Cluster Analysis
Analysis Report

Assignment 3

Syed Muzammil Ahmed- 25371

Machine Learning – 2

In this task, we are implementing different clustering models on different datasets and checking out their results by using different metrics. Datasets are diverse and belongs to different industries like health, business, airline etc. and 50% of the dataset are with labels and rest are without labels.

The models we are using in this task are centroid-based (mini batch k-means), connectivity-based (hierarchical agglomerative and divisive), and density-based (DBScan, OPTICS, MeanShift) and the metrics we are using to examine their performance are silhouette score, DBI, Fowlkes, ARI, Mutual Information, V-measure, and homogeneity. Elbow method is also used in some clustering tasks where data is without label to find out the optimum number of clusters.

First dataset, we are using is marketing campaign dataset, downloaded from Kaggle. It's a dataset with label class. On this data, the clustering algorithms are performing quite fine. Centroid based and connectivity-based models are performing quite well, but the performance of density-based models are not much effective.

Second dataset we are using is Online shoppers' dataset and it contain around 12000 records of the shoppers. This dataset also contains label class. On this dataset, all three types of clustering models are performing good with silhouette score of 0.5, but optics is performing below average.

Third dataset we have is related to health sector named as blood pressure prediction dataset. It contains around 2000 patients' records. Here on this dataset, centroid based, and connectivity-based model are performing good with Fowlkes score of around 0.6, but the Fowlkes score of density-based models are around 0.2 which mean they are not classifying binary class perfectly.

Fourth dataset we are using is also related to health named as Heart rate prediction. It is taken from Kaggle. In this, centroid based models are performing well in term of silhouette score value, but the DBI score of connectivity-based model and Density based models are better which show the optimum number of clusters used.

The next dataset we are using is Credit card data, downloaded from Kaggle. It also has a class label. On this dataset, Centroid based model is performing exceptionally well with silhouette score of 0.6 which means clusters are perfectly formed, but in other models silhouette score is very low.

After this we are using High income detection dataset. On this dataset, most of the models are performing well except Optics model. Its not forming clusters correctly that's why has very low silhouette score.

We have used one more dataset, which is an Airline traffic data. On this, DBSCAN is performing very well, but centroid based, and agglomerative model was not performing well. It's a dataset without label class and its size is also high after one hot encoding.

Next clustering data we are using for analysis are Facebook data and Sales dataset. Both datasets are without label class. On Facebook dataset. Most of the models are clustering very well with silhouette score of 0.8 and 0.7. only model with low score is optics, but this time its not in negative.

Last data we are using is sales data. It has high dimensionality with 800 rows and 700 columns. Only DBSCAN is clustering well on this dataset, rest of the model has very low score on this high dimensional data.

Pros and Cons: -

1- **Centroid Based Models: -**

1- ARI score of these models is not good enough, which means degree of agreement between clusters are not good enough.
2- Fowlkes score of these is good, which means it is classifying both class label accurately.
3- V-measure of these models are not much effective, that means datapoints are much compact in the clusters.
4- Homogeneity index of these models are good. Which means cluster contain datapoints of only one class.

2- **Connectivity Based Models: -**

1- These models are good in classifying binary classes.
2- Datapoints are not much compacted in these models.
3- Homogeneity among the data points are good in the agglomerative model.

3- **Density Based Model: -**

1- Optics model is not very good in detecting clusters, but it is good in classifying binary class
2- DBSAN is one of the best models among density-based models. Its cluster detection is good as well as its homogeneity is also good.
3- Mean Shift model is very high computational as compared to ither models.