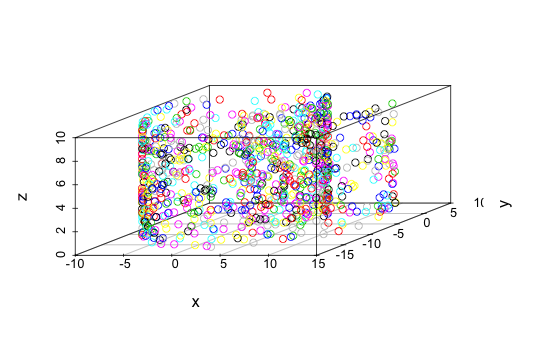
Report

**Part 1:**

Given Dataset Clustering:



Distance Based Clustering:

|  |  |
| --- | --- |
|  | This plot is obtained from applying average hierarchical clustering as distance based clustering. Here as can be observed from the 3d plot, the clusters are better organized than the given dataset clusters in regards to distance between points. I have applied average clustering as distance based clustering because generally it is considered one of the better distance clustering algorithm available as it calculates average distance between each and ever point between the clusters. So the clusters are well defined comparatively as they are scattered and clustering is able to identify which points belong to which cluster. Though some outliers are incorrectly identified as part of wrong cluster. But as can be observed from given set clusters and distance cluster the clusters generated in distance based are quite different as for distance based clustering the main criteria for finding clusters is the distance between the points, while data set was classified in cluster on other predefined conditions |

Density Based Clustering:

|  |  |
| --- | --- |
|  | This plot is obtained from applying DBSCAN clustering as density based clustering. Here as can be observed from the 3d plot, the clusters are better organized than the given dataset clusters in regards to density between points. But here as the number of clusters are not predefined. The DBSCAN has formed more than 8 clusters with respect to the distribution and density of points estimated from the graph. Thus the plots is completely different from the given dataset cluster, as density based clustering finds clusters based on density connectivity and density reachability. |

Graph Based Clustering :

|  |  |
| --- | --- |
|  | This plot is obtained from applying k cores clustering as graph based clustering. Here the dataset was first converted into graph. The k core normally is used in social networking. It is used identify well connected structures within the graph. The cores find nodes in the graph which are connected to it or to the other initial nodes in the cluster and form clusters iteratively. So the plot has clusters more concerned with connectivity between the nodes when the dataset is converted to graph. |

**Part 2 :**

|  |  |
| --- | --- |
| **User Defined** | **K Means** |
|  |  |
|  |  |

In the above to plots, it looks K Means gives a better clustering than user defined. Though point variability of user defined is higher. According to me this is happening because, due to increase in weightage of x and y dimensions less information is hidden in user defined clustering and as dimensions are increasing the distance between the nodes in data set is also increasing so, the use defined distance cluster is able to efficiently give out which node the cluster belongs to.

**Part 3:**

For the second dataset, I think 2 clusters Is the best to seek, as in the data set only two clusters are given.

As it is large data set, traditional and most frequently clustering methods cannot be used. For example distance based clustering like hierarchical clustering takes O(n^3). So it will take plenty of time which is not feasible. Also density based scanning methods like dbscan are also not much useful as they cannot find clusters well in high dimensional data. So clustering technique which can work on large Dataset has to be used. From traditional clustering, k means is very popular and one of the most used and efficient. So I have used one of scalable version derived from k means i.e. Fuzzy C means clustering. Fuzzy C means is also partitioning algorithm like k means. It is type of soft clustering though. But it has advantage that it gives the membership level of each data element in dataset. This gives us idea about the association information between node and the cluster to which it belongs. Though it has a disadvantage that, nodes may belong to more than one cluster, but since here only two clusters are to be used, we can get information about the association of the node to the cluster from membership levels.

Sources and References used for this projects :

<https://en.wikipedia.org/wiki/DBSCAN>

<https://en.wikipedia.org/wiki/K-means_clustering>

<https://en.wikipedia.org/wiki/Hierarchical_clustering>

<https://en.wikipedia.org/wiki/Fuzzy_clustering#Fuzzy_c-means_clustering>

<https://en.wikipedia.org/wiki/Cluster_analysis>

<https://en.wikipedia.org/wiki/Degeneracy_(graph_theory)>

<http://stackoverflow.com/>

Introduction to data mining

Data mining concept and techniques

[www.inside-r.org](http://www.inside-r.org)

[www.statmethods.com](http://www.statmethods.com)