Clustering: Spatial and Temporal Analysis of Traffic Data

Ethan Alexander

This report details the methodologies and decisions involved in a clustering analysis of traffic data. The goal was to investigate spatial and temporal patterns using multiple clustering approaches. All code and visualizations are included in the accompanying Jupyter Notebook.

1. **Data Loading and Preparation**
   To begin the analysis, I modified the provided sample code snippet to efficiently loop through and load data from six text files into a single Pandas DataFrame. This ensured consistency and ease of manipulation for further processing.

2. **Spatial Clustering**
   a. **Initial Clustering Using Simple Features**
      To explore the relationship between the given ‘cluster’ ID and the attributes ‘entry’ and ‘maneuver,’ I first grouped the data by ‘cluster’ and examined summary statistics. The results showed that the mean, minimum, and maximum values for both ‘entry’ and ‘maneuver’ were identical within each cluster, indicating that each cluster corresponded to a specific ‘entry’ and ‘maneuver’ combination without variation.

      A notable observation was that while there were 11 clusters, only nine unique ‘entry’-‘maneuver’ combinations existed. Two combinations—East Entry/Straight Direction and West Entry/Straight Direction—were represented by two separate clusters. This distinction arose because, in some cases, straight turns from the east and west were considered separate despite being labeled identically in ‘maneuver.’

      Using this insight, I implemented a K-means clustering algorithm with nine clusters, reflecting the nine ‘entry’-‘maneuver’ combinations. This clustering approach closely resembled the original clustering method, with the key difference being the separation of ‘Straight 1’ and ‘Straight 2.’

b. **Clustering Using Trajectories**
    i. **Minimum/Maximum X and Y Points**
    This method grouped vehicles based on their minimum and maximum x and y coordinates using K-means with 11 clusters, mirroring the original clustering approach.
    ii. **First/Last X and Y Points**
    Vehicles were clustered based on their first and last recorded x and y coordinates, again using K-means with 11 clusters.
    iii. **Model Comparisons**
    Scatter plots for each model revealed that the three clustering methods largely agreed in their cluster assignments. To quantitatively compare these models, I used confusion matrices and the Adjusted Rand Index (ARI). The confusion matrices illustrated that both the Min/Max and First/Last models closely matched the original ground-truth clustering, with the Min/Max model misclassifying only 58 points and the First/Last model showing about 2,500 misclassified points.
    ARI values further supported these observations, with the Min/Max model scoring 0.999 and the First/Last model scoring 0.971. Since ARI ranges from 0 to 1 (where values closer to 1 indicate a stronger agreement between clustering models), these results confirmed that both trajectory-based clustering approaches closely mirrored the original model.

3. **Temporal Clustering**
    To explore clustering based on speed rather than spatial differences, I developed two additional models: one clustering vehicles by total time spent in the intersection, and another by average speed while in the intersection.
    Total time was calculated as the difference between each vehicle's maximum and minimum reported time, measured in deciseconds. To address concerns that total time alone might not be an equal basis for comparison across vehicles, I also considered average speed, calculated as total distance traveled divided by total time in the intersection.
    Unlike the spatial clustering models, these temporal clustering methods produced significantly different results. The ARI scores for the Speed and Total Time models, when compared to the original clusters, were 0.076 and 0.019, respectively. These low scores indicated that speed-based clustering captured a different aspect of the data than the spatial clustering models, suggesting that spatial position and movement dynamics are not necessarily correlated with temporal behavior.

4. Conclusions

This project provided a comprehensive analysis of traffic patterns through clustering, utilizing both spatial and temporal features. The spatial clustering models—particularly those using Min/Max and First/Last x and y coordinates—are closely aligned with the original clustering methodology, as evidenced by high ARI scores. On the other hand, temporal clustering based on speed and total time in the intersection revealed distinct patterns, highlighting the difference between spatial movement and temporal dynamics.

These findings demonstrate the importance of selecting appropriate clustering features based on the specific characteristics of the data being analyzed. The project showcases skills in data preprocessing, clustering model implementation, and performance evaluation using confusion matrices and ARI metrics. The detailed comparisons and visualizations in the Jupyter Notebook provide further insights into the effectiveness of different clustering approaches.