Customer Segmentation Using K-Means Clustering: A KDD-Based Approach

Aditya Rajpurohit

October 10, 2024

Abstract

This research paper explores the use of the Knowledge Discovery in Databases (KDD) process for customer segmentation, employing K-Means clustering on a mall customer dataset. The study demonstrates how the KDD process can be used to extract meaningful insights from data, facilitating targeted marketing and personalized services.

1 Introduction

Customer segmentation is the process of dividing customers into distinct groups based on shared characteristics. This approach helps businesses better understand their customer base, enabling more effective marketing strategies. In this research, we utilize the KDD process to uncover customer segments using K-Means clustering. The KDD process consists of five key steps: data selection, preprocessing, transformation, data mining, and interpretation.

2 Methodology

2.1 Knowledge Discovery in Databases (KDD)

The KDD process is a systematic method for discovering patterns from data. It involves the following steps:

1. **Data Selection:** Identify and select relevant data.

2. **Data Preprocessing:** Clean and prepare data for analysis.
3. **Data Transformation:** Transform data for better suitability.
4. **Data Mining:** Apply algorithms (e.g., K-Means) to extract patterns.
5. **Interpretation:** Evaluate and interpret the results.

2.2 Dataset Description

The dataset used in this research consists of 200 mall customers with attributes such as:

- **Age:** Customer age in years.
- **Annual Income:** Annual income in thousand dollars.
- **Spending Score:** Spending score (1-100) assigned by the mall.

3 Data Selection

In the initial step, the dataset is loaded into a pandas DataFrame, and the relevant features are identified for clustering. The selected features are: *Age*, *Annual Income*, and *Spending Score*.

4 Data Preprocessing

The dataset is checked for missing values and duplicates, which are removed to ensure data quality. The preprocessing step ensures that the dataset is clean and ready for analysis. This is achieved through the following code:

```
import pandas as pd
df = pd.read_csv('Mall_Customers.csv')
df = df.dropna().drop_duplicates()
```

5 Data Transformation

To ensure that features contribute equally to the clustering process, the data is standardized. We use the StandardScaler to scale the features:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df[['Age', 'Annual Income (k$)', 'Spending Score
```

6 Data Mining

6.1 Elbow Method

To determine the optimal number of clusters, the Elbow Method is applied. The Within-Cluster Sum of Squares (WCSS) is calculated for different numbers of clusters, and the results are plotted.

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, randc
    kmeans.fit(df_scaled)
    wcss.append(kmeans.inertia_)

plt.plot(range(1, 11), wcss, marker='o')
plt.title('Elbow Method')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.show()
```

6.2 K-Means Clustering

Based on the Elbow Method, K-Means clustering is applied with the optimal number of clusters.

```
optimal_clusters = 4
```

```
kmeans = KMeans(n_clusters=optimal_clusters, init='k-means++', max_iter=300, n_ini
clusters = kmeans.fit_predict(df_scaled)
df['Cluster'] = clusters
```

7 Results and Interpretation

The results are visualized using a scatter plot to interpret the clusters, with each cluster representing a distinct customer segment:

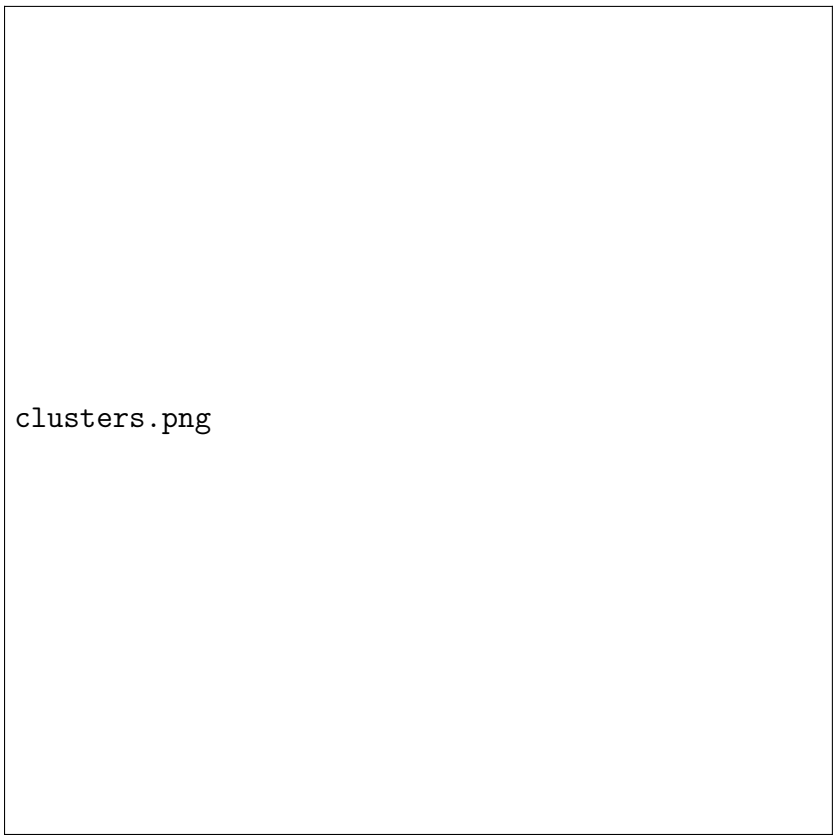

Figure 1: Visualization of Customer Clusters

The interpretation of the clusters is as follows:

- **Cluster 1:** High income, high spending customers.
- **Cluster 2:** Low income, low spending customers.

- **Cluster 3:** Moderate income, moderate spending customers.
- **Cluster 4:** High income, low spending customers.

8 Conclusion

This research demonstrates how the KDD process can be effectively used to perform customer segmentation with K-Means clustering. By following each step systematically, we were able to extract meaningful insights, helping businesses target specific customer segments for improved marketing strategies.

9 Future Work

Future studies can extend this analysis by using other clustering techniques like hierarchical clustering or DBSCAN to compare results. Additionally, incorporating demographic factors like gender or location could provide deeper insights into customer behavior.

References

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.

[2] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.