Clustering Assignment Part-II

By Subramanya Nayak

Question 1: Assignment Summary Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly (what EDA you performed, which type of Clustering produced a better result and so on).

Ans: Our main objective for this assignment is to find the countries that are in direst need of aid. Our job is to find those countries using socio-economic and heath factors which will show overall development of the country.

So, in order to do so first I analyze the dataset – It has total 167 countries with no missing values. Then I visualized the data with univariate analysis and we looked on the lowest 10 countries by each factor and then I visualized the data by finding the correlation between variables to check the multicollinearity. After that I checked for the outliers and there were outliers in all the variables of data so for better result, I treated that outliers. I checked this by describing the dataset in percentiles and by plotting boxplot. Now, it's time to do scaling the data for the KMeans and Hierarchical clustering.

So, first, I applied kmeans clustering for finding the optimal numbers of k with elbow curve and silhouette curve and found k=3 with the help of silhouette analysis. Then I initializing the Kmeans after that kmeans visualizations for better understanding the data, in which I used scatter plot as mention in problem statement. After that I checked our main columns 'income', 'child_mort' and 'gdpp' with cluster_id and did with the boxplot for outliers. So, in last I found the top 10 countries that are in direst need of aid.

After that I applied Hierarchical clustering, first I applied single linkage in which dendrogram was not clear to cut the dendrogram at appropriate number of clusters. So then I used complete linkage to decide the threshold value to cut the dengrogram there also I found k=3 from business prospective. after that kmeans visualizations for better understanding the data, in which I used scatter plot as mention in problem statement. After that I checked our main columns 'income', 'child_mort' and 'gdpp' with cluster_id and did with the boxplots for outliers. So, in last I found the top 10 countries that are in direst need of aid. The countries which is same as kmeans clustering and found the same results.

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

b) Briefly explain the steps of the K-means clustering algorithm.

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

d) Explain the necessity for scaling/standardization before performing Clustering.

e) Explain the different linkages used in Hierarchical Clustering.

Ans (a): Compare and contrast K-means Clustering and Hierarchical Clustering.

Kmeans Clustering:

- We need to have desired number of clusters ahead of time.
- It is a collection of data points in one cluster which are similar between them and not similar data points belongs to another cluster. Clusters have tree like structures and most similar clusters are first combine which continues until we reach a single branch.
- Works very good in large dataset.
- The main drawback of k-Means is it doesn't evaluate properly outliers.
- K-means only used for numerical.

Hierarchical Clustering:

- We can decide the number of clusters after completion of plotting dendrogram by cutting the dendrogram at different heights.
- Clusters have tree like structures and most similar clusters are first combine which continues until we reach a single branch.
- Works well in small dataset and not good with large dataset.
- Outliers are properly explained in hierarchical clustering.
- Hierarchical clustering is used when we have variety of data as it doesn't require to calculate any distance.

Ans (b): Briefly explain the steps of the K-means clustering algorithm.

1) Randomly select 'k' cluster centres.

2) Calculate the euclidean distance between each data point and cluster centres.

3) Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres.

4) Recalculate the new cluster centre.

5) Recalculate the distance between each data point and new obtained cluster centres.

6) If no data point was reassigned then stop, otherwise repeat from step.

Ans (c): How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. 'K' value is chosen randomly in K-Means clustering based on statistical aspect. From business aspect, we need to first understand the dataset and based on that we decide number of 'k'.

Ans(d): Explain the necessity for scaling/standardization before performing Clustering

Data standardization is about making sure that data is internally consistent; that is, each data type has the same content and format. Standardized values are useful for tracking data that isn't easy to compare otherwise. it will create a big difference as the clusters will tend to move with the variables having greater values or variances. By applying standardisation/scaling will increase the performance of our model.

Ans(e): Explain the different linkages used in Hierarchical Clustering.

Linkage is a technique used in Agglomerative Clustering. Linkage helps us to merge two data points into one using below linkage technique.

Single linkage:

The distance between two clusters is calculated by the minimum distance between two points from each cluster.

Complete linkage:

The distance between two clusters is calculated by the maximum distance between two points from each cluster.

Average linkage:

The distance between two clusters is the average distance between every point of one cluster to the another every point of other cluster.