Customer Segmentation in Credit Card Usage: A Comparative Analysis of Clustering Techniques

Lucas Teixeira Rocha - m11813 .
Computer Science and Engineering - 2nd Cycle Degree
Universidade da Beira Interior
Covilhã, Portugal
lucas.rocha@ubi.pt

***Abstract*—This study conducts a comprehensive analysis of a credit card dataset, comprising behavior of approximately 9,000 users, using various clustering algorithms. The primary objective is to categorize these users into distinct segments, employing well-known methods like K-Means, DBSCAN, MiniSom, and Hierarchical Clustering. Key preprocessing steps, including data normalization and handling missing values, are undertaken to ensure data integrity. The study also leverages dimensionality reduction techniques, namely PCA and t-SNE, to aid in the visualization and interpretation of the complex, high-dimensional dataset. The effectiveness of each clustering method is evaluated through metrics such as the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. The findings reveal unique insights into customer segmentation, providing valuable implications for targeted marketing strategies and customer service enhancements in the credit card industry. This analysis not only underscores the strengths and limitations of each clustering technique but also guides the selection of appropriate methods for similar datasets in the consumer behavior domain.**

I. Introduction

In today's data-driven world, understanding customer behavior is crucial for businesses, especially in the financial sector. Credit card usage data offers a wealth of information that can help in categorizing customers based on their spending and payment patterns. This categorization, or clustering, is valuable for personalized marketing, risk assessment, and enhancing customer service.

Our study focuses on a dataset of approximately 9,000 credit card holders, each characterized by 18 behavioral variables. The dataset is rich and multifaceted, capturing various aspects of credit card usage and user behavior. The challenge lies in effectively analyzing this high-dimensional data to identify distinct groups or clusters of customers who exhibit similar behavior.

To address this challenge, we employ several clustering algorithms: K-Means, DBSCAN, MiniSom, and Hierarchical Clustering. Each of these methods has its unique approach to grouping data. K-Means partitions users into a specified number of clusters, DBSCAN groups users based on density, MiniSom uses self-organizing maps to cluster data, and Hierarchical Clustering creates a tree of clusters. These methods are chosen for their wide use and effectiveness in different scenarios.

Before applying these clustering algorithms, we perform essential data preprocessing steps. These include normalizing the data to a common scale and handling missing values, ensuring that the subsequent analysis is robust and meaningful. Additionally, to better visualize and understand the complex, high-dimensional data, we use dimensionality reduction techniques, namely PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbor Embedding).

The effectiveness of each clustering approach is evaluated using standard metrics like the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. These metrics provide insights into the quality of the clusters formed by each method.

This report presents a comprehensive analysis of the credit card dataset using these clustering techniques, offering insights into customer segmentation and evaluating the performance of each method. Our findings aim to contribute to the understanding of customer behavior in the credit card industry and demonstrate the applicability and effectiveness of various clustering techniques in practical scenarios.

II. Methods

A. Data Preprocessing

The initial step in our analysis involved preprocessing the credit card dataset to ensure the quality and consistency of the data. This process included:

- **Handling Missing Values:** We identified and imputed missing values in the dataset. The missing values were filled with appropriate statistics, such as the mean or median, to maintain the integrity of the dataset.
- **Normalization:** Given the varying scales of the 18 behavioral variables, we applied normalization to standardize the data. This step is crucial to ensure that variables with larger scales do not unduly influence the clustering results [1].
- **Exclusion of Identifiers:** The *CUST_ID* field was removed from the dataset prior to clustering. As a unique identifier for each customer, it does not contain meaningful variance for clustering analysis and would only introduce bias into the model, leading to incorrect groupings.

B. Clustering Algorithms

We employed four different clustering algorithms, each with its approach to grouping data:

- **K-Means Clustering:** A popular partitioning method that divides the dataset into a predefined number of clusters. We experimented with different numbers of clusters and evaluated the results [1].
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** This algorithm groups points that are closely packed together and marks points in low-density regions as outliers [1].
- **MiniSom:** A type of Self-Organizing Map (SOM) that uses unsupervised learning to produce a low-dimensional representation of the input space of the training samples.
- **Hierarchical Clustering:** This method builds a hierarchy of clusters either through a bottom-up approach (agglomerative) or a top-down approach (divisive) [1].

C. Dimensionality Reduction Techniques

To aid in the visualization and interpretation of the high-dimensional data, we applied two dimensionality reduction techniques:

- **PCA (Principal Component Analysis):** A technique that reduces the dimensionality of the data by transforming the original variables into a new set of variables (principal components) that are orthogonal and capture the maximum variance in the data [1].
- **t-SNE (t-Distributed Stochastic Neighbor Embedding):** A non-linear dimensionality reduction technique particularly well-suited for embedding high-dimensional data into a space of two or three dimensions, which can then be visualized in scatter plots [2].

D. Evaluation Metrics

To assess the effectiveness of each clustering method, we used the following metrics:

- **Silhouette Score:** Measures how similar an object is to its own cluster compared to other clusters.
- **Calinski-Harabasz Index:** Evaluates the cluster validity based on the density and separation of the clusters [1].
- **Davies-Bouldin Index:** Reflects the average 'similarity' between clusters, where lower values indicate better clustering [1].

III. Results

This section presents the findings from the application of various clustering methods to the credit card dataset. The clustering results, along with the dimensionality reduction techniques applied, are visualized in Figure 1 using both PCA and t-SNE for 2D and 3D representations.

The dendrogram resulting from the Hierarchical Clustering is shown in Figure 2, providing a visual representation of the data's hierarchical structure.

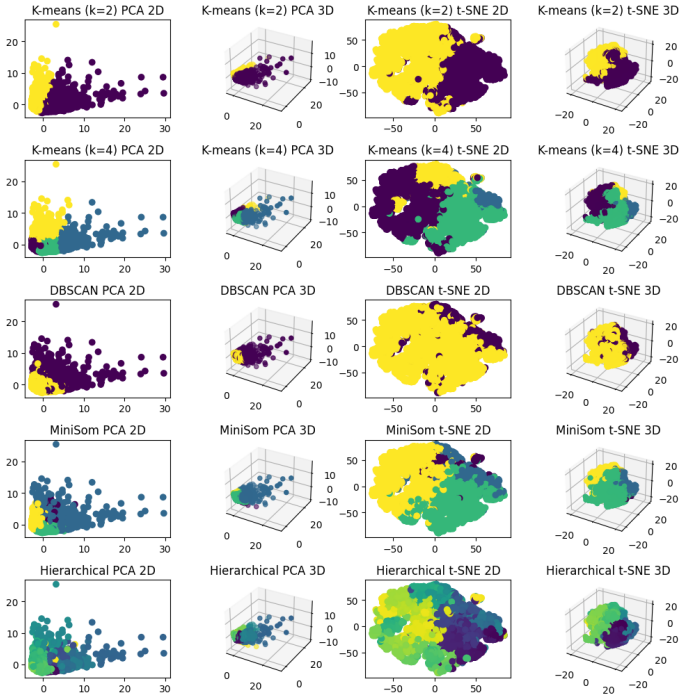

Fig. 1. Figure 1: Comparative Visualization of Clustering Algorithms Applied to Credit Card Data. The matrix displays K-Means, DBSCAN, MiniSom, and Hierarchical Clustering results. Each method is visualized in both 2D and 3D using PCA (left) and t-SNE (right) for dimensionality reduction. The distinct color palettes represent the unique clusters identified by each algorithm.

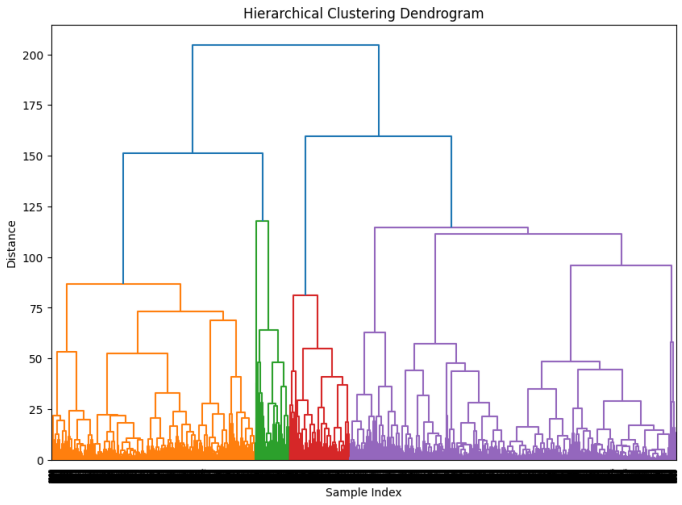

Fig. 2. Figure 2: Dendrogram from Hierarchical Clustering Analysis. The x-axis represents individual samples, while the y-axis shows the distance or dissimilarity between clusters. The color-coded lines reflect merged clusters at various distances, illustrating the hierarchical nature of the clustering process.

Each method's performance was evaluated using three metrics: the Silhouette Score, the Calinski-Harabasz Index, and the Davies-Bouldin Index. The evaluation results are as follows:

Algorithm	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-Means (k=2)	0.210	**1706.148**	1.912
K-Means (k=4)	0.198	1597.522	1.575
DBSCAN	**0.443**	1028.555	**2.183**
MiniSom	**0.142**	1367.577	1.773
Hierarchical	0.154	**139.866**	**0.866**

TABLE I
EVALUATION METRICS FOR CLUSTERING METHODS

IV. CONCLUSION

The evaluation of clustering methods on the credit card dataset has led to several noteworthy conclusions. The performance of each algorithm was measured using the **Silhouette Score**, **Calinski-Harabasz Index**, and **Davies-Bouldin Index**, with the best scores highlighted for each metric.

The **DBSCAN algorithm**'s highest Silhouette Score of **0.443** indicates its superior capability in creating clearly defined clusters. This score reflects the algorithm's effectiveness in grouping data points such that they are more closely packed within their clusters and well-separated from others, providing distinct customer segments [1].

The **K-Means algorithm**, when applied to create two clusters, achieved the highest Calinski-Harabasz Index of **1706.148**. This result suggests that K-Means was able to identify clusters with a high degree of internal homogeneity and external separation, making it a suitable method for partitioning the dataset into broader categories [1].

Furthermore, the **Hierarchical clustering** recorded the lowest Davies-Bouldin Index of **0.866**, indicative of well-separated clusters. This score demonstrates its potential in revealing more granular and distinct customer groups within the data.

In summary, the **DBSCAN algorithm** was found to be most effective for identifying distinct clusters with clear boundaries [3]. Meanwhile, **K-Means** demonstrated its versatility in achieving the best balance between cluster cohesion and separation with two clusters and distinguishing between clusters most effectively with four clusters.

These insights have significant implications for strategic decision-making in credit card customer management and targeted marketing. Recognizing the distinct behavior patterns within customer segments can lead to more personalized and effective customer engagement strategies.

For future work, it would be beneficial to explore other clustering algorithms and hybrid approaches that combine the strengths of the methods evaluated. Additionally, incorporating temporal data into the clustering process could yield dynamic customer segmentation that adapts over time.

REFERENCES

[1] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2nd ed., 2019.

[2] A. Ng, "Machine learning," 2021. Coursera course.

[3] L. Kaufman and P. J. Rousseeuw, "Finding groups in data: an introduction to cluster analysis," 2009.