Agglomerative Hierarchical Clustering for DDoS Traffic Analysis

Ahmed Yasser Ibrahim

222110758

SE495

1. Introduction

Agglomerative Hierarchical Clustering (AHC) is a bottom-up clustering method that starts with individual data points as separate clusters and merges them iteratively based on similarity. In this lab, we applied AHC to analyze DDoS (Distributed Denial-of-Service) network traffic, aiming to group patterns effectively and distinguish normal from malicious activity.

2. Dataset & Preprocessing

The dataset used is the APA-DDoS-Dataset, which contains various network traffic features. Since it was already well-structured, minimal preprocessing was needed. Instead, we focused on feature engineering to enhance clustering accuracy. The following additional features were introduced:

- **Traffic Rate** = Bytes / Frame Time
- **TCP Flags Sum** = Sum of TCP flag values (SYN, RESET, PUSH, ACK)
- **Packet Size Ratio** = Transmitted Bytes / Received Bytes
- **Interaction Intensity** = Packets / Frame Time
- **Flow Entropy** = A measure of randomness in network traffic

These engineered features helped highlight differences between normal and DDoS traffic, improving the effectiveness of clustering.

3. PCA Analysis

To reduce dimensionality and computational complexity, Principal Component Analysis (PCA) was applied. This helped retain only the most informative features while reducing redundancy.

After applying PCA, the first two principal components captured **99% of the variance**, meaning most of the dataset's information was preserved. This allowed for efficient clustering without significant information loss.

4. Clustering Analysis

Agglomerative Hierarchical Clustering was applied to the reduced dataset using the **average linkage** method to minimize computation while maintaining clustering accuracy. Due to memory constraints, clustering was performed on a subset of the dataset.

A **dendrogram** was used to visualize the hierarchical clustering process, and based on the results, we determined that **two clusters** best represented the data. The clusters successfully captured distinct network traffic patterns.

5. Visualization

To interpret the results:

- A **scatter plot** was generated using the first principal component, showing the two clusters.
- A **dendrogram** was plotted to illustrate the hierarchical structure of the clusters.

Since PCA reduced the data to one primary component, the clustering was visualized in one dimension, clearly demonstrating the separation between normal and DDoS traffic patterns.

6. Conclusion

This lab successfully applied Agglomerative Hierarchical Clustering to detect patterns in DDoS network traffic. The addition of engineered features, combined with PCA for dimensionality reduction, enhanced clustering performance while maintaining computational efficiency.

Future Improvements:

- Explore different **distance metrics and linkage methods** to refine cluster formation.
- Compare **AHC with K-Means or DBSCAN** to evaluate clustering effectiveness.
- Analyze **cluster quality using Silhouette Score or Davies-Bouldin Index**.
- Experiment with larger dataset subsets to enhance generalizability.

This analysis demonstrates the potential of hierarchical clustering in cybersecurity applications, especially for anomaly detection in network traffic.