ASSIGNMENT-5
TEJHAN BHARADWAJ RAMKUMAR
KH50252

1. K-MEANS:

Dataset : A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9)
Centroid = (2,10) , (5,8) , (1,2)

Distance:

Cluster 0:	0	5	8.48	3.6	7.07	7.21	8.06	2.23
Cluster 1:	3.6	4.24	5	0	3.60	4.12	7.21	1.41
Cluster 2:	8.06	3.16	7.28	7.21	6.70	5.38	0	7.61

MIN	0	2	1	1	1	1	2	1

	MEAN
Cluster 0: (2,10)	(2,10)
Cluster 1: (8,4), (5,8), (7,5), (6,4), (4,9)	(6,6)
Cluster 2: (2,5), (1,2)	(1,3)

New centroid : (2,10), (6,6), (1,3)

Distance:

Cluster 0:	0	5	8.48	3.6	7.07	7.21	8.06	2.23
Cluster 1:	5.65	4.12	2.82	2.23	1.41	2	6.40	3.60
Cluster 2:	7.07	2.23	7.07	6.40	6.32	5.09	1	6.70

MIN	0	2	1	1	1	1	2	0

	MEAN
Cluster 0: (2,10), (4,9)	(3,9)
Cluster 1: (8,4), (5,8), (7,5), (6,4)	(6,5)
Cluster 2: (2,5), (1,2)	(1,3)

New centroid : (3,9), (6,5), (1,3)

Distance:

Cluster 0:	1.41	4.12	7.07	2.23	5.65	5.83	7.28	1

Cluster 1:	6.40	4	2.23	3.16	1	1	5.83	4.47
Cluster 2:	7.07	2.23	7.07	6.40	6.32	5.09	1	6.70
MIN	0	2	1	0	1	1	2	0

	MEAN
Cluster 0: (2,10), (5,8), (4,9)	(5,13)
Cluster 1: (8,4), (7,5), (6,4)	(10,6)
Cluster 2: (2,5), (1,2)	(1,3)

New centroid : (5,13), (10,6), (1,3)

Distance:

Cluster 0:	4.24	8.54	9.48	5	8.24	9.05	11.70	4.12
Cluster 1:	8.94	8.06	2.82	5.38	3.16	4.47	9.84	6.70
Cluster 2:	7.07	2.23	7.07	6.40	6.32	5.09	1	6.70
MIN	0	2	1	0	1	1	2	0

The final clusters are : (2,10) (5,8) (4,9)
(8,4) (7,5) (6,4)
(2,5) (1,2)

ANSWER:

(a) (A1) (A3,B1,B2,B3,C2) (A2,C1)

(b) After 4 iterations the final cluster is : (A1,B1,C2) (A3,B2,B3) (A2,C1)

2. BIRCH:

OPTICS is a cluster analysis method which does not explicitly provides a data set cluster. It outputs a cluster ordering where the objects in a denser cluster are listed closely to each other. OPTICS does not require user to provide a specific density threshold and it extracts the basic clustering information. Whereas BIRCH requires a user to provide specific density cluster and cluster ordering is not produced. Hence BIRCH finds difficulties in finding cluster of arbitrary shape whereas OPTICS outputs the clustering order and it does not require user to provide specific density threshold. Time complexity of BIRCH is O(n) where ‘n’ is the number of objects to be clustered. Since each node can hold only limited entries because of its size, users doesn’t consider it to be a natural cluster. So we can try improving the size. If the shape is not spherical BIRCH can’t perform well.

Also introducing cluster ordering in BIRCH makes it perform for arbitrary data. These are some of the modifications that can be introduced in BIRCH.

3. DENSITY BASED CLUSTERING:

Partitioning and Hierarchical based clustering methods are only used to find spherical shaped clusters. They can't find cluster's which has dense regions in the data space separated by sparse regions. Hence, Density based clustering are used. They can find clusters of arbitrary shapes and can find where noise and outliers are included in the cluster. Three methods of density based clustering are: DBSCAN, OPTICS, DENCLUE. Using one of these methods density based clustering can be performed. Also they can be used in e-commerce applications where we want to improve the sales by recommending people their relevant products. In the beginning we don't know what the customers are looking for but based on the dataset we can predict and recommend relevant products to our customers. Here DBSCAN can be introduced to the dataset to form a cluster of the relevant products the people have bought. This can attract more people. This is one of the applications of Density based clustering.

4. CONSTRAINT AND CLUSTERING:

Cluster modification can be done is such a way that it should consider all the household and employee spaces and commuting options for people. To do this we have to consider all the places present as cluster and partition the cluster in different ways such that where most number of points are present we can conclude that is the place where most people stay and which is the most accessible place to everyone. If a particular place is most accessible it is meant that almost most of the people travel there so setting up an ATM there would be convenient for everyone. So that this avoids the constrain of rivers and highways as people have a way to commute to those places since it is largely crowded. Also the second obstacle is as there are 10,000 households in each cluster a must link constrain can be used. In this case what we can do is we can see the most crowded cluster and keep updating the cluster and the cluster will be filled with gaps which is the obstacle in constrain 1. After that it follows normal k-means cluster. We can also use hierarchical methods and form microclusters and then perform k means. Also we can use density based cluster where households, rivers and highways can be separated. Also while building ATM's time and space complexity should also be considered whether people can access it at a shorter time and there is sufficient space for more number of people.