20th International Road Safety on Five Continents Conference

Hierarchical Clustering on Graph Embeddings: A Scalable Approach to Risky Intersections

Simone Paradiso[a, *], Apostolos Ziakopoulos[a], George Yannis[a]

[a]*National Technical University of Athens, Department of Transportation Planning and Engineering, 5 Iroon Polytechneiou Street, GR-15773, Athens, Greece*

(Contact: simone_paradiso@mail.ntua.gr, apziak@central.ntua.gr, geyannis@central.ntua.gr)

Keywords: Road Safety, Hierarchical Clustering, Graph Neural Network, Risky Intersections

Abstract

Background

It is well-documented that road crashes and their consequences continue to be a serious global societal concern, as 1.19 million road users lost their lives in 2021 due to involvement in crashes (WHO, 2023). To investigate their causes and to propose solutions, Machine Learning (ML) and Deep Learning (DL) techniques have gained significant attention among road safety researchers along with the increasing availability of data (Sohail et al., 2023). Researchers have extended DL methods to graph-structured thanks to the Graph Neural Network (GNN) (Scarselli et al., 2009), by incorporating the attention mechanism, enabling the model to capture graph context and relationships more effectively, as with the Graph Attention Network (GAT) (Veličković et al., 2018). Telematics data can provide insights into road safety assessments, together with the geometric data offering more context. The present study aims to explore a hierarchy of data point groups using an agglomerative hierarchical clustering, where the data points are the nodes in the analysed graph, and improve the analysis by involving edge features employing the GAT model.

Methods

Telematics data provided by OSeven (2025), covering the northeastern Athens metropolitan area, include features such as trip ID, geographic coordinates, speed, and several binary indicators of risky driving behaviour. These data have been spatially aggregated onto the road network extracted from OpenStreetMap (2025) with aggregation methods chosen according to variable type: numerical features (e.g., speed) were averaged, while discrete events (e.g., binary flags) were summed. In the present study the GAT model was used to obtain node embeddings involving edge features of a graph. Given a simple undirected graph $G = (V, E)$ the network embedding is a mapping function:

$$\Phi(V, E) \rightarrow \mathbb{R}^{|V| \times d}$$

Where $d \ll |V|$ is the dimension of the output embeddings. At this point two datasets with equal dimension, containing respectively the raw features and the embeddings have been analysed using hierarchical clustering, that is an unsupervised ML algorithm used to group data into groups, building a hierarchy of clusters. The agglomerative approach was chosen (Murtagh & Contreras, 2011). Once two or more points have been merged, a criterion to determine the distance between clusters, called Linkage Method, must be chosen, having an impact on the shape and size of the clusters.

The hierarchical clustering output is a dendrogram displaying the hierarchical relationship between different sets of data (Benatti & Costa, 2024), analysed per each linkage method. The Cophenetic Correlation Coefficient (CCC) is used to assess the clustering performance (Saracli, 2013), defining how well the algorithm preserves the pairwise distances of the original data points. This study employed six linkage methods, presented in previous literature (Jarman, 2020), (Oti & Olusola, 2024): (1) Single, (2) Complete, (3) Average, (4) Centroid, (5) Median, (6) Ward's.

Results

Key Insights from Dendrogram Analysis of Raw Features

The single method points merging at low heights, related to bridging points leading to the chain effect, whereas the complete method points merging at various heights. The average and centroid methods yield similar dendrograms having short stems leading to a growing main cluster. The median method highlights a small cluster separated from the central one. Ward's method yields a good CCC, enabling a separation of the dense data points. However, it does not rely on pairwise distances, sacrificing distance accuracy in favour of forming clusters, identifying clusters even when only statistical fluctuations are present. Yet, different stem lengths in the dendrogram indicate a certain level of separation in the data (Benatti & Costa, 2024). Thus, Ward's method was chosen for achieving a more distinct separation in the data.

Key Insights from Dendrogram Analysis of Feature Embeddings

The single method exhibits an emphasized chaining effect, indicating a continuous structure in the data, expected from the embedding process, further highlighted by high CCC. Centroid and average methods show similar results with points gradually forming a singular group, the median method produces similar results as well, strengthening the notion that the data are tightly packed.

However, the complete method shows a small clustering emerging. Once again, Ward's method successfully separates the data, as previously observed, identifying two groups at relatively high linkage heights. This results in a higher Cophenetic Correlation Coefficient (CCC) compared to clustering based on the raw features, though with a different group size distribution. Thus, Ward's method appeared as the most suitable choice.

Discussions and Conclusion

Discussion focuses on the selected Ward's method applied to feature embeddings. The dendrogram encourages separation in 2 clusters, giving as output one risky cluster displaying increased speeding and mobile usage and more frequent harsh events, while the other one exhibits lower risk traits. Downwards, the number of clusters increased to 3, resulting in the previous safer cluster being split into two subgroups: one with intermediate characteristics in terms of speeding and harsh events, yet interestingly displaying the highest mobile usage, besides the highest trip frequency.

The current work proposes an approach to understand the nature of graph-structured data, showing how node representations through GNNs improve clustering performance. It explores the resulting hierarchy from the hierarchical clustering, which allows for clearer selection of the number of clusters by inspecting the dendrogram and also demonstrates the potential to scale the analysis for deeper insights into road safety. The division of the safer cluster into two subgroups highlights an alternative risk perspective, emphasizing the value of a multi-level clustering approach to capture nuanced behavioural patterns relevant to road safety analysis. Further research could extend this work by investigating other linkage methods, or by investigating different GNN architectures for further insights into graph partitioning.

REFERENCES

Benatti, A., & Costa, L. da F. (2024). *Agglomerative Clustering in Uniform and Proportional Feature Spaces* (No. arXiv:2407.08604). arXiv. https://doi.org/10.48550/arXiv.2407.08604

Jarman, A. (2020). *Hierarchical Cluster Analysis: Comparison of Single linkage,Complete linkage, Average linkage and Centroid Linkage Method*. https://doi.org/10.13140/RG.2.2.11388.90240

Murtagh, F., & Contreras, P. (2011). *Methods of Hierarchical Clustering* (No. arXiv:1105.0121). arXiv. https://doi.org/10.48550/arXiv.1105.0121

OpenStreetMap. (2025). https://wiki.openstreetmap.org/wiki/About_OpenStreetMap

OSeven. (2025). https://oseven.io/

Oti, E., & Olusola, M. (2024). Overview of Agglomerative Hierarchical Clustering Methods. *British Journal of Computer, Networking and Information Technology*, *7*, 14–23. https://doi.org/10.52589/BJCNIT-CV9POOGW

Saracli, S. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*. https://www.academia.edu/32926376/Comparison_of_hierarchical_cluster_analysis_methods_by_cophenetic_correlation

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80. IEEE Transactions on Neural Networks. https://doi.org/10.1109/TNN.2008.2005605

Sohail, A., Cheema, M. A., Ali, M. E., Toosi, A. N., & Rakha, H. A. (2023). Data-driven approaches for road safety: A comprehensive systematic literature review. *Safety Science*, *158*, 105949. https://doi.org/10.1016/j.ssci.2022.105949

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph Attention Networks* (No. arXiv:1710.10903). arXiv. https://doi.org/10.48550/arXiv.1710.10903

WHO. (2023). *Global status report on road safety 2023*. https://www.who.int/publications/i/item/9789240086517