Types of clustering

1) Partitioning methods –

(a) Given a database of n objects or data tuples a partitioning method constructs K partitions of data, where each partition represents a cluster and $K \leqslant n$.

(b) Given K, the number of partitions to construct it creates an initial partitioning.

(c) It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another.

(d) The criteria of a good partitioning is that objects in the same cluster are close to each other, whereas objects of different clusters are far apart.

(e) Ex – K-mean, K-medoid

2) Hierarchical methods

(a) This method creates a hierarchical decomposition of the given set of data objects. It can be either agglomerative or divisive.

(b) The agglomerative (bottom-up) approach starts with each object forming a separate group. It successively merge the objects that are close to one another, until all of the groups are merged into one.

(c) The divisive (top-down) approach starts with all of the objects in the same cluster. On each successive iteration, a cluster is split up into small clusters, until each object forms its own cluster or until a termination criteria meets.

EX - AGNES, DIANA, BIRCH.

(3) Density based clustering

(a) This method based on the notion of density.

(b) For each data point within a given cluster, the neighborhood of a given radius have to contain at least a minimum no. of points.

(c) This method can be used to filter out noise and discover clusters of arbitrary shape.

EX - DBSCAN, OPTICS

(4) Grid based methods

(a) These methods quantize the object space into a finite no. of cells that form a grid structure.

(b) All of the clustering operations are performed on the grid.

(c) Advantage is fast processing time.

EX - STING

(5) Model-based methods

(a) This approach hypothesizes a model for each of the clusters and finds the best fit of the data to the given model.

(b) It also leads to a way of automatically determining the no. of clusters based on standard statistics.

(c) It takes noise or outliers into account. thereby contributing to the robustness of approach.

Ex- COBWEB, self-organizing feature maps

(6) Methods for high dimensional data

(a) High dimensional data can typically have many irrelevant dimensions.

(b) subspace clustering methods - which search for cluster in subspaces of the data, rather than over the entire data.

Ex - CLIQUE, PROCLUS

(c) Frequent pattern based clustering - It extracts distinct frequent patterns among subsets of dimensions that occur frequently.

Ex- pCluster

(7) constraint based methods

These methods perform clustering by incorporating user-specified or application oriented constraint.

K-mean clustering

Q. suppose that the data mining task is to cluster the following eight points into 3-clusters.

$A_1(2,10)$, $A_2(2,5)$, $A_3(8,4)$, $B_1(5,8)$, $B_2(7,5)$

$B_3(6,4)$, $C_1(1,2)$, $C_2(4,9)$

The distance function is Euclidean distance. suppose initially we assign A_1, B_1 and C_1 as the center of each cluster, respectively. Use the K-means algorithm to show only

(a) the 3-cluster centers after the first round of execution

(b) The final three clusters.

Ans

C_1 $A_1(2,10)$

~~$A_1A_1 = \sqrt{(2-2)^2+(10-10)^2} = 0$~~

$A_2A_1 = \sqrt{(2-2)^2+(5-10)^2} = 5$

$A_3A_1 = \sqrt{(8-2)^2+(4-10)^2} = \sqrt{36+36} = \sqrt{72} = 8.485$

$B_2A_1 = \sqrt{(7-2)^2+(5-10)^2} = \sqrt{50} = 7.0710$

$B_3A_1 = \sqrt{(6-4)^2+(4-10)^2} = \sqrt{4+36} = \sqrt{40} = 6.324$

$C_2A_1 = \sqrt{(4-2)^2+(9-10)^2} = \sqrt{4+1} = \sqrt{5} = 2.236$

C_2 $B_1(5,8)$

$A_2B_1 = \sqrt{(2-5)^2+(5-8)^2} = \sqrt{9+9} = 4.2426$

$A_3B_1 = \sqrt{(8-5)^2+(4-8)^2} = \sqrt{9+16} = \sqrt{25} = 5$

$B_2B_1 = \sqrt{(7-5)^2+(5-8)^2} = \sqrt{4+9} = 3.605$

$B_3B_1 = \sqrt{(6-5)^2+(4-8)^2} = \sqrt{1+16} = \sqrt{17} = 4.1231$

$C_2B_1 = \sqrt{(4-5)^2+(9-8)^2} = \sqrt{1+1} = \sqrt{2} = 1.414$

C_3 $C_1(1,2)$

$A_2C_1 = \sqrt{(2-1)^2+(5-2)^2} = \sqrt{1+9} = \sqrt{10} = 3.1622$

$A_3C_1 = \sqrt{(8-1)^2+(4-2)^2} = \sqrt{49+4} = 7.28$

$B_2C_1 = \sqrt{(7-1)^2+(5-2)^2} = \sqrt{36+9} = 6.708$

$B_3C_1 = \sqrt{(6-1)^2+(4-2)^2} = \sqrt{25+4} = \sqrt{29} = 5.385$

$C_2C_1 = \sqrt{(4-1)^2+(9-2)^2} = \sqrt{9+49} = 7.615$

Clusters

$\{A_1\}$

$(2,10)$

$\{B_1, A_3, B_2, B_3, C_2\}$ ✓ ✓

$(5,8)$
$(8,4)$
$(7,5)$
$(6,4)$
$(4,9)$

$(6,6)$

New cluster head

$(6,6)$

$\{C_1, A_2\}$

$(1,2)$
$(2,5)$

$(1.5, 3.5)$ → New cluster head.

C_1	C_2	C_3
$A_1(2, 10)$	$B_1(5, 8)$	$C_1(1, 2)$
$A_2A_1 = 5$	$A_2B_1 = 4.2426$	$A_2C_1 = 3.1622$
$A_3A_1 = 8.485$	$A_3B_1 = 5$	$A_3C_1 = 7.28$
$B_2A_1 = 7.0710$	$B_2B_1 = 3.605$	$B_2C_1 = 6.708$
$B_3A_1 = 6.324$	$B_3B_1 = 4.1231$	$B_3C_1 = 5.385$
$C_2A_1 = 2.236$	$C_2B_1 = 1.414$	$C_2C_1 = 7.615$

Clusters

$\{A_1\}$, $\{B_1, A_3, B_2, B_3, C_2\}$ $\{C_1, A_2\}$

update cluster head

$(2, 10)$

$(5, 8)$
$(8, 4)$
$(7, 5)$
$(6, 4)$
$(4, 9)$
$(6, 6)$

$(1, 2)$
$(2, 5)$
$(1.5, 3.5)$

Repeat the same process from there new cluster heads.

K-mean clustering for IRIS data.

IRIS data - (150 X 4) (PL, PW, SL, SW)
containg - 3 classes of data.

1. Load the data.
2. Choose randomly three cluster heads from 150 tuples.
3. Calculate the Euclidean distance of 150 tuples from these cluster head and assign the tuple to that cluster from which the distance is minimum.
4. Update the cluster heads.
5. Repeat the steps 2 and 3 with the new cluster heads.
6. Stoping criteria - until the no. of data objects does not change in each cluster.

Comments on K-means

Strength Relatively efficient

$O(tkn)$, $k, t << n$

n = objects.
k = no. of clusters.
t = no. of iterations.

Weakness

1) Terminates at a local solution.
2) Need to specify k, the no. of clusters in advance.
3) Unable to handle noisy data and outliers.
4) Not suitable to discover clusters with non-convex shapes.

K-Medoid (partitioning around Medoids) (PAM), CLARA, CLARANS

1. Initialize: select k of the n data points as the medoids
2. Associate each data point to the closest medoid.
3. While the cost of the configuration decreases
   (a) For each medoid m, for each non-medoid data point, o
      > swap m and o, recompute the cost
      > if the total cost of the configuration increased in the previous step, undo the step

Ex cluster the following data set of 10 objects into two clusters i.e $K=2$

x_1	2, 6
x_2	3, 4
x_3	3, 8
x_4	4, 7
x_5	6, 2
x_6	6, 4
x_7	7, 3
x_8	7, 4
x_9	8, 5
x_{10}	7, 6

Step-1

Two observations $c_1 = x_2 = (3,4)$ and $c_2 = x_8 = (7,4)$ are randomly selected as medoids (cluster centers).

Step 2

Manhattan distances are calculated to each center to associate each data object to its nearest medoid.

Manhattan distance

$d(c_i, x_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|$

PAM is more robust than k-mean.

Work well for small data sets.

x_i	Distance to	
	$c_1(3,4)$	$c_2(7,4)$
(2,6)	3 ✓	7
(3,4)	0 ✓	4
(3,8)	4 ✓	8
(4,7)	4 ✓	6
(6,2)	5	3 ✓
(6,4)	3	1 ✓
(7,3)	5	1 ✓
(7,4)	4	0 ✓
(8,5)	6	2 ✓
(7,6)	6	2 ✓
cost	11	9

Since the points (2,6), (3,8) and (4,7) are closer to c_1 hence they form one cluster while the remaining points form another cluster.

Cluster 1 = {(3,4), (2,6), (3,8), (4,7)}

cluster 2 = {(7,4), (6,2) (6,4) (7,3) (8,5) (7,6)}

The total cost of this clustering is the sum of the distance between a data point and its cluster center =

$3+0+4+4+3+1+1+0+2+2 = 20$.

Step 3

select one of the nonmedoids O'

Let $x_7 = (7,3) = O'$

so now medoids are $c_1(3,4)$ and $O'(7,3)$

Calculate the total cost involved.

x_i	$c_1(3,4)$	$O'(7,3)$
x_1	3	8 ①
~~x_2~~		9 ⑥
x_3	4	7 ①
x_4	4	2 ⑥
x_5	5	2 ②
x_6	3	1 ②
x_7		
x_8	4	3 ⑧
x_9	6	3 ⑨
x_{10}	6	

total cost = $3+4+4+2+2+1+3+3$

$= 22$

so the cost of swapping medoid from c_2 to O' is

$S = 22 - 20 = 2 > 0$

so moving O' would be a bad idea.

So we try other non-medoid points to get minimum distance.