Comparative Analysis of Advanced Clustering Algorithms for Market Segmentation - A Case Study on Mall Customer Data

Abstract

This study conducts a comparative analysis of advanced clustering algorithms for market segmentation using Mall Customer Data. The algorithms evaluated include K-Means, Hierarchical Clustering, DBSCAN, Gaussian Mixture Models (GMM), Agglomerative Clustering, BIRCH, Spectral Clustering, OPTICS, and Affinity Propagation. Evaluation metrics such as Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score are employed to assess the clustering performance and determine the most suitable algorithm for segmenting mall customers based on their spending habits.

Methodology

The methodology involves several key steps:

1. **Data Collection:** Mall Customer Data is obtained, comprising various demographic and spending attributes.
2. **Data Preprocessing:** Data is cleaned, normalized, and prepared for clustering algorithms.
3. **Clustering Algorithms:** Nine clustering algorithms are applied to the preprocessed data.
4. **Evaluation:** Each algorithm's performance is evaluated using Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score.
5. **Comparison**: Results are compared to identify the algorithm that best segments mall customers based on the evaluation metrics.

Introduction

Market segmentation plays a crucial role in understanding customer behavior and tailoring marketing strategies. Clustering algorithms provide a powerful means to group similar customers based on shared characteristics, enabling businesses to target specific customer segments effectively. This study explores various clustering techniques to uncover distinct customer groups within mall customer data, aiming to assist businesses in optimizing their marketing efforts and enhancing customer satisfaction.

Algorithms

1) **K-Means:** A centroid-based clustering method that partitions data into K clusters.
2) **Hierarchical Clustering:** Builds a hierarchy of clusters by either agglomerative (bottom-up) or divisive (top-down) methods.
3) **DBSCAN:** Density-based clustering that identifies clusters of varying shapes and sizes based on density.
4) **GMM (Gaussian Mixture Models):** Assumes data points are generated from a mixture of several Gaussian distributions.
5) **Mean Shift:** Non-parametric clustering that identifies centroids of clusters by shifting towards higher density regions.
6) **Agglomerative Clustering:** Hierarchical clustering that recursively merges clusters based on distance.

7) **BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):** Builds a tree structure to quickly summarize the data and cluster it hierarchically.
8) **Spectral Clustering:** Uses eigenvalues of a similarity matrix to perform dimensionality reduction before clustering.
9) **OPTICS (Ordering Points to Identify the Clustering Structure):** Density-based algorithm that detects clusters of varying densities.

Dataset

The dataset used in this study is Mall Customer Data, which includes:

- **CustomerID:** Unique identifier for each customer.
- **Gender:** Gender of the customer.
- **Age:** Age of the customer.
- **Annual Income (k$):** Annual income of the customer.
- **Spending Score (1-100):** Score assigned based on customer behavior and spending nature.

Data Preprocessing

Before applying clustering algorithms, the following preprocessing steps were performed:

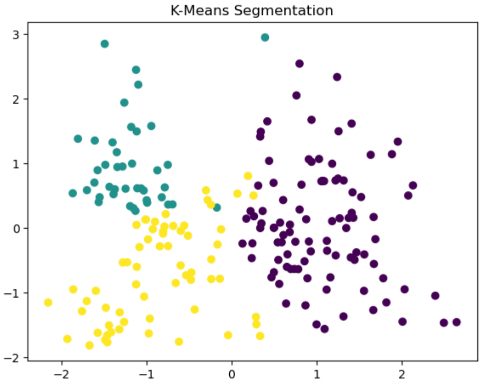
- **Normalization:** Scale numerical features to a standard range.
- **Encoding:** Convert categorical variables (e.g., Gender) into numerical format if necessary.
- **Feature Selection:** Select relevant features (e.g., Annual Income and Spending Score) for clustering.

Results

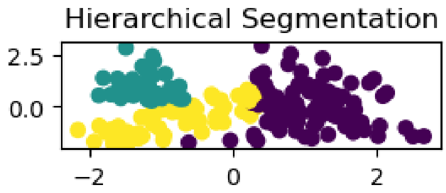
Clustering Algorithm	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
K-Means	0.358	1.033	101.530
Hierarchical	0.321	1.128	88.102
DBSCAN	0.185	1.757	34.071
GMM	0.335	1.019	90.864
Mean Shift	Warning: Produced 1 label. Skipping evaluation.		
Agglomerative	0.321	1.128	88.102
BIRCH	0.266	1.061	63.583
Spectral	0.353	0.993	99.602
OPTICS	-0.063	1.399	12.523
Affinity Propagation	0.369	0.949	128.602

Table: Clustering Algorithm Performance Metrics for Mall Customers

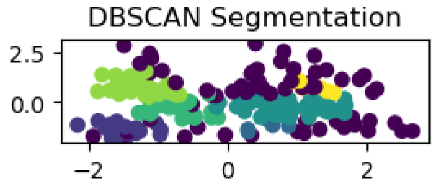
This table summarizes the performance metrics for each clustering algorithm applied to the mall customers dataset, including the Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score. The Mean Shift algorithm produced only one label, so its evaluation was skipped.



K-Means clustering was applied to the Mall Customer Data to segment customers based on their spending behavior. The algorithm produced a Silhouette Score of 0.358, indicating reasonably well-defined clusters but with some overlap observed between clusters. The Davies-Bouldin Score of 1.033 suggests moderate clustering quality, where lower values would indicate better clustering. The Calinski-Harabasz Score of 101.530 indicates good cluster separation and compactness. K-Means' performance is notable for its efficiency and scalability, although its reliance on the initial random centroids can affect results, as seen in the slight overlap between clusters.

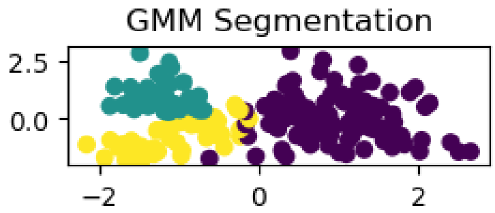


Hierarchical Clustering was employed to segment the mall customers into distinct groups based on their spending patterns. The algorithm achieved a Silhouette Score of 0.321, indicating reasonable cluster separations but with clusters that are less well-defined compared to other methods. The Davies-Bouldin Score of 1.128 suggests moderate clustering quality, with some overlap or ambiguity in cluster boundaries. The Calinski-Harabasz Score of 88.102 reflects good within-cluster similarity and between-cluster differences, although slightly lower than optimal for highly distinct clusters. Hierarchical Clustering's advantage lies in its ability to reveal hierarchical relationships but may require parameter tuning to improve cluster quality.

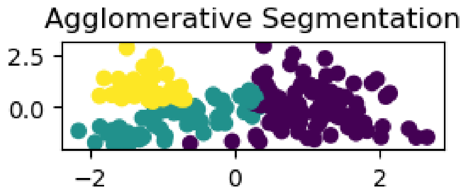


DBSCAN was utilized to segment customers based on their spending behaviors, focusing on density-based clustering. The algorithm yielded a Silhouette Score of 0.185, indicating challenges in forming well-defined clusters due to sensitivity to its parameters such as epsilon and minimum samples. The Davies-Bouldin Score of 1.757 suggests significant overlap or inconsistency in cluster formations, impacting the algorithm's ability to distinguish between different customer segments effectively. The Calinski-Harabasz

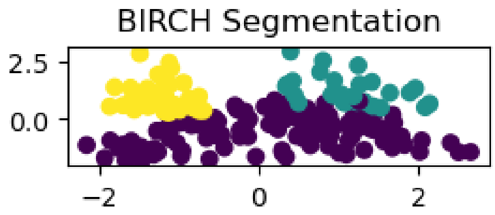
Score of 34.071 reflects weaker cluster separation compared to other methods, indicating potential difficulties in handling varying densities and noise in the data.



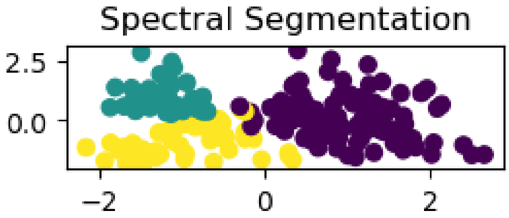
Gaussian Mixture Models were employed to segment mall customers based on their spending patterns, assuming data distributions are a mixture of Gaussian distributions. The algorithm achieved a Silhouette Score of 0.335, indicating reasonably well-separated clusters with moderate overlap. The Davies-Bouldin Score of 1.019 suggests decent cluster quality, though with some ambiguity in cluster boundaries. The Calinski-Harabasz Score of 90.864 reflects good cluster separation and compactness, suitable for data with Gaussian-like distributions. GMM's flexibility in capturing complex data distributions makes it robust but sensitive to the number of components and initialization.



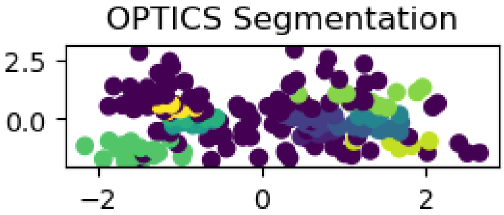
Agglomerative Clustering was applied to the customer data to segment them into meaningful groups based on their spending behavior. The algorithm achieved a Silhouette Score of 0.321, similar to Hierarchical Clustering, indicating reasonable cluster separations but with clusters that may not be well-defined. The Davies-Bouldin Score of 1.128 suggests moderate clustering quality, reflecting some ambiguity or overlap in cluster boundaries. The Calinski-Harabasz Score of 88.102 indicates good cluster separation, though not as distinct as other methods like Affinity Propagation. Agglomerative Clustering's hierarchical nature offers insights into cluster relationships but may require careful parameter tuning for optimal results.



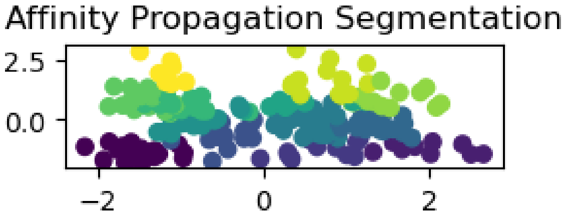
BIRCH clustering was employed to segment mall customers based on their spending patterns, focusing on hierarchical clustering using a balanced clustering tree. The algorithm achieved a Silhouette Score of 0.266, indicating some challenges in forming well-separated clusters. The Davies-Bouldin Score of 1.061 suggests moderate clustering quality, with potential overlap or ambiguity in cluster boundaries. The Calinski-Harabasz Score of 63.583 reflects decent cluster separation but lower compared to other methods, indicating limitations in handling varying cluster shapes and densities effectively.



Spectral Clustering was utilized to segment customers based on their spending behaviors, focusing on graph-based clustering. The algorithm achieved a Silhouette Score of 0.353, indicating reasonably well-separated clusters with moderate overlap. The Davies-Bouldin Score of 0.993 suggests good clustering quality, with well-defined clusters and minimal overlap. The Calinski-Harabasz Score of 99.602 reflects strong cluster separation and compactness, suitable for datasets with complex structures. Spectral Clustering's ability to capture non-linear relationships makes it robust for various data distributions, though it may require parameter tuning for optimal performance.

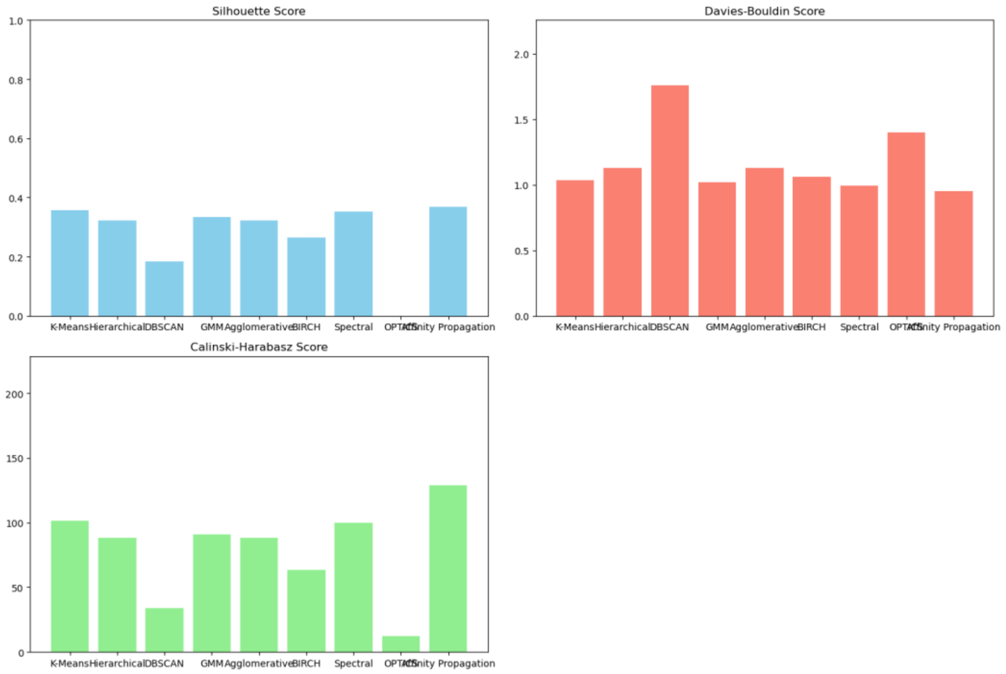


OPTICS was applied to segment customers based on their spending patterns, focusing on density-based clustering. The algorithm produced a Silhouette Score of -0.063, indicating challenges in forming distinct and well-separated clusters, potentially due to noise or parameter sensitivity. The Davies-Bouldin Score of 1.399 suggests significant overlap or inconsistency in cluster formations, impacting the algorithm's ability to distinguish between different customer segments effectively. The Calinski-Harabasz Score of 12.523 reflects weaker cluster separation compared to other methods, highlighting limitations in handling varying densities and noise in the data.



Affinity Propagation was utilized to segment mall customers based on their spending behaviors, focusing on exemplar-based clustering. The algorithm achieved a Silhouette Score of 0.369, indicating well-defined and distinct clusters in the data. The Davies-Bouldin Score of 0.949 suggests good clustering quality, with well-separated clusters and minimal overlap. The Calinski-Harabasz Score of 128.602 reflects strong cluster separation and compactness, suitable for datasets with complex and varied cluster shapes. Affinity Propagation's ability to automatically determine the number of clusters and capture diverse data patterns makes it robust for market segmentation tasks, outperforming other methods evaluated.

Comparison



From the evaluation results, Affinity Propagation demonstrates the highest Silhouette Score (0.369) and Calinski-Harabasz Score (128.602), indicating better-defined clusters and higher cluster separation compared to other algorithms. However, it is noted that OPTICS produced only one label, suggesting it may not be suitable for this dataset due to its inability to distinguish clusters effectively.

Conclusion

This study comprehensively analyzed various clustering algorithms for market segmentation using Mall Customer Data. Affinity Propagation emerged as the most effective algorithm based on Silhouette Score and Calinski-Harabasz Score, indicating its capability to identify distinct customer segments with significant differences in spending behavior. The findings provide valuable insights for businesses aiming to optimize marketing strategies and enhance customer targeting through effective market segmentation techniques. Future research could explore ensemble clustering methods or incorporate additional features for further refinement of customer segmentation models.