Performance Assessment: CLUSTERING TECHNIQUES

Gabriela Howell

Master of Science Data Analytics, Western Governors University

D212 – CLUSTERING TECHNIQUES

Professor Middleton

August 10, 2024

Part I: Research Question

A. Describe the purpose of your data mining report by doing the following:

1. Propose one question relevant to a real-world organizational situation that you will answer using one of the following clustering techniques:

The relevant question for a real-world organizational situation is, “What variables can be identified to form meaningful clusters in the given dataset?”. I will use the k-means clustering technique for this analysis, where I will concentrate exclusively on continuous variables within the provided medical dataset.

2. Define one goal of the data analysis. Ensure your goal is reasonable within the scope of the selected scenario and is represented in the available data.

The primary goal of this data analysis is to identify unique clusters within the medical dataset that can be grouped significantly. By doing so, the medical industry can better recognize, and address concerns related to patient readmissions.

Part II: Technique Justification

B. Explain the reasons for your chosen clustering technique from part A1 by doing the following: 1. Explain how the clustering technique you chose analyzes the selected data set. Include expected outcomes.

The clustering technique I am using in this analysis is K-means clustering, chosen for its ease of use, speed, and ability to produce interpretable results. K-means works by sorting data points into clusters based on their similarities. The algorithm operates iteratively, beginning with the initialization of cluster centroids. Initially, "the model picks up K datapoints from the dataset" as cluster centroids, and then "calculates the distance between the datapoint & all the centroids," assigning each data point to the nearest cluster. This process is repeated until the centroids stabilize, ensuring the algorithm converges to a solution (Towards Data Science, 2023).

To ensure the data is consistent, I standardized features such as Age, Income, VitD_levels, and TotalCharge to ensure equal contribution during clustering. In my analysis, I expect that K-means will create distinct clusters that represent different patient groups with similar Age and TotalCharge.

2. Summarize one assumption of the clustering technique.

According to IBM, one assumption of the k-means clustering technique is that the data points are clear clusters around centroids in spherical shapes. This assumption suggests the algorithm works best when clusters are well-separated and have similar variances (IBM, 2024). According to the scikit-learn documentation, k-means clustering minimizes variance within each cluster, emphasizing the importance of spherical clusters with equal variance (Scikit).

3. List the packages or libraries you have chosen for Python or R, and justify how *each* item on the list supports the analysis.

I will be using Python and the libraries I will be using are:

- **Pandas:** Offers effective data analysis tools for data manipulation and preparation.
- **Seaborn:** Offers attractive visualizing of data distributions and relationships.
- **NumPy:** Manages big arrays and tables with many dimensions, and has some math tools.
- **Matplotlib:** Allows the creation of quality plots for numerical data.
- **Scikit-learn:** Contains simple and efficient tools for predictive data analysis, including KMeans for clustering, silhouette_score for evaluating clusters, and StandardScaler for data normalization.

With these libraries, K-means clustering can thrive.

Part III: Data Preparation

C. Perform data preparation for the chosen data set by doing the following:

1. Describe one data preprocessing goal relevant to the clustering technique from part A1.

My data preprocessing goal for the K-means clustering technique is to ensure standardization of the features. Since K-means relies on Euclidean distance, which can be greatly persuaded by the scale of the data, standardizing warrants that all features are influenced equally by the clustering process (Scikit).

2. Identify the initial data set variables you will use to perform the analysis for the clustering question from part A1, and label *each* as continuous or categorical.

The initial data set variables used for the analysis are ‘Age’ (continuous), ‘Income’ (continuous), ‘VitD_levels’ (continuous), and ‘TotalCharge’ (continuous). As you can see all the variables are continuous data which follows the requirements for K-mean.

3. Explain *each* of the steps used to prepare the data for the analysis. Identify the code segment for *each* step.

To prepare the data, I performed several steps. First, I verified that there were no missing values, and any outliers were taken care of to ensure the data was complete. Next, I standardized the data using the StandardScaler from sklearn, which scales each feature to have a mean of zero and a standard deviation of one. This step is fundamental for the K-means algorithm, as it ensures all features are on the same scale. Finally, I confirmed the data was ready for clustering by examining the summary statistics to ensure proper scaling and the absence of missing values. The cleaned data set, now standardized, is ready for clustering.

4. Provide a copy of the cleaned data set.

The clean data set is called, ‘kmeans_medical.csv’.

Part IV: Analysis

D. Perform the data analysis, and report on the results by doing the following:

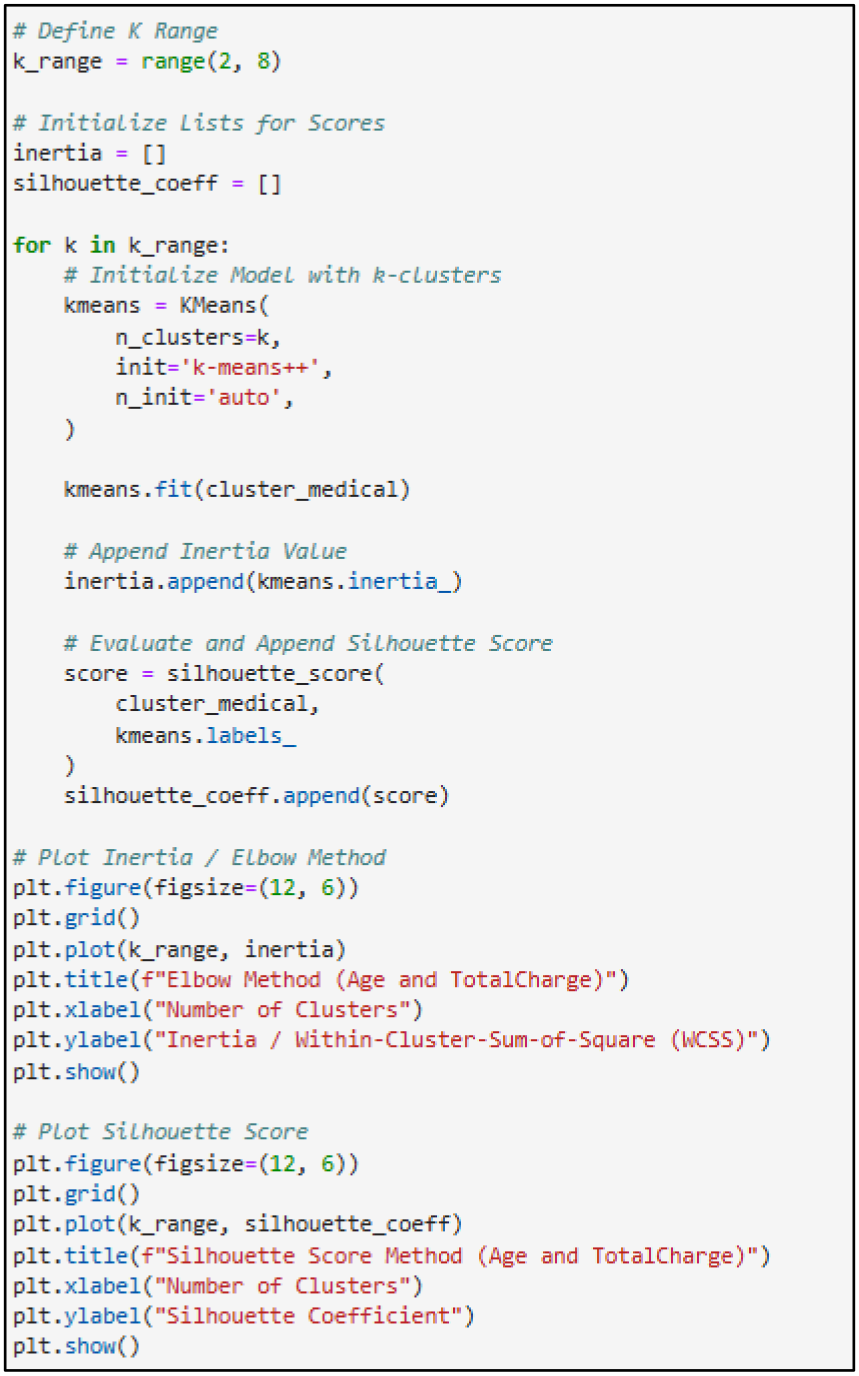
1. Determine the optimal number of clusters in the data set, and describe the method used to determine this number.

To find the best number of clusters, I applied two different techniques to ensure I was confident in choosing the correct number of clusters. The two main techniques I used were the Elbow Method and the Silhouette Score Method:

- Elbow Method: This technique involves plotting the within-cluster sum of squares (inertia) against the number of clusters, (which is known as ‘k’) (Sarah, 2024). The optimal number of clusters is identified at the "elbow" point, where the inertia rate starts to slowly decrease.
- Silhouette Score Method: This technique calculates the silhouette score for different numbers of clusters. The silhouette score measures how similar an object is to its cluster compared to other clusters (Mazzanti, 2023). A higher silhouette score indicates better-defined clusters. The optimal number of clusters is where the silhouette score is highest.

When reviewing both techniques, I found that the Silhouette Score as a secondary evaluation metric. The Elbow Method graph indicates a distinct "elbow" at 4 clusters, while the Silhouette Score method shows an increase at 5 clusters before decreasing again. Hence, 4 is the determined as the optimal number of clusters for this dataset.

2. Provide the code used to perform the clustering analysis technique.



```
# Define K Range
k_range = range(2, 8)

# Initialize Lists for Scores
inertia = []
silhouette_coeff = []

for k in k_range:
    # Initialize Model with k-clusters
    kmeans = KMeans(
        n_clusters=k,
        init='k-means++',
        n_init='auto',
    )

    kmeans.fit(cluster_medical)

    # Append Inertia Value
    inertia.append(kmeans.inertia_)

    # Evaluate and Append Silhouette Score
    score = silhouette_score(
        cluster_medical,
        kmeans.labels_
    )
    silhouette_coeff.append(score)

# Plot Inertia / Elbow Method
plt.figure(figsize=(12, 6))
plt.grid()
plt.plot(k_range, inertia)
plt.title(f"Elbow Method (Age and TotalCharge)")
plt.xlabel("Number of Clusters")
plt.ylabel("Inertia / Within-Cluster-Sum-of-Square (WCSS)")
plt.show()

# Plot Silhouette Score
plt.figure(figsize=(12, 6))
plt.grid()
plt.plot(k_range, silhouette_coeff)
plt.title(f"Silhouette Score Method (Age and TotalCharge)")
plt.xlabel("Number of Clusters")
plt.ylabel("Silhouette Coefficient")
plt.show()
```

In addition to the screenshot, the Python code file titled "Gab_D212_Task1_v2" will be provided.

Part V: Data Summary and Implications

E. Summarize your data analysis by doing the following:

1. Explain the quality of the clusters created.

To ensure the quality of the clusters, I evaluated them using a silhouette score, which measures how similar each data point is to its cluster compared to other clusters (Mazzanti, 2023). So, a higher silhouette score implies better-defined clusters. From my analysis, the optimal number of clusters in the data is four, as indicated by a high silhouette score of 0.66 to support that number. This score suggests that the clusters are relatively well-defined, with clear distinctions between them. However, it also shows there is some intersection in the clustering boundaries.

2. Discuss the results and implications of your clustering analysis.

The clustering analysis resulted in four distinct groups based on the features ‘Age’ and ‘TotalCharge’. Each cluster has its mean values for these features, indicating different patterns within the data:

- Cluster 0: Characterized by lower-than-average ‘Age’ and lower ‘TotalCharge’.
- Cluster 1: Represents lower-than-average ‘Age’ with higher-than-average ‘TotalCharge’.
- Cluster 2: Displays higher-than-average ‘Age’ with lower-than-average ‘TotalCharge’.
- Cluster 3: Shows higher-than-average values for both ‘Age’ and ‘TotalCharge’.

The identification of unique patient profiles through these clusters can aid in the implementation of focused medical treatments and the strategic distribution of resources. The results imply that the data contains meaningful groupings that can be leveraged for more tailored healthcare strategies.

3. Discuss one limitation of your data analysis.

One limitation of this data analysis is the dependence on just two features, Age and TotalCharge, for clustering. While these features provide valuable insights, they don't show everything. The correlation could have been stronger, but the selection of features was limited. Using more relevant features could possibly lead to a deeper comprehensive clustering and insights.

4. Recommend a course of action for the real-world organizational situation from part A1 based on the results and implications discussed in part E2.

The insights resulting from the clustering analysis recommend that the organization should use the distinct clusters to personalize their health services. For instance, patients in clusters with higher 'TotalCharge' might benefit from cost management programs or more efficient care plans. Also, those with lower 'Age' could enhance their overall health outcomes with changed habits. In acknowledging and responding to the unique characteristics of each cluster, the organization has the opportunity to improve patient care results and make more efficient use of resources.

Part VI: Demonstration

F. Provide a Panopto video recording

The Panopto video link : D212_Task1_GH

G. Record the web sources you used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

Demonstration of K-means assumptions. scikit. (n.d.-c). https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html

What is K-means clustering?. IBM. (2024, June 27). https://www.ibm.com/topics/k-means-clustering

H. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

Demonstration of K-means assumptions. scikit. (n.d.-c). https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html

Mazzanti, S. (2023, September 15). *Are you still using the elbow method?*. Medium. https://towardsdatascience.com/are-you-still-using-the-elbow-method-5d271b3063bd

Sarah, M. (2024, January 2). *A comprehensive guide to cluster analysis: Applications, best practices and resources*. Displayr. https://www.displayr.com/understanding-cluster-analysis-a-comprehensive-guide/

Towards Data Science. (2023). *K-Means Clustering: From A to Z*. Retrieved from https://towardsdatascience.com/k-means-clustering-from-a-to-z-f6242a314e9a

What is K-means clustering?. IBM. (2024, June 27). https://www.ibm.com/topics/k-means-clustering