UNIT IV-Clustering Techniques

Topics : Clustering and Applications: Cluster analysis–Types of Data in Cluster Analysis–Categorization of Major Clustering Methods– Partitioning Methods, Hierarchical Methods– Density–Based Methods, Grid–Based Methods, Outlier Analysis.

Question 1 : Write a note on cluster analysis and applications of clustering ?

Cluster Analysis:

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression.

Cluster analysis tools based on k-means, k-medoids, and several methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SAS.

Applications:

1. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing.
2. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns.
3. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations.
4. Clustering may also help in the identification of areas of similar land use in an earth observation database and in the identification of groups of houses in a city according to house type, value, and geographic location, as well as the identification of groups of automobile insurance policy holders with a high average claim cost.
5. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity.

Question 2: Which type of data is used in Cluster Analysis ?

Main memory-based clustering algorithms typically operate on either of the following two data structures.

Data matrix (or object-by-variable structure):

This represents n objects, such as persons, with p variables (also called measurements or attributes), such as age, height, weight, gender, and so on. The structure is in the form of a relational table, or n-by-p matrix (n objects X p variables):

Dissimilarity matrix (or object-by-object structure):

This stores a collection of proximities that are available for all pairs of n objects. It is often represented by an n-by-n table:

The rows and columns of the data matrix represent different entities, while those of the dissimilarity matrix represent the same entity. Thus, the data matrix is often called a two-mode matrix, whereas the dissimilarity matrix is called a one-mode matrix. Many clustering algorithms operate on a dissimilarity matrix. If the data are presented in the form of a data matrix, it can first be transformed into a dissimilarity matrix before applying such clustering algorithms.

Interval-scaled (numeric) variables :

Interval-scaled (numeric) variables are continuous measurements of a roughly linear scale. Examples – weight and height, latitude and longitude coordinates (e.g., when clustering houses), and weather temperature.

The measurement unit used can affect the clustering Types of Data in Cluster analysis – For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead to a very different clustering structure.

Binary variables:

A binary variable has only two states: 0 or 1, where 0 means that the variable is absent, and 1 means that it is present.

Given the variable smoker describing a patient 1 indicates that the patient smokes and 0 indicates that the patient does not. Treating binary variables as if they are interval-scaled can lead to misleading clustering results. Therefore, methods specific to binary data are necessary for computing dissimilarities.

Categorical (nominal) variables:

A categorical (nominal) variable is a generalization of the binary variable in that it can take on more than two states. – Example: map_color is a categorical variable that may have five states: red, yellow, green, pink, and blue. The states can be denoted by letters, symbols, or a set Types of Data in Cluster Analysis The states can be denoted by letters, symbols, or a set of integers.

Ordinal variables:

A discrete ordinal variable resembles a categorical variable, except that the M states of the ordinal value are ordered in a meaningful sequence.

Example: professional ranks are often enumerated in a sequential order, such as assistant, associate, and full for professors. Ordinal variables may also be obtained from the discretization of interval-scaled quantities by splitting the value range into a finite number of classes. The values of an ordinal variable can be mapped to ranks.

Example: Suppose that we have the sample data:

There are three states for test-2, namely fair, good, and excellent.

3. Explain the classification of clustering methods in detail ?

a. Partitioning Methods
b. Hierarchical Methods
c. Density-Based Methods
d. Grid-Based Methods
e. Model-Based Methods

Partitioning Methods:

A partitioning method constructs k partitions of the data, where each partition represents a cluster and k <= n. That is, it classifies the data into k groups, which together satisfy the following requirements:

Each group must contain at least one object, and Each object must belong to exactly one group. A partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The general criterion of a good partitioning is that objects in the same cluster are close or related to each other, whereas objects of different clusters are far apart or very different.

Hierarchical Methods:

A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on

How the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one or until a termination condition holds.

The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each objectis in one cluster, or until a termination condition holds. Hierarchical methods suffer fromthe fact that once a step (merge or split) is done, it can never be undone. This rigidity is useful in that it leads to smaller computation costs by not having toworry about a combinatorial number of different choices.

Density-based methods:

Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering

clusters of arbitrary shapes. Other clustering methods have been developed based on the notion of density. Their general idea is to continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold; that is, for each data point within a given cluster, the neighbourhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise (outliers)and discover clusters of arbitrary shape.

DBSCAN and its extension, OPTICS, are typical density-based methods that grow clusters according to a density-based connectivity analysis. DENCLUE is a method that clusters objects based on the analysis of the value distributions of density functions.

Grid-Based Methods:

Grid-based methods quantize the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on the grid structure i.e., on the quantized space. The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space.

STING is a typical example of a grid-based method. Wave Cluster applies wavelet transformation for clustering analysis and is both grid-based and density-based.

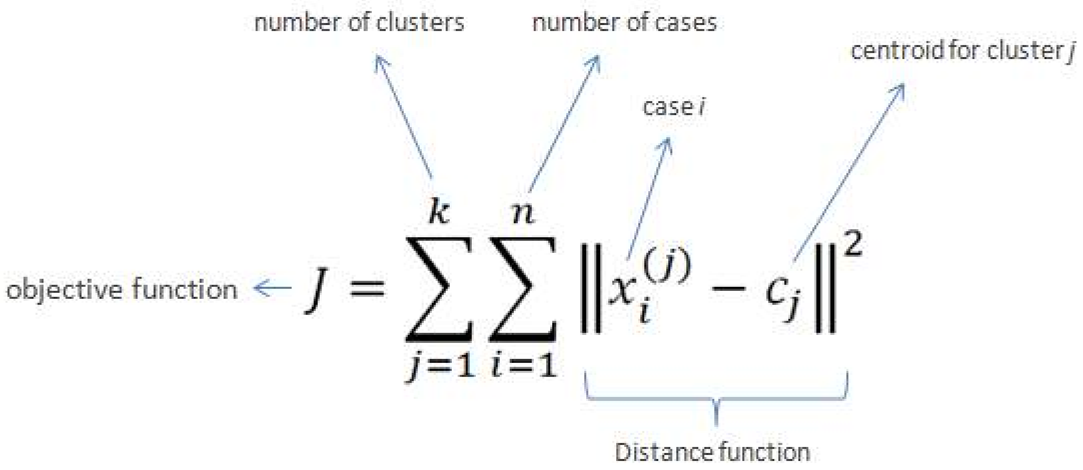
Model-Based Methods:

Model-based methods hypothesize a model for each of the clusters and find the best fit

of the data to the given model. A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points.It also leads to a way of automatically determining the number of clusters based onstandard statistics, taking noise or outliers into account and thus yielding robust clustering methods.

4.Explain K-Means Clustering algorithm with an example ?

K-Means clustering intends to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly *k* different clusters of greatest possible distinction. The best number of clusters *k* leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:



$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$

Algorithm

1. Clusters the data into *k* groups where *k* is predefined.
2. Select *k* points at random as cluster centers.
3. Assign objects to their closest cluster center according to the *Euclidean distance* function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

K-Means is relatively an efficient method. However, we need to specify the number of clusters, in advance And the final results are sensitive to initialization and often terminates at a local optimum. Unfortunately there is no global theoretical method to find the optimal number of clusters. A practical approach is to compare the outcomes of multiple runs with different *k* and choose the best one based on a predefined criterion.In general, a large *k* probably decreases the error but increases the risk of overfitting.

Example:

Suppose we want to group the visitors to a website using just their age (one-dimensional space) as follows:

n **= 19**

15,15,16,19,19,20,20,21,22,28,35,40,41,42,43,44,60,61,65

Initial clusters (random centroid or average):

***k* = 2**

$c_1 = 16$

$c_2 = 22$

$$Distance\ 1 = |x_i - c_1|$$

$$Distance\ 2 = |x_i - c_2|$$

Iteration 1:

$c_1 = 15.33$

$c_2 = 36.25$

x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	16	22	1	7	1	**15.33**
15	16	22	1	7	1	
16	16	22	0	6	1	
19	16	22	3	3	2	**36.25**
19	16	22	3	3	2	
20	16	22	4	2	2	
20	16	22	4	2	2	
21	16	22	5	1	2	
22	16	22	6	0	2	
28	16	22	12	6	2	
35	16	22	19	13	2	
40	16	22	24	18	2	
41	16	22	25	19	2	
42	16	22	26	20	2	
43	16	22	27	21	2	
44	16	22	28	22	2	
60	16	22	44	38	2	
61	16	22	45	39	2	
65	16	22	49	43	2	

Iteration 2:

c_1 = 18.56

c_2 = 45.90

x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	15.33	36.25	0.33	21.25	1	**18.56**
15	15.33	36.25	0.33	21.25	1	
16	15.33	36.25	0.67	20.25	1	
19	15.33	36.25	3.67	17.25	1	
19	15.33	36.25	3.67	17.25	1	
20	15.33	36.25	4.67	16.25	1	
20	15.33	36.25	4.67	16.25	1	
21	15.33	36.25	5.67	15.25	1	
22	15.33	36.25	6.67	14.25	1	
28	15.33	36.25	12.67	8.25	2	**45.9**
35	15.33	36.25	19.67	1.25	2	
40	15.33	36.25	24.67	3.75	2	
41	15.33	36.25	25.67	4.75	2	
42	15.33	36.25	26.67	5.75	2	
43	15.33	36.25	27.67	6.75	2	
44	15.33	36.25	28.67	7.75	2	
60	15.33	36.25	44.67	23.75	2	
61	15.33	36.25	45.67	24.75	2	
65	15.33	36.25	49.67	28.75	2	

Iteration 3:

c_1 = 19.50

c_2 = 47.89

x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	18.56	45.9	3.56	30.9	1	**19.50**
15	18.56	45.9	3.56	30.9	1	
16	18.56	45.9	2.56	29.9	1	
19	18.56	45.9	0.44	26.9	1	
19	18.56	45.9	0.44	26.9	1	
20	18.56	45.9	1.44	25.9	1	
20	18.56	45.9	1.44	25.9	1	
21	18.56	45.9	2.44	24.9	1	
22	18.56	45.9	3.44	23.9	1	
28	18.56	45.9	9.44	17.9	1	
35	18.56	45.9	16.44	10.9	2	**47.89**
40	18.56	45.9	21.44	5.9	2	
41	18.56	45.9	22.44	4.9	2	
42	18.56	45.9	23.44	3.9	2	
43	18.56	45.9	24.44	2.9	2	
44	18.56	45.9	25.44	1.9	2	
60	18.56	45.9	41.44	14.1	2	
61	18.56	45.9	42.44	15.1	2	
65	18.56	45.9	46.44	19.1	2	

Iteration 4:

$c_1 = 19.50$

$c_2 = 47.89$

x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	19.5	47.89	4.50	32.89	1	**19.50**
15	19.5	47.89	4.50	32.89	1	
16	19.5	47.89	3.50	31.89	1	
19	19.5	47.89	0.50	28.89	1	
19	19.5	47.89	0.50	28.89	1	
20	19.5	47.89	0.50	27.89	1	
20	19.5	47.89	0.50	27.89	1	
21	19.5	47.89	1.50	26.89	1	
22	19.5	47.89	2.50	25.89	1	
28	19.5	47.89	8.50	19.89	1	
35	19.5	47.89	15.50	12.89	2	**47.89**
40	19.5	47.89	20.50	7.89	2	
41	19.5	47.89	21.50	6.89	2	
42	19.5	47.89	22.50	5.89	2	
43	19.5	47.89	23.50	4.89	2	
44	19.5	47.89	24.50	3.89	2	
60	19.5	47.89	40.50	12.11	2	
61	19.5	47.89	41.50	13.11	2	
65	19.5	47.89	45.50	17.11	2	

No change between iterations 3 and 4 has been noted. By using clustering, 2 groups have been identified 15-28 and 35-65. The initial choice of centroids can affect the output clusters, so the algorithm is often run multiple times with different starting conditions in order to get a fair view of what the clusters should be.

5 . Write a note on K-Medoids Clustering Algorithm ?

Having an overview of K-Medoids clustering, let us discuss the algorithm for the same.

1. First, we select K random data points from the dataset and use them as medoids.
2. Now, we will calculate the distance of each data point from the medoids. You can use any of the Euclidean, Manhattan distance, or squared Euclidean distance as the distance measure.
3. Once we find the distance of each data point from the medoids, we will assign the data points to the clusters associated with each medoid. The data points are assigned to the medoids at the closest distance.
4. After determining the clusters, we will calculate the sum of the distance of all the non-medoid data points to the medoid of each cluster. Let the cost be C_i.
5. Now, we will select a random data point D_j from the dataset and swap it with a medoid M_i. Here, D_j becomes a temporary medoid. After swapping, we will calculate the distance of all the non-medoid data points to the current medoid of each cluster. Let this cost be C_j.
6. If $C_i > C_j$, the current medoids with D_j as one of the medoids are made permanent medoids. Otherwise, we undo the swap, and M_i is reinstated as the medoid.
7. Repeat 4 to 6 until no change occurs in the clusters.

K-Medoids Clustering Numerical Example With Solution

Now that we have discussed the algorithm, let us discuss a numerical example of k-medoids clustering.

The dataset for clustering is as follows.

Point	Coordinates
A1	(2, 6)
A2	(3, 8)
A3	(4, 7)
A4	(6, 2)
A5	(6, 4)
A6	(7, 3)
A7	(7,4)
A8	(8, 5)
A9	(7, 6)

A10	(3, 4)

Dataset For K-Medoids Clustering

Iteration 1

Suppose that we want to group the above dataset into two clusters. So, we will randomly choose two medoids.

Here, the choice of medoids is important for efficient execution. Hence, we have selected two points from the dataset that can be potential medoid for the final clusters. Following are two points from the dataset that we have selected as medoids.

- M1 = (3, 4)
- M2 = (7, 3)

Now, we will calculate the distance between each data point and the medoids using the Manhattan distance measure. The results have been tabulated as follows.

Point	Coordinates	Distance From M1 (3,4)	Distance from M2 (7,3)	Assigned Cluster
A1	(2, 6)	3	8	Cluster 1
A2	(3, 8)	4	9	Cluster 1
A3	(4, 7)	4	7	Cluster 1
A4	(6, 2)	5	2	Cluster 2
A5	(6, 4)	3	2	Cluster 2
A6	(7, 3)	5	0	Cluster 2
A7	(7,4)	4	1	Cluster 2
A8	(8, 5)	6	3	Cluster 2
A9	(7, 6)	6	3	Cluster 2
A10	(3, 4)	0	5	Cluster 1

Iteration 1

The clusters made with medoids (3, 4) and (7, 3) are as follows.

- Points in cluster1= {(2, 6), (3, 8), (4, 7), (3, 4)}
- Points in cluster 2= {(7,4), (6,2), (6, 4), (7,3), (8,5), (7,6)}

After assigning clusters, we will calculate the cost for each cluster and find their sum. The cost is nothing but the sum of distances of all the data points from the medoid of the cluster they belong to.

Hence, the cost for the current cluster will be 3+4+4+2+2+0+1+3+3+0=22.

Iteration 2

Now, we will select another non-medoid point (7, 4) and make it a temporary medoid for the second cluster. Hence,

- M1 = (3, 4)
- M2 = (7, 4)

Now, let us calculate the distance between all the data points and the current medoids.

Point	Coordinates	Distance From M1 (3,4)	Distance from M2 (7,4)	Assigned Cluster
A1	(2, 6)	3	7	Cluster 1
A2	(3, 8)	4	8	Cluster 1
A3	(4, 7)	4	6	Cluster 1
A4	(6, 2)	5	3	Cluster 2
A5	(6, 4)	3	1	Cluster 2
A6	(7, 3)	5	1	Cluster 2
A7	(7,4)	4	0	Cluster 2
A8	(8, 5)	6	2	Cluster 2
A9	(7, 6)	6	2	Cluster 2

A10	(3, 4)	0	4	Cluster 1

Iteration 2

The data points haven’t changed in the clusters after changing the medoids. Hence, clusters are:

- Points in cluster1:{(2, 6), (3, 8), (4, 7), (3, 4)}
- Points in cluster 2:{(7,4), (6,2), (6, 4), (7,3), (8,5), (7,6)}

Now, let us again calculate the cost for each cluster and find their sum. The total cost this time will be 3+4+4+3+1+1+0+2+2+0=20.

Here, the current cost is less than the cost calculated in the previous iteration. Hence, we will make the swap permanent and make (7,4) the medoid for cluster 2. If the cost this time was greater than the previous cost i.e. 22, we would have to revert the change. New medoids after this iteration are (3, 4) and (7, 4) with no change in the clusters.

Iteration 3

Now, let us again change the medoid of cluster 2 to (6, 4). Hence, the new medoids for the clusters are M1=(3, 4) and M2= (6, 4).Let us calculate the distance between the data points and the above medoids to find the new cluster. The results have been tabulated as follows.

Point	Coordinates	Distance From M1 (3,4)	Distance from M2 (6,4)	Assigned Cluster
A1	(2, 6)	3	6	Cluster 1
A2	(3, 8)	4	7	Cluster 1
A3	(4, 7)	4	5	Cluster 1
A4	(6, 2)	5	2	Cluster 2
A5	(6, 4)	3	0	Cluster 2
A6	(7, 3)	5	2	Cluster 2
A7	(7,4)	4	1	Cluster 2
A8	(8, 5)	6	3	Cluster 2
A9	(7, 6)	6	3	Cluster 2

A10	(3, 4)	0	3	Cluster 1

Iteration 3

Again, the clusters haven’t changed. Hence, clusters are:

- Points in cluster1:{(2, 6), (3, 8), (4, 7), (3, 4)}
- Points in cluster 2:{(7,4), (6,2), (6, 4), (7,3), (8,5), (7,6)}

Now, let us again calculate the cost for each cluster and find their sum. The total cost this time will be 3+4+4+2+0+2+1+3+3+0=22.

The current cost is 22 which is greater than the cost in the previous iteration i.e. 20. Hence, we will revert the change and the point (7, 4) will again be made the medoid for cluster 2.

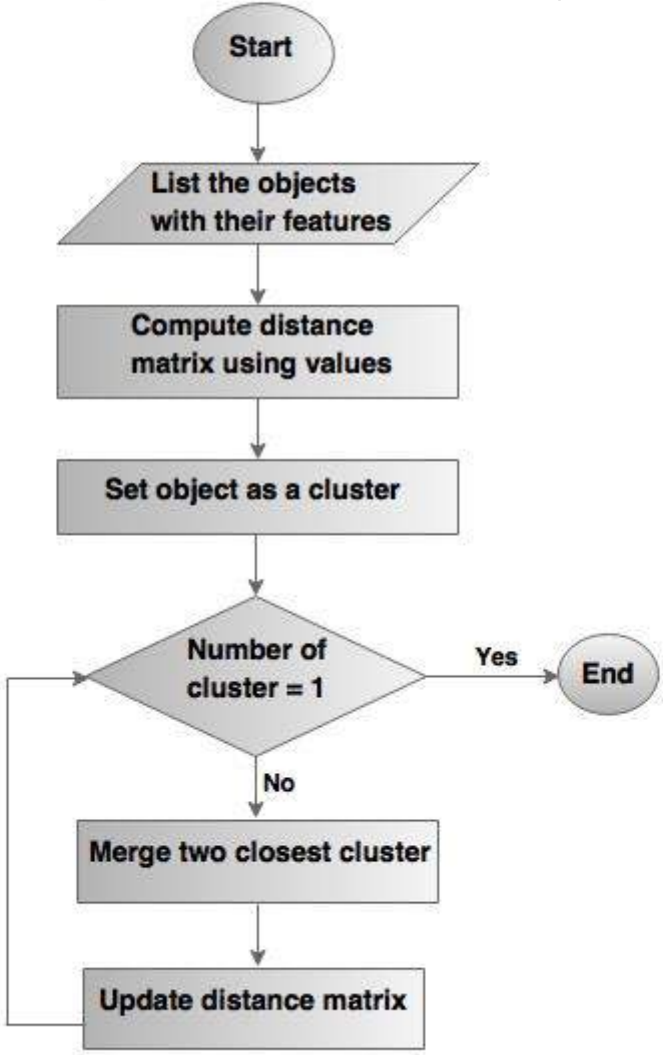
So, the clusters after this iteration will be cluster1 = {(2, 6), (3, 8), (4, 7), (3, 4)} and cluster 2= {(7,4), (6,2), (6, 4), (7,3), (8,5), (7,6)}. The medoids are (3,4) and (7,4).

We keep replacing the medoids with a non-medoid data point. The set of medoids for which the cost is the least, the medoids, and the associated clusters are made permanent. So, after all the iterations, you will get the final clusters and their medoids.

The K-Medoids clustering algorithm is a computation-intensive algorithm that requires many iterations. In each iteration, we need to calculate the distance between the medoids and the data points, assign clusters, and compute the cost. Hence, K-Medoids clustering is not well suited for large data sets.

6. How Agglomerative hierarchical clustering works ?

- It is a bottom-up approach, in which clusters have sub-clusters.
- The process is explained in the following flowchart.


Start
List the objects with their features
Compute distance matrix using values
Set object as a cluster
Number of cluster = 1
Yes
End
No
Merge two closest cluster
Update distance matrix

Agglomerative hierarchical clustering flowchart

Example: For given distance matrix, draw single link, complete link and average link dendrogram.

	A	B	C	D	E
A	0				
B	2	0			
C	6	3	0		
D	11	9	7	0	
E	9	8	5	4	0

Solution: Single link: From given distance matrix.

Step 1:

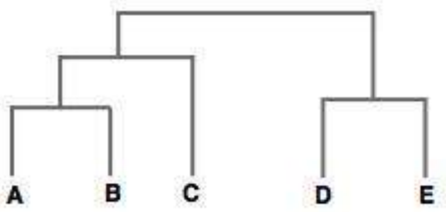
	A, B	C	D	E
A, B	0			
C	3	0		
D	9	7	0	
E	8	5	4	0

Step 2:

	A, B, C	D	E
A, B, C	0	0	
D	7	4	
E	5	0	0

Step 3:

	A, B, C	D, E
A, B, C	0	
D, E	5	0



Single link Dendrogram

2. Complete link: Find maximum distance to draw complete link.

Step 1:

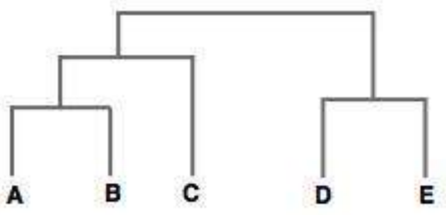
	A, B	C	D	E
A, B	0			
C	6	0		
D	11	7	0	
E	9	5	4	0

Step 2:

	A, B	C	D, E
A, B	0		
C	6	0	
D, E	11	7	0

Step 3:

	A, B, C	D, E
A, B, C	0	
D, E	11	0



Single link Dendrogram

3. Average link: To draw average link, compute the average link.

Step 1:

I) 6+3/2 = 4.5

ii) 11+9/2 = 10

iii) 9+8/2 = 8.5

	A, B	C	D	E
A, B	0			
C	4.5			
D	10	7	0	
E	8.5	5	4	0

Step 2:

I) 11+9+9+8/2 = 9.25

II) 7+5/ 2 = 6

	A, B	C	D, E
A, B	0		
C	4.5	0	
D, E	9.25	6	0

Step 3:

11+9.25+9.25+8+7+5/6 = 8.20

	A, B, C	D, E
A, B, C	0	
D, E	8.20	0

Comparison between Single link, Complete link and Average link based on distance formula.

Single link	Complete link	Average link
Handles non-elliptical shapes. It is sensitive to noise and outliers.	Less sensitive to noise and outliers.	It strikes a balance between single and complete links.

7. Explain Divisive Clustering method with an example ?

Divisive clustering **starts with one, all-inclusive cluster**. At each step, it **splits a cluster until each cluster contains a point** (or there are k clusters).

The following is an example of Divisive Clustering.

Distance	**a**	**b**	**c**	**d**	**e**
a	0	2	6	10	9
b	2	0	5	9	8
c	6	5	0	4	5
d	10	9	4	0	3
e	9	8	5	3	0

Step 1. Split whole data into 2 clusters

1. Who hates other members the most? (in Average)
   - A to others: mean(2,6,10,9)=6.75 →a(2,6,10,9)=6.75 → goes out! (Divide a into a new cluster)
   - B to others: mean(2,5,9,8)=6.0 (2,5,9,8)=6.0
   - C to others: mean(6,5,4,5)=5.0 (6,5,4,5)=5.0
   - D to others: mean(10,9,4,3)=6.5 (10,9,4,3)=6.5
   - E to others: mean(9,8,5,3)=6.25 (9,8,5,3)=6.25
2. Everyone in the old party asks himself: *“In average, do I hate others in old party more than hating the members in the new party?”*
   - If the answer is “Yes”, then he will also go to the new party.

	α==**distance to the old party**	β==**distance to the new party**	α−β
b	5+9+83=7.335+9+83=7.33	2	>0>0 (b also goes out!)
c	5+4+53=4.675+4+53=4.67	6	<0<0
d	9+4+33=5.339+4+33=5.33	10	<0<0
e	8+5+33=5.338+5+33=5.33	9	<0<0

3. Everyone in the old party ask himself the same question as above again and again until everyone’s got the answer “No”.

	α==**distance to the old party**	β==**distance to the new party**	α−β
c	…	…	<0<0
d	…	…	<0<0
e	…	…	<0<0

Step 2. Choose a current cluster and split it as in **Step 1.**

1. Choose a current cluster
   - o If split the cluster with the largest number of members, then the cluster {c,d,e} will be split.
   - o If split the cluster with the largest diameter, then the cluster {c,d,e} will be split.
   - o

cluster	**diameter**
{a,b}	2
{c,d,e}	5

2. Split the chosen cluster as in **Step 1.**

Step 3. Repeat **Step 2.** until each cluster contains a point (or there are k clusters)

8. Write a note on density based clustering ?

Density-based clustering refers to **unsupervised ML approaches that find discrete clusters in the dataset**, based on the notion that a cluster/group in a dataset is a continuous area of high point density that is isolated from another cluster by sparse regions. Typically in data points in the dividing, sparse zones are regarded as noise or outliers.

- **The issue of clustering is crucial in the field of data analysis.**

Data scientists utilize clustering for a wide variety of purposes such as pinpointing faulty servers, classifying genes based on expression patterns, spotting outliers in biological pictures, and many more.

You may be acquainted with some of the most common data clustering algorithm families: **DBSCAN and k-Means**. K-Means clusters point by allocating them to the closest centroid.

Applications

- Urban water distribution networks are a significant subsurface asset. Clusters of pipe ruptures and bursts may signal impending issues. Using the technique for clustering density, an engineer may locate these clusters and take preventative measures in high-risk areas of water supply networks.
- Consider that you have position data for every successful and failed NBA shot. The density-based clustering method may reveal the various patterns of successful and unsuccessful shot placements for each player. This data may then be used to guide the game's strategy.
- Hypothetically, you have a point dataset where each point represents a home in your research region, and some of those homes are plagued with pests while others are not. The greatest groups of infected homes may be located with the use of Density-based Clustering in r, narrowing down the search for an effective treatment and elimination strategy.
- As a result of geo-locating tweets after natural disasters or terrorist acts, rescue and evacuation requirements may be determined depending on cluster size and location.

Clustering Techniques

The Clustering Methods parameter of the density-based clustering tool gives three possibilities for locating clusters in your point data:

- **Defined distance (DBSCAN clustering)** is used to differentiate between dense clusters and sparser noise. The DBSCAN algorithm is the fastest of the clustering algorithms, but it can only be used if there is a clear Search Distance that applies to all candidate clusters and performs effectively. This implies that all significant clusters possess comparable densities. The Search Time Interval and Time Field parameters allow you to locate spatiotemporal groups of points.
- **Self-adjusting (HDBSCAN)** uses a range of distances to distinguish clusters of different densities from noise with sparser coverage. The HDBSCAN clustering algorithm is a data-driven approach that needs the least amount of human input.
- **Multi-scale (OPTICS)** uses the distance between nearby features to generate a reachability plot, which is subsequently utilized in distinguishing clusters of different densities from noise. The OPTICS technique provides the greatest versatility in fine-tuning the discovered clusters, but it is computationally costly, especially when the Search Distance is significant. You may use the approach to locate time and space clusters by using Search Time Interval and the Time Field parameters.

This tool requires Input Point Features, a route for Output Features, and a minimum amount of features necessary for a cluster to be evaluated. Depending on the chosen Clustering Method, you may need to supply additional parameters as indicated below.

9.Write about Grid-Based Clustering ?

Grid-Based Clustering method uses a multi-resolution grid data structure.

Several interesting methods

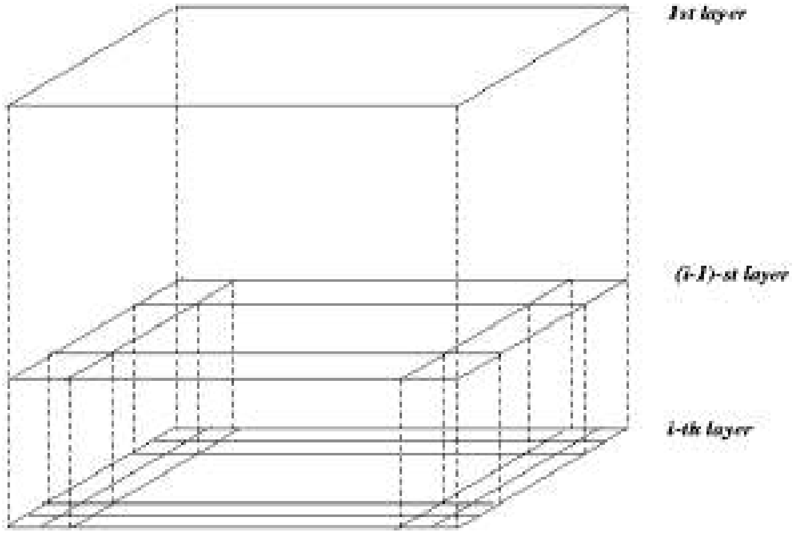
- **STING** (a **ST**atistical **IN**formation **G**rid approach) by Wang, Yang, and Muntz (1997)
- **WaveCluster** by Sheikholeslami, Chatterjee, and Zhang (VLDB’98) - A multi-resolution clustering approach using wavelet method
- **CLIQUE** - Agrawal, et al. (SIGMOD’98)

STING - A Statistical Information Grid Approach

STING was proposed by Wang, Yang, and Muntz (VLDB’97).

In this method, the spatial area is divided into rectangular cells.

There are several levels of cells corresponding to different levels of resolution.



For each cell, the high level is partitioned into several smaller cells in the next lower level.

The statistical info of each cell is calculated and stored beforehand and is used to answer queries.

The parameters of higher-level cells can be easily calculated from parameters of lower-level cell

- Count, mean, s, min, max
- Type of distribution—normal, uniform, etc.

Then using a top-down approach we need to answer spatial data queries.

Then start from a pre-selected layer—typically with a small number of cells.

For each cell in the current level compute the confidence interval.

Now remove the irrelevant cells from further consideration.

When finishing examining the current layer, proceed to the next lower level.

Repeat this process until the bottom layer is reached.

Advantages:

It is Query-independent, easy to parallelize, incremental update.

O(K), where K is the number of grid cells at the lowest level.

Disadvantages:

All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected.

WaveCluster

It was proposed by Sheikholeslami, Chatterjee, and Zhang (VLDB'98).

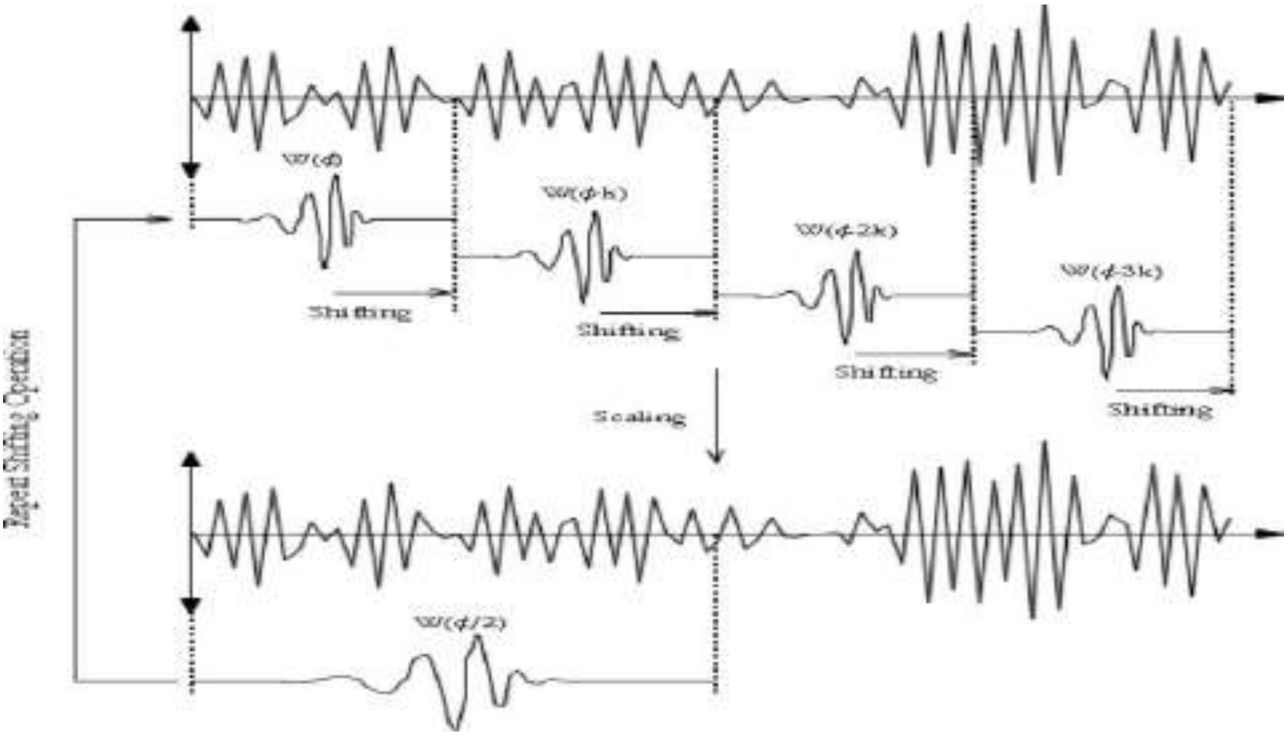
It is a multi-resolution clustering approach which applies wavelet transform to the feature space

- A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub-band.

It can be both grid-based and density-based method.

Input parameters:

- No of grid cells for each dimension
- The wavelet, and the no of applications of wavelet transform.



How to apply the wavelet transform to find clusters

- It summaries the data by imposing a multidimensional grid structure onto data space.
- These multidimensional spatial data objects are represented in an n-dimensional feature space.
- Now apply wavelet transform on feature space to find the dense regions in the feature space.
- Then apply wavelet transform multiple times which results in clusters at different scales from fine to coarse.

Why is wavelet transformation useful for clustering

- It uses hat-shape filters to emphasize region where points cluster, but simultaneously to suppress weaker information in their boundary.
- It is an effective removal method for outliers.
- It is of Multi-resolution method.
- It is cost-efficiency.

Major features:

- The time complexity of this method is O(N).
- It detects arbitrary shaped clusters at different scales.
- It is not sensitive to noise, not sensitive to input order.
- It only applicable to low dimensional data.

CLIQUE - Clustering In QUEst

It was proposed by Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98).

It is based on automatically identifying the subspaces of high dimensional data space that allow better clustering than original space.

CLIQUE can be considered as both density-based and grid-based:

- It partitions each dimension into the same number of equal-length intervals.
- It partitions an m-dimensional data space into non-overlapping rectangular units.
- A unit is dense if the fraction of the total data points contained in the unit exceeds the input model parameter.
- A cluster is a maximal set of connected dense units within a subspace.

Partition the data space and find the number of points that lie inside each cell of the partition.

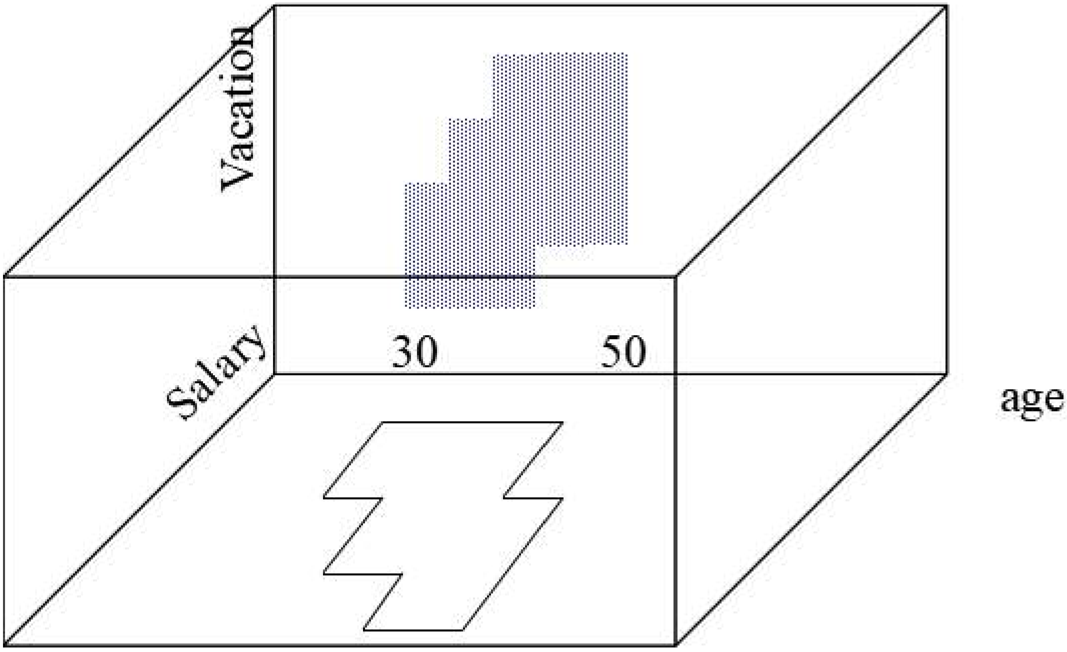
Identify the subspaces that contain clusters using the Apriori principle.

Identify clusters:

- Determine dense units in all subspaces of interests.
- Determine connected dense units in all subspaces of interests.

Generate minimal description for the clusters:

- Determine maximal regions that cover a cluster of connected dense units for each cluster.
- Determination of minimal cover for each cluster.



10.Explain outlier analysis in data mining ?

Outliers in Data mining is a very hot topic in the field of data mining. Let's discuss the outliers.

The data which deviates too much far away from other data is known as an outlier. The outlier is the data that deviate from other data.

The outlier shows variability in an experimental error or in measurement. In other words, an outlier is a data that is far away from an overall pattern of the sample data.

Outliers can indicate that the population has a heavy-tailed distribution or when measurement error occurs.

Outliers can be categorized as;

1. Collective outliers.
2. Point outliers
3. Contextual outliers

Collective outliers can be subsets of outliers when we introducing the novelties in data. For example, a signal that may indicate the discovery of a new phenomenon for the data set.

Point outliers are the data points that are far from the other distribution of the data.

Contextual outliers are the outliers just like noisy data. One example of noise data is when data have a punctuation symbol and suppose we are analyzing the background noise of the voice when doing speech recognition.

There are two types of Outliers.

1. Univariate outliers
2. Multivariate outliers

A univariate outlier is a data outlier that differs significantly from one variable. A multivariate outlier is an outlier when a combination of values on two or more than two variables have a significant difference. The univariate outlier and Multivariate outliers can influence the overall outcome of the data analysis.

Outliers can have many different causes. Some of these causes are mentioned below.

- Ther instruments used in the experiments for taking measurements suddenly malfunctioned.
- The error in data transmission.
- Due to changes in system behavior.
- Due to fraudulent behavior
- Due to human error
- Due to natural deviations in populations.
- Due to flaws in the assumed theory.
- Incorrect data collection.

Algorithm to Detect Outlier in data mining.

1. Calculate the mean of each cluster of the data.
2. Initialize the Threshold value of the data.

3. Calculate the distance of the test data from each cluster mean
4. Find the nearest cluster to the test data
5. Now, if we found that Distance is greater than Threshold, then it is a signal of Outlier.

There are many methods of outlier detection. Some of the outlier detection methods are mentioned below;

- Z-Score Normalizatoin
- Linear Regression Models (PCA, LMS)
- Information Theory Models
- High Dimensional Outlier Detection Methods (high dimensional sparse data)
- Proximity Based Models (non-parametric)
- Probabilistic and Statistical Modeling (parametric)
- Probabilistic and Statistical Modeling (parametric)
- Numeric Outlier

Numeric Outlier

Numeric Outlier is the nonparametric outlier detection technique in a one-dimensional feature space. TheNumeric outliers calculation can be performed by means of the InterQuartile Range (IQR).

Z-Score

Z-score is a data normalization technique and assumes a Gaussian distribution of the data. Outliers detection can be performed by Z-Score.

DBSCAN

The DBSCAN technique is based on the DBSCAN clustering algorithm. DBSCAN is a density-based, nonparametric outlier detection technique in a 1 or multi-dimensional feature space. In DBSCAN, all the data points are defined in the following points.

1. Core Points
2. Border Points
3. Noise Points.