- Portscan is very similar to benign cluster than any other malicious class in CIC-IDS 2017. It had a distance of only 0.22 while all other classes had a distance greater than 1.5
- Most of the classes in the USB-IDS dataset are very close to each other, i.e, they are overlapping.

Conclusions

- The benign classifier was able to perform very well with classifying benign data as benign even with different datasets
- It was able to classify 14/15 benign clusters accurately. That is an accuracy of 93.33%
- Clustering is a very good method to classify IDS data based on the different clustering algorithms available

Takeaways

- Learned about different cybersecurity concepts and terms.
- This project helped me understand the importance of preprocessing to build a ML model with good performance.
- Learned about how clustering works and its benefits in the context of cybersecurity.
- More experiments can be performed with different datasets and fine tuning this method.

Github Repository

- https://github.com/jazeelmohd/generic-benign-ids-model

Thank you!