Iyana Jones & Coreen Mullen

Midterm Report

Objectives & Key Findings

- **Clustering Map:** Develop a clustering map that visualizes accident hotspots across North Carolina, using accident data over a specified time period.
- **Environmental Analysis:** Examine the potential correlation between environmental factors (e.g., road types, traffic density, weather conditions) and accident hotspots to understand why these areas experience higher accident rates.
- **Awareness:** Inform local communities, city planners, and government officials about the locations of these accident hotspots. The map generated from this research will help decision-makers prioritize areas for intervention, such as traffic control measures, road improvements, or increased law enforcement in high-risk zones.
- **Impact on Policy:** Contribute to the development of public policies aimed at reducing accidents, enhancing traffic safety measures, and guiding infrastructure investments toward areas with high accident rates.

Methodology/Solution

The project employed a two-pronged approach for clustering analysis, utilizing the K-means and DBSCAN algorithms.

1. **Data Collection:** Accident data was gathered from state traffic databases, which included details such as accident location (latitude and longitude), time, severity, and environmental conditions like weather or road type.
2. **Data Preprocessing:** The data was cleaned by handling missing values, removing outliers, and normalizing geographical coordinates to ensure consistency.
3. **Clustering Algorithms:**

- ○ **K-means Clustering:** This algorithm partitions the dataset into k distinct clusters based on accident location. The number of clusters (k) was determined using methods like the elbow method to find an optimal value for k.
- ○ **DBSCAN Clustering:** This density-based algorithm identifies clusters based on the density of data points, making it especially useful in identifying irregularly shaped clusters and handling outliers.

4. **Comparative Analysis:** We compared the results from both algorithms, evaluating their effectiveness in identifying meaningful accident hotspots and their ability to handle noise and irregularly distributed data.
5. **Visualization:** The final output was a map of North Carolina displaying the accident hotspots. The hotspots were color-coded and highlighted to visually represent high-risk areas, with an overlay of potential environmental factors.

Literature Review

Research on clustering methods has demonstrated their effectiveness in various domains, including traffic analysis. For example, in "A Novel Classification Algorithm Based on the Synergy Between Dynamic Clustering with Adaptive Distances and K-Nearest Neighbors," Sabri, Verde, and Balzanella (2024) propose a dynamic clustering approach that allows for continuous updates to the clustering model as new data becomes available. This flexibility makes dynamic clustering particularly beneficial in traffic-related studies where accident data is constantly changing. Their methodology of adaptive distances can be particularly useful for handling data where the density and distribution of accidents are not uniform across the region.

Additionally, Alam (2023) presents a hybridization of K-means clustering with an improved Firefly algorithm (ODFA) for automatic clustering, particularly in high-dimensional datasets. algorithm determines the optimal number of clusters **K** by leveraging the Firefly Algorithm (FA) to search for the best clustering configuration. The Firefly Algorithm is an optimization technique inspired by the flashing behavior of fireflies. In this hybrid approach, the FA is used to optimize the K-means clustering process. Specifically, FA evaluates various potential values of **K** by adjusting the position of fireflies (representing possible solutions). The fitness function for the fireflies is based on the clustering performance, typically measured by an objective function like the sum of squared errors (SSE) or intra-cluster distance. As the fireflies move towards better solutions (lower SSE or better compactness), the optimal value of **K** emerges through this search.

The fitness function is used to evaluate the quality of the clustering solution represented by the position of each firefly. For clustering, the **objective function** could be the **sum of squared errors (SSE)** or the **within-cluster distance**.

Fitness=$\sum N \sum K \|x_i - \mu k\|^2$ where:

- **N** is the total number of data points.
- **K** is the current number of clusters (represented by a firefly's position).
- μk is the centroid of cluster **k**.
- x_i i is the *i*-th data point.

The algorithm balances exploration and exploitation to find the best clustering structure, thus determining the most suitable **K** for the given high-dimensional data set. It also helps with clustering accuracy and we are able to obtain more robust results in identifying hotspots and understanding underlying patterns in the data.

Expected Outcomes/Impact

The outcome of our research will be a detailed, visual representation of accident hotspots in North Carolina. By applying both DBSCAN and K-means clustering methods, we aim to:

- Provide a comprehensive map showing the areas of highest risk for traffic accidents, which can be used by local governments and transportation agencies for targeted interventions.
- Identify patterns and relationships between environmental factors and accident hotspots, providing valuable insights into the causes behind these high-risk areas.
- Promote the use of data-driven decision-making in traffic safety policies and urban planning. This can lead to more informed actions, such as adjusting speed limits, improving road signage, increasing law enforcement presence, or implementing road redesigns in high-risk zones.

- Ultimately, the project is expected to have a positive impact on reducing traffic-related accidents and fatalities, benefiting both the local community and the broader public.

Citations

Sabri, M., Verde, R., Balzanella, A. et al. A Novel Classification Algorithm Based on the Synergy Between Dynamic Clustering with Adaptive Distances and K-Nearest Neighbors. *J Classif* 41, 264–288 (2024). https://doi.org/10.1007/s00357-024-09471-5
Alam, Afroj. (2023).

Hybridization of K-means with Improved Firefly Algorithm for Automatic Clustering in High Dimension. *arXiv preprint* arXiv:2302.10765. https://arxiv.org/abs/2302.10765

Yang, XS. (2009). Firefly Algorithms for Multimodal Optimization. In: Watanabe, O., Zeugmann, T. (eds) Stochastic Algorithms: Foundations and Applications. SAGA 2009. Lecture Notes in Computer Science, vol 5792. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04944-6_14