DATA MINING ASSIGNMENT-6

Parampreet Singh

Ans:- The technique of splitting a set of data objects into multiple groups is known as clustering. Each group is a cluster with objects in one cluster like those in another but differing from those in other clusters.

Clustering is the term used to describe a group of clusters that occur as a result of clustering analysis.

The many clustering approaches are listed below:

→ Partitioning methods: Assume database D including n items and that the partitioning method creates k data partitions. Each of the 'k' partitions is a cluster with $k \leq n$.

It means that the data should be divided into k groups with each group containing at least one object and each object belonging to one group only.

These methods use the means of moving to represent cluster centre and find

distances b/w the items to construct clusters.
Ex: k-means, k medoids & CLARANS

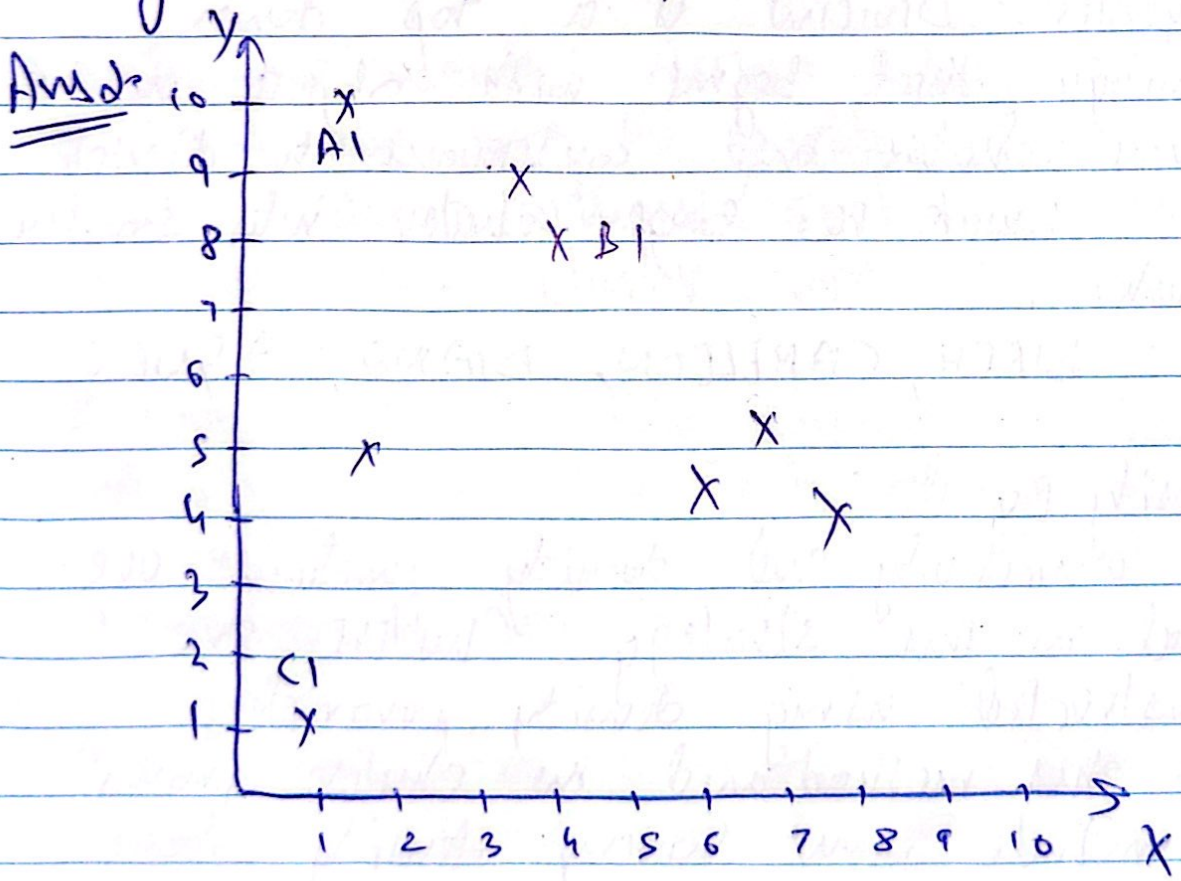
→ Hierarchical Methods:
This method provides a hierarchical decomposition of a set of objects. The way hierarchical decomposition is generated can be used to classify this agglomerative and divisive techniques are the two options is generated can be used to classify this. Agglomerative is a bottom up strategy that begins by grouping objects into distinct groups and then continue to merge objects that are close together. Divisive is a top down strategy that begins with objects in the same cluster and continues to divide and break the large cluster into smaller clusters.
Eg: BIRCH, CAMELEON, DIANA, AGNES

→ Density Based:
The connectivity and density functions are used in this strategy. Clusters are constructed using density parameters in this method and the cluster grows to include items having density above

the threshold values.
It is divided into 2 sections, minpts & Eps.
the greatest radius of a neighbourhood is Eps.
outliers are detected using this strategy.
Eg. DBSCAN, OPTICS, Dendue

→ Grid based methods:
The objects form a grid in this grid based method and the object space is quantized into a finite no. of cells that form a grid structure. It has quick processing speed.
Eg. STING, CLIQUE, wave cluster

Ans:



(a) Here $k = 3$, A_1, B_1, C_1 → initial mean points. These are the centers of the clusters.

K means: Find distance b/w data points and mean point

Distance metric: Euclidean Distance

$$d(P_1, P_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

I st iteration

$d(A_1, A_1) = 0$

$d(A_1, A_2) = \sqrt{(2-2)^2 + (10-5)^2} = 5$

$d(A_1, A_3) = \sqrt{(2-8)^2 + (10-4)^2} = 8.49$

$d(A_1, B_2) = 7.07$

$d(A_1, B_3) = 7.21$, $d(A_1, C_2) = 2.236$

$d(B_1, B_1) = 0$ $d(B_1, A_2) = 4.24$

$d(B_1, A_3) = 5$ $d(B_1, B_2) = 3.36$

$d(B_1, B_3) = 4.12$ $d(B_1, C_2) = 1.414$

$d(C_1, C_2) = 0$ $d(C_1, A_2) = 3.16$ $d(C_1, A_3) = 7.28$

$d(C_1, B_2) = 6.7$ $d(C_1, B_3) = 5.385$ $d(C_1, C_2) = 7.62$

So, Assigning points to the clusters they have min distance to

$A_2(2, 5) \longrightarrow C_1(1, 2)$

$A_3(8, 4) \longrightarrow B_1(5, 8)$

$B_2(7,5) \longrightarrow B_1(5,8)$
$B_3(6,4) \longrightarrow B_1(5,8)$
$C_2(4,9) \longrightarrow B_1(5,8)$

Updating mean points

Center O_1 : $(2,10)$

Center O_2 : $\left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6,6)$

Center O_3 : $\left(\frac{3+4}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

(b) distance for other iterations :

$d(O, A1) = 0$
$d(O_2, A1) = \sqrt{(6-2)^2+(6-10)^2} = 5.656$
$d(O_3, A1) = \sqrt{(1.5-2)^2+(3.5-10)^2} = 6.519$

$d(O, A2) = 5 \quad d(O_2 A_2) = 4.12 \quad d(O_3 A3) = 2.5$

$d(O_1, A_3) = 8.49 \quad d(O_2 A_3) = 2.83 \quad d(O_3 A_3) = 6.52$
$d(O_1 B_1) = 3.61 \quad d(O_2 B_1) = 2.24 \quad d(O_3, B_1) = 5.7$
$d(O_1 B_2) = 7.07 \quad d(O_2 B_2) = 1.414 \quad d(O_3 B_2) = 5.7$
$d(O_1 B3) = 7.21 \quad d(O_2 B_3) = 2 \quad d(O_3 B_3) = 4.52$
$d(O_1 C_1) = 8.06 \quad d(O_2 C_1) = 6.4 \quad d(O_3, C_1) = 1.58$
$d(O_1 C_2) = 2.23 \quad d(O_2 C_2) = 3.65 \quad d(O_3 C_3) = 6.04$

Reassigning points to clusters

$A_1, C_2 \longrightarrow O_1\ (2, 10)$

$A_3, B_1, B_2, B_3 \longrightarrow O_2\ (5, 6)$

$A_2, C_1 \longrightarrow O_3\ (1.5, 3.5)$

clusters :

$\{A_1, C_2\}$ → now we have to update the centers

$\{A_3, B_1, B_2, B_3\}$

$\{A_2, C_1\}$

$$\bar{O}_1 = \left(\frac{2+4}{2}, \frac{10+9}{2}\right) = (3, 9.5)$$

$$\bar{O}_2 = \left(\frac{3+5+7+6}{4}, \frac{4+8+5+4}{4}\right) = (6.5, 5.25)$$

$$\bar{O}_3 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$$

After 3rd iteration -

$$d(\bar{O}_1, A_1) = \sqrt{(3-2)^2 + (9.5-10)^2} = 1.12$$

$$d(\bar{O}_2, A_1) = \sqrt{(6.5-2)^2 + (5.25-10)^2} = 6.54$$

$$d(\bar{O}_3, A_1) = \sqrt{(1.5-2)^2 + (3.5-10)^2} = 6.52$$

$d(\bar{O}_1, A_2) = 4.61$, $d(\bar{O}_2, A_2) = 4.51$, $d(\bar{O}_3, A_3) = 1.58$

$d(\bar{O}_1, A_3) = 7.43$, $d(\bar{O}_2, A_3) = 1.95$, $d(\bar{O}_3, A_3) = 6.52$

$d(\bar{O}_1, B_1) = 2.5$ $d(\bar{O}_2, B_1) = 3.13$ $d(\bar{O}_3, B_1) = 5.7$

$d(\bar{O}_1, B_2) = 6.02$ $d(\bar{O}_2, B_2) = 0.55$ $d(\bar{O}_3, B_2) = 5.5$

$d(O_1, B_3) = 6.26$ $d(\bar{O}_2, B_3) = 1.346$ $d(\bar{O}_3, B_3) = 4.52$

$d(\bar{O}_1, C_1) = 7.76$ $d(\bar{O}_2, C_1) = 6.39$ $d(\bar{O}_3, C_1) = 1.58$

$d(\bar{O}_1, C_2) = 1.18$ $d(\bar{O}_2, C_2) = 4.506$ $d(\bar{O}_3, C_2) = 6.04$

Assigning points to clusters:

$A_1, B_1, C_2 \rightarrow O_1$

$A_3, B_2, B_3 \rightarrow O_2$

$A_2, C_1 \rightarrow O_3$

Updating centroids,

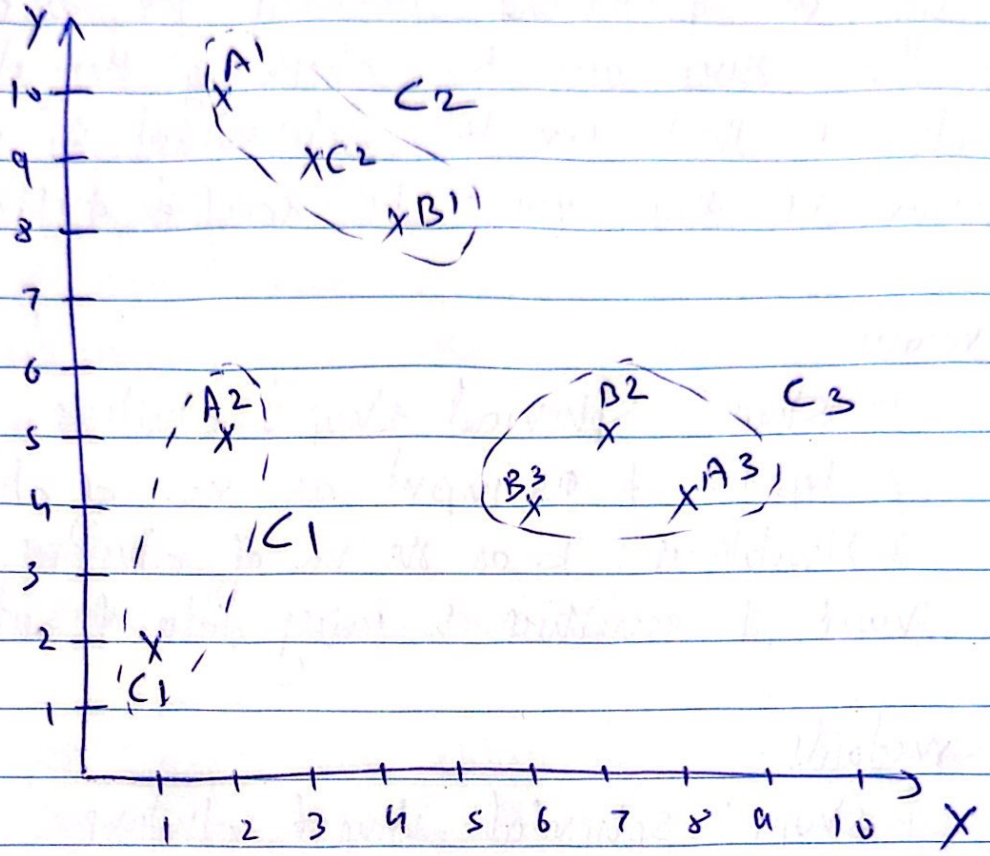
$$\bar{O}_1 = \left(\frac{4+5+4}{3}, \frac{10+8+9}{3}\right) = (3.67, 9)$$

$$\bar{O}_2 = \left(\frac{8+7+6}{3}, \frac{4+5+4}{3}\right) = (7, 4.3)$$

$$\bar{O}_3 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$$

After this iteration, same clusters are formed again and same centroids are achieved.

final clusters



ANSWER 3:

Both k-means and k-medoids are popular clustering techniques, each with distinct strengths and weaknesses.

a) K-means vs. K-medoids: K-means is known for its computational efficiency, especially suitable for large datasets with a time complexity of O(tkn), where t is the number of iterations, k is the number of clusters, and n is the number of data points. However, it is sensitive to outliers and noise since it uses the mean of cluster members to determine the cluster center, which can be significantly influenced by extreme values. In contrast, K-medoids, which operates with a complexity of O(k(n-k)^2), is more robust against outliers as it uses actual data points as the centers (medoids). This makes K-medoids computationally expensive and less scalable for larger datasets compared to K-means.
b) Comparison with Hierarchical Clustering: Hierarchical clustering, such as the Agglomerative Nesting (AGNES) method, doesn't require the number of clusters to be defined a priori, which is a key advantage over K-means and K-medoids where the number of clusters must be specified beforehand. Hierarchical clustering builds a dendrogram to represent data hierarchically, allowing the model to capture relationships at different scales. However, hierarchical methods can be slower and less efficient for large datasets since they typically involve recursive calculations that can be computationally intensive. In contrast, K-means and K-medoids are more efficient in handling large datasets but assume clusters are spherical, which can be a limitation if the natural cluster shapes vary significantly.

ANSWER 4:

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) struggles with identifying clusters of arbitrary shapes primarily because it constructs a clustering feature tree based on predefined parameters that summarize the data points. This method inherently assumes spherical cluster shapes due to its reliance on centroid or radius-based summaries, which makes it less flexible in adapting to non-linear cluster boundaries.

On the other hand, OPTICS (Ordering Points To Identify the Clustering Structure) does not face this limitation because it examines the spatial density of data points, creating an ordered reachability distance plot that naturally adapts to varying densities and shapes of clusters. This method does not presume any specific shape of the clusters, allowing it to effectively handle diverse cluster configurations.

Proposed Modifications to BIRCH:

Integrate Density-Based Analysis: Incorporate density-based metrics into the BIRCH algorithm, allowing it to adjust the clustering feature tree based on local density variations rather than just simple centroid or radius measurements. This can help BIRCH adapt to clusters of varying shapes and sizes, similar to how OPTICS operates.

Flexible Thresholding: Implement a dynamic threshold for cluster merging that adjusts based on the density of the surrounding data points, rather than a static threshold. This would help in recognizing and merging clusters with non-spherical shapes more effectively.

Hybrid Approach: Combine the hierarchical structure of BIRCH with a post-processing phase using a technique like DBSCAN or OPTICS to re-evaluate and possibly restructure the clusters formed in the initial phases. This can help refine the cluster shapes to better match the actual data distribution.

Ans:-

These are following criteria for the clustering algorithm may are by checking the shape of clusters that can be determined & input parameters that must be specified & limitations.

(a) K-means:

1- Shape: Spherical shaped clusters
2- Input: k as input as no. of clusters
3- Limitations: k as the no. of clusters before hand & sensitive to noisy data & outliers.

(b) k-medoids:

1- Shape: Spherical shaped clusters
2- Input: no. of clusters k
3- Limitations: Not suitable for clustering non-spherical groups of objects. Good for small data set only, obtains different results for different runs.

(C) CLARA:

1- Shape: Convex shaped clusters
2- Input: no. of clusters k
3- Limitations: If sampling is biased, we don't get good clusters & the but k mediods may not be selected during sampling.

4- BIRCH:

1- Shape: Spherical shape as it uses radius & diameter to measure.

2- Input: n dimensional data points

3- Limitation: Sensitive to insertion order of data points, clusters may not be so natural because the size of leaf nodes are fixed.

(e) CHAMELEON:

① Input: n dimensional categorical data points

② Shape: non spherical shaped clusters

③ limitation: worst case complexity for quadratic & doesn't handle the noise at all

(f) DBSCAN:

① Input: 2 inputs - Eps, minpts

Eps → max. radius of neighbourhood

minpts → min. of points in an eps neighbourhood of that point in that particular cluster.

② Shape: clusters of arbitary shapes

③ limitations: It has worst case complexity for quadratic