Assignment II: Clustering and PCA

SMRUTI RANJAN PARIDA

Assignment Summary

Problem Statement:

- An International humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.
- After the recent funding programmes, they have been able to raise around 10 million USD. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.
- The Objective is to cluster the countries by the socio-economic and health factors and determine the overall development of the country.
- **Available data :**
  - datasets containing socio-economic factors for 167 countries.

Solution Approach

- The data set contains data for **167 countries with 9 Socio Economic parameters**.
- After EDA we could see outlier in some of the data sets and the outlier treatment was done using IQR which resulted in **164 countries** data.
- Heat Map of these parameters suggested that there is high co-relation between the parameters in the data set. So PCA was considered for reducing the dimension and capturing variance in data.
- **Standardization** was performed on the dataset as the parameters are in **different units of measurement** and the variance in the datasets were High.
- PCA was performed on the data set to create new dimensions of the data set.
- **5 PCA components** were considered for clustering as the Scree suggested that **95% of the variability** can be explained by these 5 components.
- Clustering was performed on the datasets using "**Kmeans clustering**" with **k = 3** as inferred from the **Elbow Curve** and **Silhouette score**.
- Clustering was performed on the datasets using "**Hierarchical clustering**" and from the dendrogram we can see 3 distinct clusters with similar clustering of countries shown by Kmeans method.
- As one of the cluster was for the countries with **Low GDP, Low Income and High Child Mortality** the same was selected as the countries which in direst need of aid.
- As the no of countries were close to 30, the final countries were selected which were having the lowest GDP and Highest Child Mortality.

Clustering

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

Kmeans Clustering

- K- means is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.
- we have to define the number of clusters to be created beforehand
- Hierarchical is Flexible but can not be used on large data
- K-means clustering is usually more efficient run-time wise
- The Centroids for different clusters are easy to understand and explain.

Hierarchical Clustering

- Clusters have a tree like structure or a parent child relationship. Here, the two most similar clusters are combined together and continue to combine until all objects are in the same cluster
- Data is automatically formed into a tree shape form (dendrogram) and we can chose which trees are significant.
- Hierarchical is Flexible but can not be used on large data
- Hierarchical clustering can be slow (has to make several merge/split decisions)
- Hierarchical clustering gives deep insight of each step of converging different clusters and create dendrogram which helps you figure out which clusters combination makes sense

Clustering

- Steps K-means clustering Algorithm:

  Step 1: The First step in Kmeans clustering is to specify the number of clusters *K*.

  Step 2: Initialize centroids randomly by selecting *K* data points for the centroids.

  Step 3: The following sub- steps are done in step3

  - Compute the sum of the squared distance between data points and all centroids.
  - Assign each data point to the closest cluster (centroid).
  - Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster

  Step 4: Keep iterating step 3 until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

Clustering

- Choosing value of ‘k’ in K-means clustering:
- **Elbow method :** The Elbow method looks at the total WSS (within-cluster sum of square) as a function of the number of clusters. The no of clusters are selected such that adding another cluster doesn’t improve the total WSS. The total WSS is calculated for different values of k. The location of a bend (knee) in the plot between total WSS and K is generally considered as an indicator of the appropriate number of clusters.
- **Average silhouette method :** Average silhouette method determines how well each object lies within its cluster by computing the average silhouette of observations for different values of k. The location of the maximum is considered as the appropriate number of clusters.
- **Business Context:** Value of K can also be driven based on business strategy which require certain no of clusters. E.g For marketing department wants to cluster the customers into certain groups and send different Promotion offers.

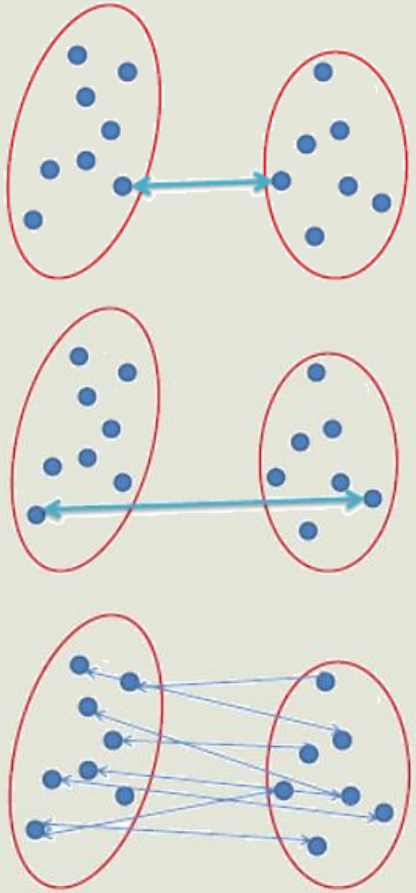
Clustering

- Scaling/standardisation before performing Clustering.:
  - While doing clustering if the variables in the data set are in different units then these values can not be compared with each other (e.g. some variables are in Kg and some variables are in meter). If standardization of data is not done then the unit of dimension can distort relative near-ness of observations.
  - By doing Scaling/standardization improves the Convergence of cluster canters.

Clustering

Different linkages used in Hierarchical Clustering:

- **Single linkage:** This measure defines the distance between two clusters as the minimum distance found between two cluster. A concern of using single linkage is that it can sometimes produce chaining amongst the clusters. This means that several clusters may be joined together simply because one of their cases is within close proximity of case from a separate cluster.
- **Complete linkage:** This measure defines the distance between two clusters as the furthest distance between two cluster. In complete linkage, outlying cases prevent close clusters to merge together because the measure of the furthest neighbour exacerbates the effects of outlying data.
- **Average linkage**. This method is supposed to represent a natural compromise between the linkage measures to provide a more accurate evaluation of the distance between clusters. For average linkage, the distances between each case in the first cluster and every case in the second cluster are calculated and then averaged.



Principal Component Analysis

Applications of using PCA:

1. Dimensionality reduction : PCA is used as dimensionality reduction technique in image recognition by comparing the principal components of each of the images
2. Component/Feature Selection: By plotting a scree plot we can identify how many principal components are required to describe the variance in given data.
3. Visualization of data: PCA improves the visualization of data as the same information can be seen using fewer PCA components and which can be plotted in 2D space.

Principal Component Analysis

- Basis transformation and variance as information:
- **Basis Transformation**: Any vector space has multiple bases. Basis transformation is process of translate vectors in terms of one basis into terms of the other, given two bases of a vector space. This is done by applying the transformation matrix described in one basis to the vectors in the other vector space.
- **Variance as information**: In a given dataset, the parameters which has more variance or variety contains more information. To measure the importance of a column can be determined by checking its variance of the values. The direction of PCA is determined by the direction in which maximum variance is observed.

Principal Component Analysis

Shortcomings of using Principal Component Analysis:

1. Independent variables become less interpretable: After implementing PCA on the dataset, the original features will turn into Principal Components. Principal Components are not as readable and interpretable as original features.

2. Data standardization is must before PCA: standardizing the data becomes mandatory before implementing PCA. If standardization of data is not done before PCA, the resultant principal components will be biased towards the features with variance. PCA will not be able to find the optimal Principal Components and leading to false results.

3. Information Loss: if the number of Principal Components are not selected with care, it may miss some information as compared to the original list of features which may be critical. e.g. Parameters with very less variation may be important in detecting fraud.