BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

Assignment-6_Profiling Network Traffic

Group -102 – Shiva Bansal, Ashish Kumar Singh, Hariom Yadav,
Introduction to Data Science
M.Tech Data Science and Engineering

Overview

- Objective
  The ability to use the data science to differentiate and profile different types of data traversing in the Internet traffic, that is essential for ensuring effective bandwidth distribution and safeguarding network security.
- Methodology
  Clustering techniques is used to categorize data traversing through network. This will help to better understand as similar data will fall into same cluster.

Dataset

- How many features: 40
- Size of the dataset: 22544 (Number of rows in provided dataset)
- Multiple files: No
- What kind of data*: Numerical variable type*
- Balanced or imbalanced – *Highly imbalanced and skewed data*
- Distribution of Training set, validation set, testing set
  - As clustering technique is used, data is not split into different subsets. How….
- Missing data and Preprocessing challenges.
  - *Data is highly skewed and imbalanced*
  - *Higley junk data is present*
-

Feature Engineering Techniques

- Features removed
  - *Highly correlated features are removed during feature engineering using Pearson correlation coefficient and Principal component Analysis*
- Feature creation
  - *All Categorical features are one hot encoded during the analysis*
- Feature ranking
  - *feature importance ranking (FIR) refers to a task that measures contributions of individual input features (variables) to the performance of a supervised learning model.*
- Class imbalance treatment – None Applied as imbalance treatment has its clustering tendency
- Any other
  - *As PCA converts a matrix of n features into a new dataset of less than n features, so we applied to keep only feature which represent maximum variation in data.*
  - *It also helped to improve the accuracy of the model.*

Methodology

- Various type of data exploration techniques are applied to understands the distribution of the data using box plot (for numerical feature) and bar graph (for categorical features.
- Pearson correlation is used to understand feature correlation.
- Following attributes are removed “num_compromised”, “srv_rerror_rate”, “srv_serror_rate”, “dst_host_srv_serror_rate”,”dst_host_srv_rerror_rate” as these features are highly correlated to other attributes.
- Missing values are backfilled with previous value.
- PCA is applied to further reduce feature set to 50.
- Three clustering techniques are tried to categorize the data
  1. K-Means
  2. DBScan
  3. Optics
- ML pipeline is built, which applies all predefined hyperparameters like “number of clusters”, “number of iterations” and it gives model with best cluster quality.
- Below hyper parameters are used:

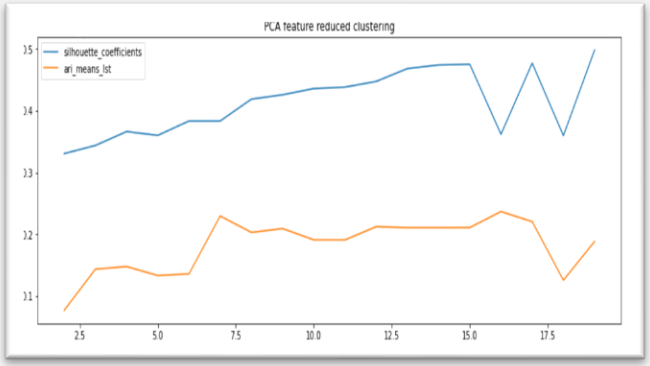
K-Means:
- {"init": "k-means++", "n_init": 10, "max_iter": 100, "random_state": 42}

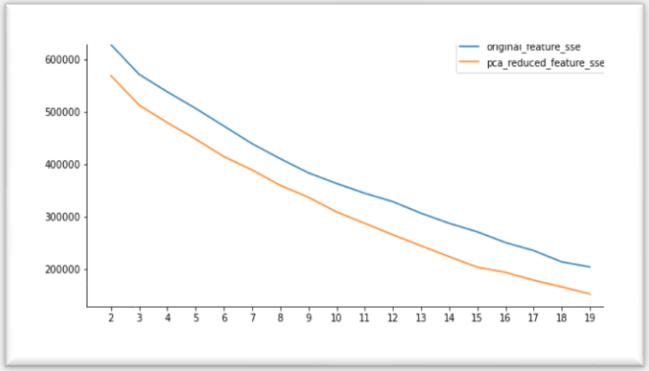
DBScan:
- Eps = 5

Optics:
- min_samples=50

Results

- “Silhouette” and “Adjusted Rand Index” measures are used to understand the *cluster quality.*
- Below are metrics values with different cluster values for KMEANS
- PCA Feature Clustering is done as below





- Silhouette coefficient of DBScan and Optics is 0.29, -0.26 respectively
- Adjusted Rand Index of DBScan and Optics is -0.007 and -0.014 respectively
- Conclusion – Based on the model analysis seems 15th cluster is having good feature clustering. Model shows K-Means having better score than DBScan and Optics, so as conclusion K-Means is better than DBScan and optics.