ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

CSPC54

PROJECT REPORT

COUNTRIES CLOSENESS

TEAM MEMBERS:

A Lokanush	106120009
Ashwath Vasudevan	106120017
I Viswanath	106120043
Skandan R	106120121

CODE

https://github.com/Machine-Impossible/Countries-Closeness

DATASET

https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data

PROBLEM STATEMENT

The aim of this project is to cluster countries based on the economic and development similarities. We have used and compared various clustering algorithms by their silhouette score. The clustering algorithms used are,

- K - Means
- DBSCAN
- Fuzzy
- Hierarchical

K MEANS

The goal of k-means clustering, a vector quantization technique that originated in signal processing, is to divide n observations into k clusters, where each observation belongs to the cluster that has the closest mean (also known as the cluster centroid or cluster centre), which serves as a prototype for the cluster. As a result, the data space is divided into Voronoi cells. The geometric median is the only one that minimises Euclidean distances; k-means clustering minimises within-cluster variances (squared Euclidean distances), but not regular Euclidean distances,

which would be the more challenging Weber problem. For instance, k-medians and k-medoids can be used to find better Euclidean solutions.

In order to identify the value of k such that the result obtained is accurate, we run an iterative loop which changes the value of k from 1 to an upper limit and find the wcss (within cluster squared sum error) value. We need to choose a value for k such that the wcss value is neither too high which could mean that the result obtained is not accurate or too low which could mean that we have a lot of clusters and all data items might be clustered into unique clusters which is not our goal. Hence we choose a value for k that lies on the elbow of the k vs wcss graph.

DBSCAN

It is a density based clustering algorithm. It takes in two parameters eps which are the distance to look for points in the cluster and min_samples which is no. of points to consider it as a core point. The min_samples parameter is always greater than equal to the dimension of the dataset which is 9 in our case. Using this and by plotting the k-distance graph we can find the eps parameter. For the K-Distance Graph we will be using the KNN method where k = 9.

FUZZY

Fuzzy c-means clustering has been carried out on the dataset. In this clustering, every point in the dataset belongs to every cluster to a certain degree. The number of

clusters is fixed and a partition matrix is initialised assigning each point an arbitrary cluster. A fuzzy parameter is chosen which will be used in an equation used to calculate the new cluster centroids. Based on the new cluster centroids, the partition matrix is updated and the process is repeated till the cluster centroids don't change.

In this project, the number of clusters are chosen to be 5 which is passed to the function FCM() which performs fuzzy c-means clustering. The optimal number of clusters can be found using the elbow method (similar to K-means clustering) which plots the WCSS against the number of clusters. The point where the slope of the curve changes drastically corresponds to the optimal number of clusters.

HIERARCHICAL

The Agglomerative (bottom-up/additive) approach has been carried out on the dataset. Initially, each data point is considered as a cluster. Then, the closest pair of clusters are considered and combined together. This is repeated until we end up with a single cluster and the dendrogram is constructed. Once this is achieved, The dendrogram is used to divide this cluster into the appropriate number of clusters to solve our problem.

The optimal number of clusters is determined from the dendrogram as follows:

- In the dendrogram, the largest vertical difference between nodes is located and the vertical line that satisfies this is chosen.
- In the middle of the identified line,an horizontal line is passed.
- The number of vertical lines intersecting it is the optimal number of clusters

In this project, the optimal number of clusters turned out to be about 5, which is passed to the AgglomerativeClustering function which performs agglomerative clustering.

CONCLUSION

For the given dataset hierarchical clustering had the highest silhouette score thus the better clustering algorithm for this case.