Alexandria Univeristy
Faculty of Engineering
Computer and Systems Engineering
Department

Pattern Recognition
Assignment 2
Due: March 27th, 2023

Assignment #2: Network Anomaly Detection

Problem Statement

The exponential growth of network traffic has led to an increase in network anomalies, such as cyber attacks, network failures, and hardware malfunctions. Network anomaly detection is a critical task for maintaining the security and stability of computer networks. The objective of this assignment is to help students understand how K-Means and Normalized Cut algorithms can be used for network anomaly detection.

1 Download Datset and Understand the Format

For this assignment, we will use the ”KDD Cup 1999” dataset, which is a widely used benchmark dataset for network anomaly detection. This dataset contains network traffic data collected from a simulated environment, including features such as protocol type, service, source and destination IP addresses, source and destination ports, and attack types. The data is available at the following link

You will use the kddcup.data.gz for training and corrected.gz for testing.

Analyze the dataset and preprocess the dataset to be ready for clustering. ”Change the categorical features to numerical”

2 Clustering Using K-Means and Normalized Cut (Your implementation)

We will use K-Means to cluster the network traffic data and identify anomalies. Every data traffic is a feature vector of 41 dimension. We will use this feature representation to do the clustering

- We will change the K of the K-means algorithm between {7, 15, 23, 31, 45} clusters. You will produce different clusters.

3 Normalized Cut (Your implementation)

Here, we will use Normalized cut algorithm to cluster the network traffic data and identify anomalies. For this experiments we need to decrease the size of the dataset for it to run successfully.

- Set the random seed across all experiments to 42
- Split the training dataset used using train_test_split in sklearn, and take only 0.5% of the data in the new training set. Be sure you set stratify = True.
- Apply Normalized Cut algorithm to the preprocessed data to cluster the data into **23** clusters.
- Rerun the experiments on K-Means when K=23
- Compare the results of K-Means and Normalized Cut clustering in terms of the number of detected anomalies and their characteristics.

4 Evaluation

We will evaluate models based on their ability to detect network anomalies accurately. You will be required to use the following metrics to evaluate the quality of their models:

- Precision
- Recall
- F1 score
- Conditional Entropy

5 New Clustering Algorithm (Your implementation)

Your goal is to get to know to other clustering techniques and how they are working and differ than K-Means and Normalized cut. Choose any clustering technique of your own choice, implement it and repeat the above experiments.

6 Submission Notes

- Work in groups of 3 students.
- You are required to submit a clear and detailed report [in PDF format] illustrating every step in the assignment