CPS803 Assignment 5 Report

Done By: Kris Soni

Student#501147159

Background

The dataset used in this analysis is the **Car Evaluation Data Set**, which was obtained from the UCI Machine Learning Repository link to dataset.[1] This dataset contains information on various car attributes, such as buying price, maintenance cost, number of doors, number of persons it can accommodate, luggage boot size, and safety ratings. The target variable is the car evaluation class, which categorizes the cars as either “unacceptable”, “acceptable”, “good”, or “very good”.[1]

This dataset is interesting because it provides valuable insights into how different factors influence the evaluation of cars, making it relevant for car manufacturers, car buyers, and market analysts. The categorical nature of the data (such as 'buying', 'maint', 'doors') makes it a suitable candidate for clustering analysis, as we can group similar types of cars and identify patterns based on their features.

Methods

For this analysis, we employed the **K-Modes clustering algorithm**. The K-Modes algorithm is particularly suited for categorical data, unlike K-Means, which is primarily used for continuous data. K-Modes works by minimizing the dissimilarity between objects in a cluster based on their modes (the most frequent category in a cluster) rather than their means, making it ideal for the car evaluation dataset, which consists entirely of categorical attributes.

Clustering Process

1. **Preprocessing**:
    - We first loaded the dataset and selected relevant features. The target variable 'class' was excluded from the clustering process, as we aimed to identify patterns and groupings in the data based on the other attributes.
    - We used the **K-Modes** implementation from the kmodes Python package to perform clustering.
2. **Cluster Initialization**:
    - We initialized the K-Modes algorithm with 8 clusters (n_clusters = 8) and used the **Huang initialization method** to assign initial cluster centroids randomly.[2]
    - The model was run with 10 initializations (n_init=10) to ensure the solution was stable and robust.

3. **Other Clustering Methods**:

    - Before settling on K-Modes, we attempted other clustering techniques, such as **Hierarchical clustering** and **DBSCAN**. However, both methods faced challenges with the categorical nature of the data:
        - **Hierarchical clustering**: This method requires a distance metric that works well with categorical data. The usual distance metrics like Euclidean distance did not effectively capture the dissimilarities between categorical attributes, leading to less meaningful clusters.
        - **DBSCAN**: DBSCAN, which relies on density-based clustering, also struggled with categorical data. It requires a meaningful way to measure density and distance between points, which was not suitable for the discrete nature of the features in our dataset.
    - **Why K-Modes?**: K-Modes clustering was selected due to its effectiveness with categorical data. Unlike K-Means, which calculates distances between points based on means, K-Modes focuses on modes, which makes it more natural for grouping categorical data. This is why K-Modes was the best choice for clustering this dataset.[2]
4. **Evaluation**:

    - The evaluation strategy involved examining the cluster distribution, reviewing the cluster centroids, and visualizing the results using bar plots. This approach helped assess whether the clusters were meaningful based on the car attributes and the target variable 'class'. The cluster distribution and centroids provided insights into how well the algorithm grouped cars with similar characteristics, while the visualizations offered a clear overview of the clustering results.

Results

The K-Modes clustering algorithm was applied to the **Car Evaluation Data Set** using 8 clusters, and the following results were obtained from 10 runs. Each run initialized the centroids and started the iterations for clustering. The cost values for each run reflect the overall dissimilarity between the cluster centroids and the data points assigned to them. The best run was determined based on the lowest cost value, which was observed in **Run 3** with a final cost of 4393.0. This cost metric indicates how well the clusters have been formed based on the chosen categorical attributes.

Cluster Distribution

The distribution of instances across the clusters was as follows:

Cluster distribution:
cluster
0 374
1 304
2 229
3 198

6 183
5 161
4 159
7 120

Cluster 0 had the largest number of instances (374), while Cluster 7 had the fewest (120). This shows that the data is somewhat distributed across the clusters, though some clusters contain a significantly larger number of instances than others.

Cluster Centroids

The cluster centroids, representing the most frequent attribute values in each cluster, were as follows:

Cluster centroids:
[['low' 'vhigh' '2' 'more' 'big' 'med']
 ['vhigh' 'vhigh' '4' '4' 'small' 'low']
 ['low' 'low' '3' '2' 'small' 'low']
 ['med' 'high' '2' '4' 'big' 'low']
 ['vhigh' 'high' '3' 'more' 'small' 'med']
 ['vhigh' 'low' '3' '2' 'big' 'high']
 ['high' 'low' '2' '4' 'med' 'high']
 ['low' 'high' '5more' 'more' 'small' 'high']]

Each centroid corresponds to a specific combination of values for the categorical attributes, such as buying price, maintenance cost, number of doors, number of persons, luggage boot size, and safety rating. These centroids provide insight into the typical characteristics of the cars in each cluster.

Cluster Analysis

The following table summarizes the most frequent values (modes) for each cluster, highlighting the distinct characteristics of the cars within each group:

Cluster	Buying	Maint	Doors	Persons	Lug Boot	Safety
0	low	vhigh	2	more	big	med
1	vhigh	vhigh	4	4	small	low
2	low	low	3	2	small	low
3	med	high	2	4	big	low
4	vhigh	high	3	more	small	med

5	vhigh	low	3	2	big	high
6	high	low	2	4	med	high
7	low	high	5more	more	small	high

The clusters show clear patterns based on the different attributes. For example:

- **Cluster 0** contains cars that are inexpensive to buy and maintain but have a large luggage boot and medium safety.
- **Cluster 1** represents cars that are highly rated for both buying and maintenance, with 4 doors and a small luggage boot.
- **Cluster 7** features cars with a high safety rating, small luggage boot, and a large number of passengers.

These distinct clusters highlight the variation in car characteristics based on the features provided.

Cluster Summary and Visualization

A summary of the clusters was generated, displaying the most frequent attributes for each group. This was visualized through a count plot, showing the distribution of instances across the different clusters. The clustering revealed meaningful patterns in car evaluation, which could be useful for various stakeholders in the automotive industry.

In conclusion, the K-Modes algorithm effectively grouped cars with similar characteristics, providing a useful categorization of the cars in the dataset.

Conclusion

The K-Modes clustering algorithm successfully identified distinct groups of cars based on their attributes, although one cluster was dominant (Cluster 0), which could be attributed to the imbalance in the data. We encountered challenges with other clustering methods like Hierarchical and DBSCAN, which were not well-suited for the categorical nature of the dataset. The K-Modes algorithm, with its ability to handle categorical data, emerged as the most effective method for this analysis.

References

1. UCI Machine Learning Repository. Car Evaluation Data Set. Link to Dataset.
2. Huang, Z. (1998). "Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values". *Data Mining and Knowledge Discovery*.

Used for coding:

- Python KModes Documentation. KModes Documentation.
- Seaborn Documentation. Seaborn Documentation.
- Matplotlib Documentation. Matplotlib Documentation.