DATA MINING HOMEWORK 2 REPORT

OZAN GECKIN -1801042103

About Homework

Using and analyzing Frequent Pattern Growth, K-means, DBSCAN and Chameleon clustering technologies in my homework. I need to use 2 datasets to use the algorithms. To use in these algorithms, I created datasets with 2 dimentions and 20 dimentions with 1000 elements each. I used the make_classificiation function of the sklearn.datasets library to create these datasets. In my homework, I printed the graphs of the results using data sets with 2 dimentions. I could not print the output of the Frequents pattern Growth clustering technology to the screen. But I printed the graphs of other K-Means, Chameleon and DBSCAN. I will add screenshots to the rest of my report. There are 2 things I need to calculate in my homework they are silhouette coefficient and computational time.

Silhouette Coefficient:

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

Compute the average distance from all data points in the same cluster (ai).

Compute the average distance from all data points in the closest cluster (bi).

Compute the coefficient:

$$\frac{b^i - a^i}{max(a^i, b^i)}$$

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.

Computation time:

Computation time (also called "running time") is the length of time required to perform a computational process. Representation a computation as a sequence of rule applications, the computation time is proportional to the number of rule applications.

Information about cluster technologies:

K Means Algorithm

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

K-Means Algorithm time complexity O(n k d)

DBSCAN Algorithms

Density-Based **Clustering** refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.

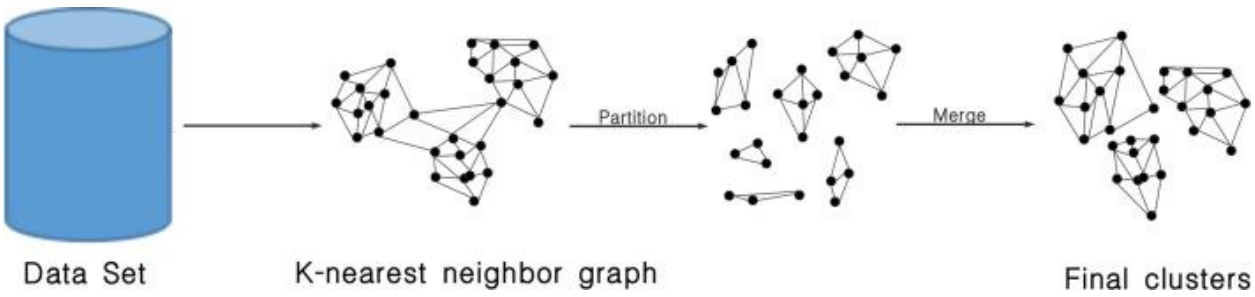
The DBSCAN algorithm uses two parameters:

- **minPts:** The minimum number of points (a threshold) clustered together for a region to be considered dense.
- **eps (ε):** A distance measure that will be used to locate the points in the neighborhood of any point.

DBSCAN Algorithm time complexity O(nlogn)

Chameleon Algorithm :

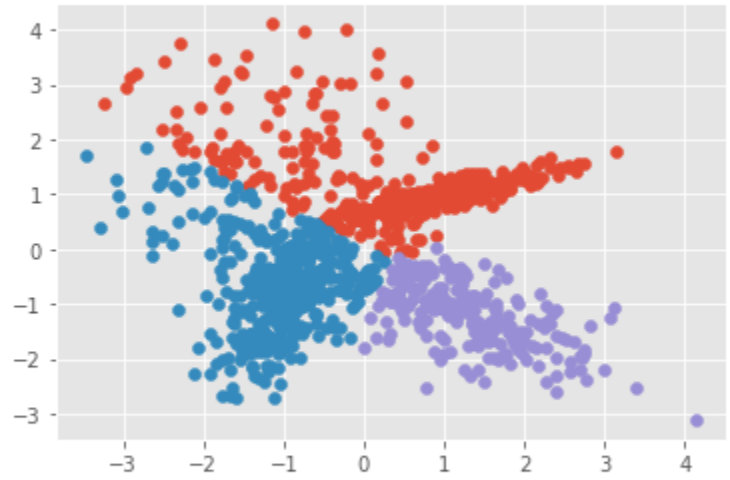
Chameleon algorithm is a hierarchical agglomerative clustering algorithm. It combines small clusters to achieve the final clustering. It is a bottom-up clustering algorithm. Its computing load is relatively small.



Chameleon Algorithm time complexity O(n^2)

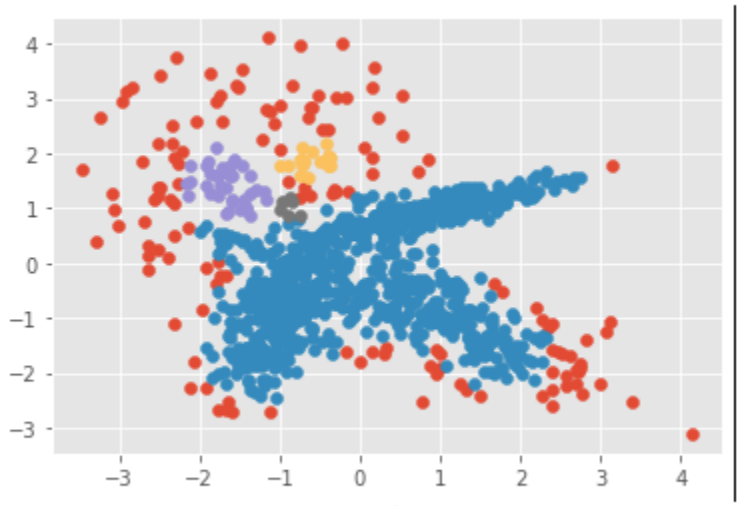
Graphs, Silhoutte score and Computational time:

K-Means 2D dataset Graph



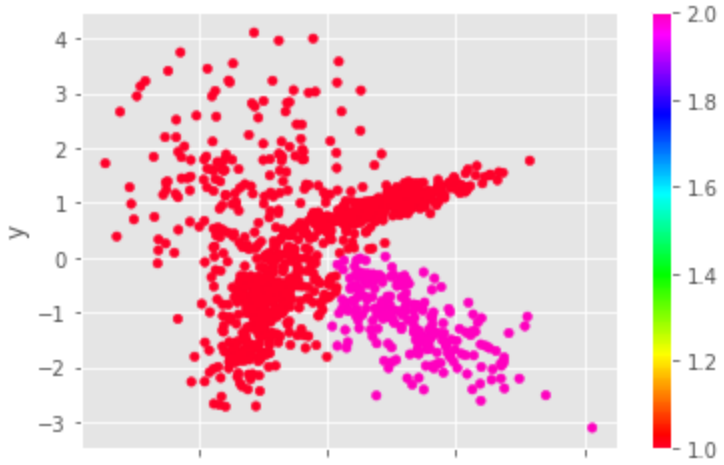

```
Computational time for 2D dataset KMeans : 0.130859
Silhouette score for 2D dataset KMeans : 0.433986
Computational time for 20D dataset KMeans : 1.045494
Silhouette score for 20D dataset KMeans : 0.052752
```

DBSCAN 2D dataset Graph




```
Computational time for 2D dataset DBSCAN : 0.021944
Silhouette score for 2D dataset DBSCAN : -0.024356
```

```
Chameleon 2D dataset Graph
```




```
Computational time for 2D dataset Chameleon : 84.033945
```

```
Computational time for 20D dataset Chameleon : 107.427087

Computational time for 2D dataset Fpgrowth : 0.002973
Computational time for 20D dataset Fpgrowth : 0.011442
```

Which clustering technique is more suitable for your dataset? Write a discussion about it using the results mentioned above and characteristics of the clusters and the dataset.

I could not draw fpgrowth and calculate its silhouette. I added the data I have to my output report. I use the K-Means algorithm according to the datasets I have. Because I have 1000 pieces of data, the Chameleon algorithm worked slowly. K-Means was faster. When I compared the DBSCAN algorithm with the K-Means, the Silhoutte score of the K-Means algorithm was better. But if my dataset was much larger I could use DBSCAN. Because DBSCAN algorithm is faster. I couldn't predict how the K-Means algorithm would perform in very large data. Another reason affecting my preference is graphs. Clusters formed more smoothly in the K-Means graph.