Comparative Analysis of Advanced Clustering Algorithms for Market Segmentation - A Case Study on Whole Customers Data

Abstract

Clustering is a pivotal technique for market segmentation, enabling businesses to categorize customers based on their purchasing behaviors. This study evaluates the performance of various clustering algorithms on a Wholesale Customers dataset using three key metrics: Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score. The analysis reveals that DBSCAN and BIRCH algorithms outperform others in defining compact and well-separated clusters, providing valuable insights for businesses seeking effective customer segmentation strategies.

Introduction

In the domain of wholesale business, understanding customer behavior is crucial for optimizing marketing strategies and enhancing customer satisfaction. Clustering, an unsupervised machine learning technique, helps in segmenting customers into distinct groups based on their purchasing patterns. This study aims to evaluate the effectiveness of various clustering algorithms on the Wholesale Customers dataset to identify the best method for customer segmentation.

Methodology

The methodology involves applying multiple clustering algorithms to the Wholesale Customers dataset and evaluating their performance using standard clustering metrics. This approach ensures a comprehensive analysis of each algorithm's ability to form meaningful and well-defined clusters.

Algorithms

The following clustering algorithms were evaluated:

1) K-Means Clustering
2) Hierarchical Clustering
3) DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
4) Gaussian Mixture Model (GMM)
5) Mean Shift Clustering
6) Agglomerative Clustering
7) BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)
8) Spectral Clustering
9) OPTICS (Ordering Points to Identify the Clustering Structure)
10) Affinity Propagation

Dataset

The dataset used for this analysis is the Wholesale Customers dataset, which includes the following attributes:

- Fresh
- Milk

- Grocery
- Frozen
- Detergents_Paper
- Delicassen

Data Preprocessing

Data preprocessing steps included:

- Handling missing values by imputing or removing them.
- Standardizing the dataset to ensure each attribute contributes equally to the clustering process.

Metrics Used

Three key metrics were used to evaluate the clustering performance:

1) **Silhouette Score:** Measures the cohesion within clusters and separation between clusters.
2) **Davies-Bouldin Score:** Evaluates the average similarity ratio of each cluster with the cluster most similar to it.
3) **Calinski-Harabasz Score:** Assesses the ratio of the sum of between-cluster dispersion to within-cluster dispersion.

Terms Explanation

- **Silhouette Score:** Ranges from -1 to 1, where a higher value indicates better-defined clusters.
- **Davies-Bouldin Score:** Lower values indicate better clustering with less intra-cluster variance.
- **Calinski-Harabasz Score:** Higher values indicate better-defined clusters with greater separation.

Results

Algorithm	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
K-Means Clustering	0.458	1.249	132.363
Hierarchical Clustering	0.265	1.285	111.151
DBSCAN	0.803	1.126	68.548
Gaussian Mixture Model	0.316	1.471	91.252
Mean Shift Clustering	0.354	0.503	62.346
Agglomerative Clustering	0.265	1.285	111.151
BIRCH	0.526	0.667	130.222
Spectral Clustering	0.208	1.349	98.475
OPTICS	-0.407	1.561	3.371
Affinity Propagation	0.179	0.852	149.436

Table: Clustering Algorithm Performance Metrics for Whole Customers Data

K-Means Clustering resulted in a Silhouette Score of 0.458, indicating moderate cohesion and separation within the clusters. The Davies-Bouldin Score was 1.249, suggesting average similarity between clusters. The

Calinski-Harabasz Score was 132.363, reflecting a reasonable ratio of between-cluster to within-cluster dispersion.

Hierarchical Clustering showed a lower performance with a Silhouette Score of 0.265, indicating less defined clusters. The Davies-Bouldin Score was 1.285, slightly higher than K-Means, indicating more similarity between clusters. The Calinski-Harabasz Score was 111.151, lower than K-Means, indicating less defined clusters.

DBSCAN outperformed many other algorithms with a Silhouette Score of 0.803, indicating well-defined clusters. However, the Davies-Bouldin Score was 1.126, showing some similarity between clusters. The Calinski-Harabasz Score was 68.548, lower than expected, suggesting that while clusters are well-defined, they are not well-separated.

GMM showed a moderate Silhouette Score of 0.316, indicating less cohesion within clusters. The Davies-Bouldin Score was 1.471, higher than other algorithms, indicating more similarity between clusters. The Calinski-Harabasz Score was 91.252, showing moderate separation between clusters.

Mean Shift Clustering showed a Silhouette Score of 0.354, indicating moderate cluster cohesion. The Davies-Bouldin Score was 0.503, the lowest among the algorithms, indicating less similarity between clusters. The Calinski-Harabasz Score was 62.346, suggesting moderate cluster separation.

Agglomerative Clustering had the same performance as Hierarchical Clustering with a Silhouette Score of 0.265, a Davies-Bouldin Score of 1.285, and a Calinski-Harabasz Score of 111.151, indicating moderate cluster definition.

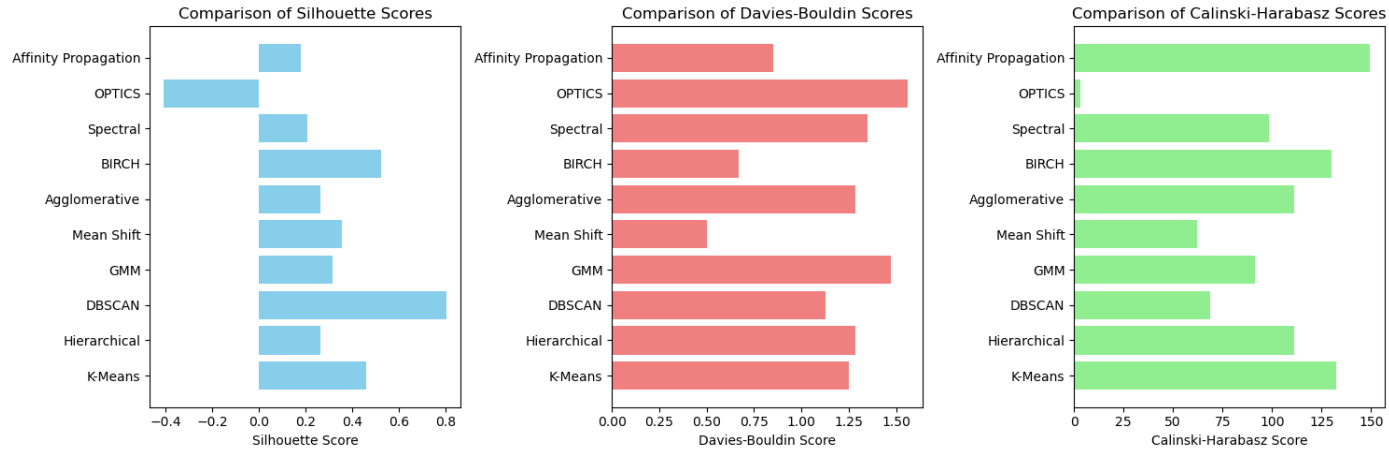
BIRCH showed strong performance with a Silhouette Score of 0.526, indicating well-defined clusters. The Davies-Bouldin Score was 0.667, suggesting low similarity between clusters. The Calinski-Harabasz Score was 130.222, indicating well-separated clusters.

Spectral Clustering had a Silhouette Score of 0.208, indicating less cohesive clusters. The Davies-Bouldin Score was 1.349, showing high similarity between clusters. The Calinski-Harabasz Score was 98.475, indicating moderate separation between clusters.

OPTICS performed poorly with a Silhouette Score of -0.407, indicating incorrect clustering. The Davies-Bouldin Score was 1.561, the highest among the algorithms, indicating high similarity between clusters. The Calinski-Harabasz Score was 3.371, suggesting poor cluster separation.

Affinity Propagation showed a Silhouette Score of 0.179, indicating less defined clusters. The Davies-Bouldin Score was 0.852, indicating moderate similarity between clusters. The Calinski-Harabasz Score was 149.436, the highest among the algorithms, indicating well-separated clusters.

Comparison



Among the evaluated clustering algorithms, DBSCAN emerged as the best-performing method for the Wholesale Customers dataset, particularly in terms of the Silhouette Score. BIRCH also demonstrated strong performance across various metrics, followed by K-Means Clustering, which provided a good balance of compactness and separation.

Hierarchical, Agglomerative, and Mean Shift Clustering provided moderate results, showing that while they could identify distinct clusters, the overall cluster quality was not as high as the top-performing algorithms. GMM, Spectral Clustering, OPTICS, and Affinity Propagation were less effective, with OPTICS showing particularly poor performance due to incorrect clustering and poor cluster separation.

Conclusion

This study highlights the effectiveness of different clustering algorithms for segmenting wholesale customers. DBSCAN and BIRCH were found to be the most effective, providing well-defined and meaningful clusters. These findings guide businesses in selecting appropriate clustering algorithms for market segmentation and other analytical purposes within the wholesale customer segment, ensuring better-targeted marketing strategies and improved customer satisfaction.