O. Filat, O. Kiseleva, Yu. Honcharova

EFFICIENT SET SEGMENTATION USING CLUSTERING ALGORITHMS FOR NETWORK TRAFFIC ANALYSIS

Network traffic analysis is a complex and multifaceted issue that encompasses a range of technologies and implementations designed to capture, process, categorize, control, and modify network packets based on their content. It is also a critical aspect of modern network security, involving the real time collection, processing, and analysis of network packets to detect and prevent security threats.

By using efficient set segmentation techniques, cluster analysis can provide more comprehensive understanding of network traffic patterns, allowing for the identification of abnormal behavior that may indicate a security breach. This is particularly useful for identifying network attacks, which can be distinguished from normal traffic patterns through the use of data analysis methods such as entropy. In this article, we will explore the role of cluster analysis in network traffic analysis and its applications in efficient set segmentation for improved security, with a focus on the identification of network attacks and normal traffic patterns.

The primary objective of the study is to develop and apply methods and algorithms for accurate and proactive identifying various types of network attacks within network traffic. Recent research has demonstrated that data analysis techniques can differentiate efficiently and accurately normal and abnormal traffic, using the concept of entropy. In this study, we propose to use clustering methods such as hierarchical clustering, DBSCAN, OPTICS, and K-MEANS to detect attacks. Our results demonstrate that DBSCAN, OPTICS along with hierarchical clustering provide the best outcomes.

During the research, a dataset with 4,898,431 rows and 42 columns was utilized. One column describes the type of traffic and is not involved in the clustering process, leaving 41 columns to be processed. 15 columns contain floating-point data, 23 columns contain integer data, and 3 columns contain object data. The data preprocessing revealed that most users employ private and http services, so we focused our analysis on these two types of services. Similarly, we selected only SO, SF, and REJ flags from the flag column for analysis.

Also, after conducting the analysis of boxplot, distplot, and violin plots for each feature, we identified several features that were candidates for removal from the dataset. Specifically, based on our analysis, we determined that the su_attempted and num_file_creations features could be dropped from the dataset.

Hierarchical clustering with single, average, and total linkage was the first clustering method used in the study. The distances between the points were measured using Euclidean, Manhattan, or Canberra distances. Principal component analysis was also used to reduce the data dimension. A confusion matrix was used to evaluate clusters that contained attacks, and the silhouette score was used to identify an appropriate number of clusters. DBSCAN and OPTICS algorithms were implemented in the same way. Below you can a short description of the best clustering results.

1. Hierarchical clustering with single linkage

The Euclidean metric is used to present the results of clustering for 5 clusters. Parallel coordinate plots were constructed for all features, and as an example, we present a plot for the feature of duration (connection time).

2. DBSCAN

The best interpretational results were achieved with clustering using the Manhattan distance metric. The data was divided into 5 clusters using this algorithm with input parameters of ε= 3 and a minimum number of points of 10.

3. OPTICS

During clustering with the OPTICS algorithm, a value of εwas chosen as 0.5. The best results were obtained using the Canberra distance metric, although the Euclidean and Manhattan metrics also produced fairly accurate results. In all three cases, the data were clustered into 2 clusters using the algorithm. Below are the results using the Canberra distance metric.

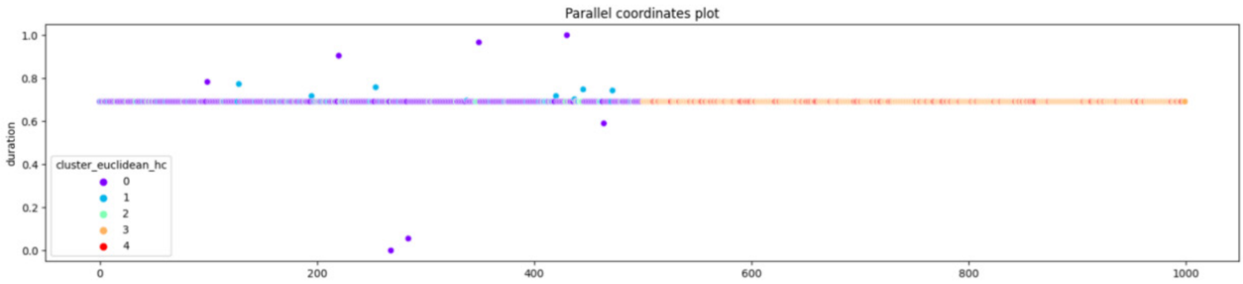

Fig. 1 – The plot for the duration feature of data clustered by the hierarchical method (Euclidean distance)

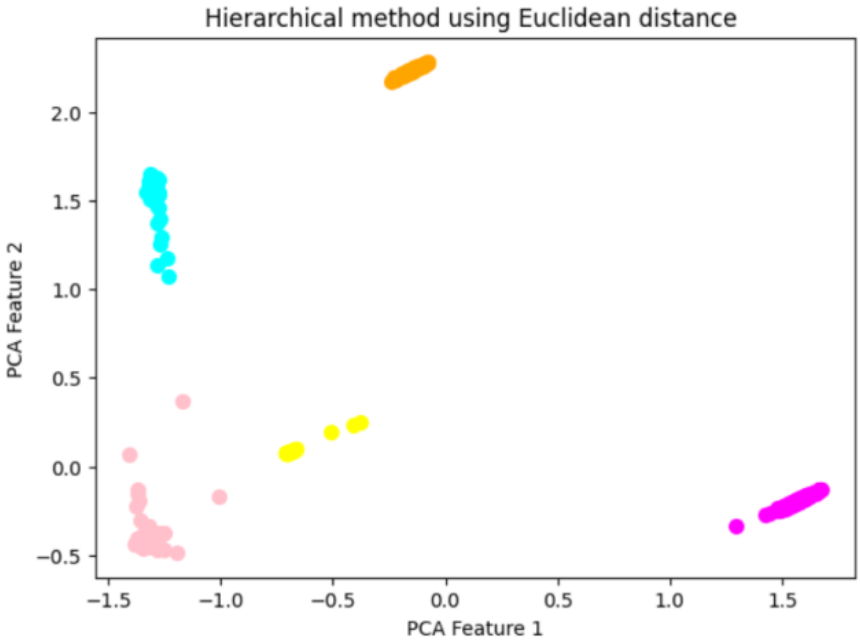

Fig. 2 – The plot of data clustered by the hierarchical method using Euclidean distance and principal component analysis

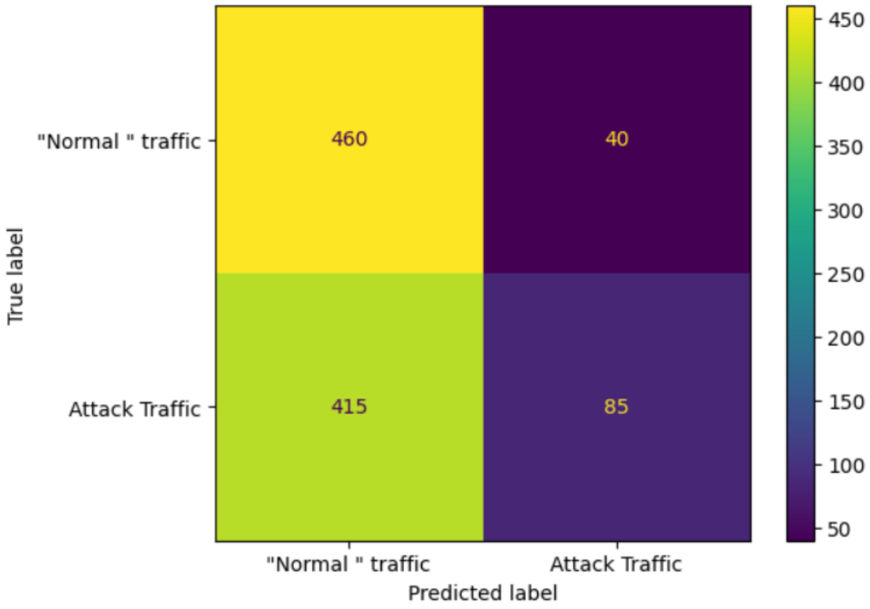

Fig. 3 – The confusion matrix for data clustered by the hierarchical method using Euclidean distance

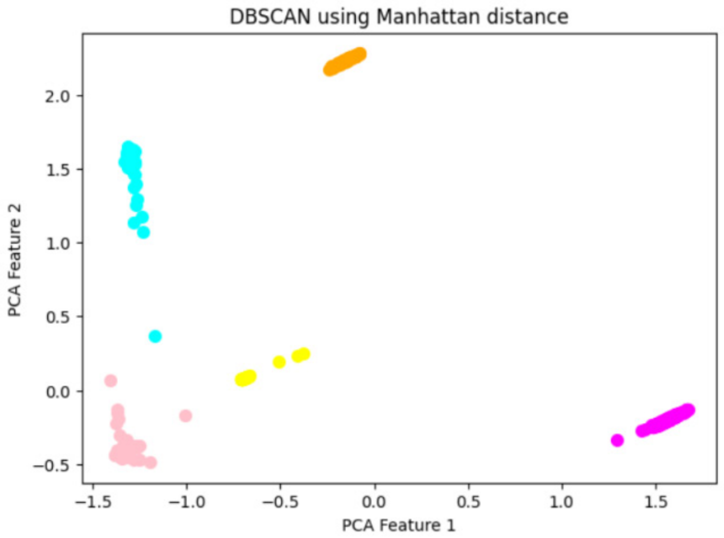

Fig. 4 – Graph of data clustered using the DBSCAN method with Manhattan distance and applying Principal Component Analysis (PCA)

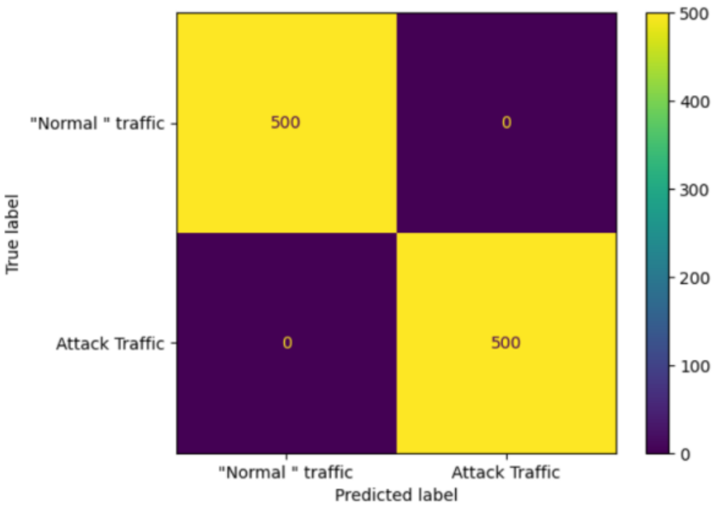

Fig. 5 – Confusion matrix for data clustered by DBSCAN method using Manhattan distance metric

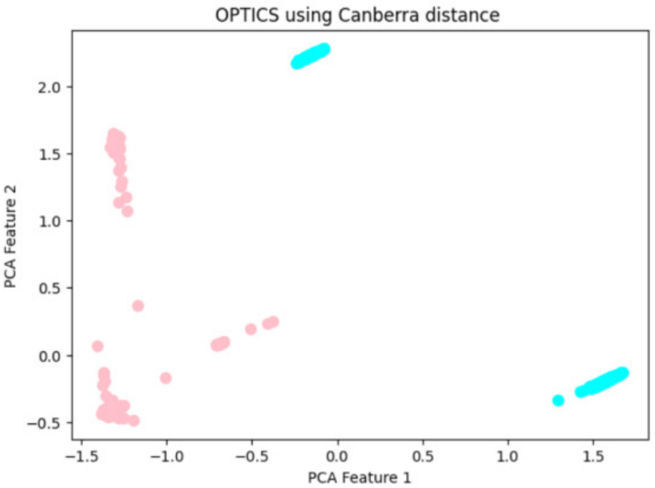

Fig. 6 – Graph of data clustered by OPTICS method, using Canberra distance and principal component analysis (PCA) method

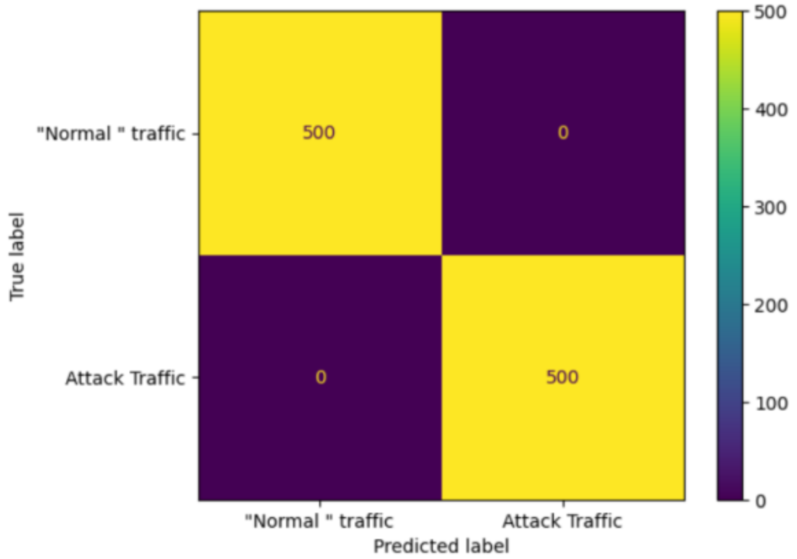

Fig. 7 – Confusion matrix for data clustered by OPTICS method, using Canberra distance

To sum up, clustering methods are important tools in detecting changes in network traffic and identifying attacks. Density-based methods, which can identify outliers and uncover the underlying clustering structure, are particularly useful in identifying traffic related to attacks. With the increasing complexity of the field of efficient set segmentation, there is a growing expectation for clustering to have a more prominent role in its applications. Further research and innovation in clustering techniques will be the key points for unlocking their potential in this area.

REFERENCES

1. Filat O., Tonkoshkur I. Network Traffic Analysis Using Clustering Algorithms. The *latest problems of modern science and practice. Proceedings of the I International Scientific and Practical Conference*. Boston, USA. 2022. Pp. 430-434.
2. Rahmah N. Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra. *IOP Conf. Series: Earth and Environmental Science,* №31. 2016.