TeamID: DTC4872 | Team Name: D-Wing: Vishal Kumar | Shivam Sood | Harsh Sharma | Soumyo Bhattacharjee

Price Setter and Follower Detection – KDAG Hackathon

Report

Data Preprocessing and Feature Engineering

The Dataset demands the use of unsupervised learning methods, clustering in this case. To apply clustering numeric data must be generated.

The following features were generated:
1. Absolute time difference between the service date and the date at which the prices were recorded.
2. Whether the service date was a weekend or not.
3. The fare types (type 1 or type 2)
4. The average and the standard deviation of fare price.

The data was then grouped by the 'bus-ID', taking the mean of all of the above features and the maximum of the absolute time difference. The data was then normalized.

The total dimensions generated were six. Since clustering algorithms do not run well on high dimensional data, Principal Component Analysis (PCA) was performed and the data was projected onto two dimensions, which was then used for the model.

Model Training and Output Generation:

K-means clustering (K-means++) was used in two stages to find similarity between the data points. The metric used to generate confidence metric (probability of x being in cluster c_i) was:

$$\frac{\frac{1}{||x - c_i||}}{\frac{1}{||x - c_i||} + \sum_j \frac{1}{||x - c_j||}}$$

where, x is a data point, c_i is its corresponding cluster center and c_j are the remaining cluster centers.

This metric was further used to re-cluster each original cluster based on the silhouette score. Then the confidence metric was recalculated based on the secondary clusters.

The leader follower relationship was calculated by assigning those closest to the secondary cluster centers as leaders and their adjacent neighbour as their follower. The final confidence score was adjusted as required.

The output file was generated and has been submitted.

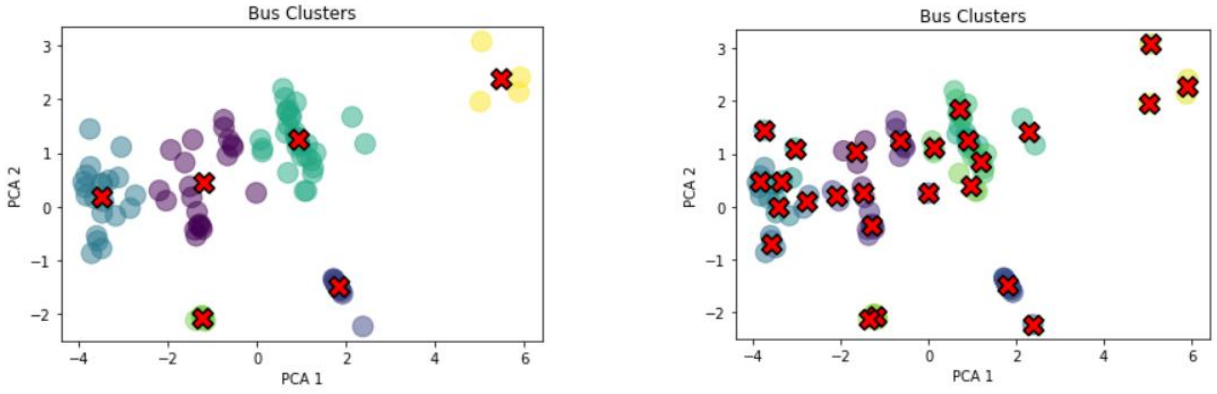

Fig: The original Cluster Centers (left) and Secondary Cluster Centers (right).