11/9/19

Cluster Analysis

For ex: At a bank,
Customers - 100 (Age, Income, Area, Crown
Cust. manager - 4 ⟸ Input para. consist of the 4 attr. of cust. mentioned
Serves as clusters

→ Clustering is the process of grouping the data into classes or clusters so that objects with cluster have high similarity in comparison to one another but, varying dissimilarities with objects of other clusters.

⇒ Clustering gives dense or sparse representation of the data.

Generally, in Statistics distance-based clustering algo. are used.

→ Classification classifies the given data into pre-defined labels only. It is supervised learning.
Whereas, clustering is unsupervised learning.

Requirements of clustering in DM

1) Scalability - The algo. should be able to work on any no. of data / tuples.

2) Ability to deal with different type of attributes - Multiple attr. should be

processed simultaneously.

3) Identifying clusters with arbitrary shape

4) Minimal / less knowledge for input parameters
- clusters should be identified by using a min. no. of attr. or knowledge base

5) Deals with noisy data
- Preprocessing should be performed in cases where noisy data is present

6) Insensitivity towards the ordering of input parameters
- whether data is processed from 1 to 100 or 100 to 1, the output should be same
even if some data is added, it should be processed in a similar way as its predecessors.

7) High dimensionality
- should be able to handle data from multiple tables or data in same table with multiple attr.

8) Constraint-based clustering

9) Interpretability & usability
- Output of clustering should be

easy to interpret and easy to use

Types of data used in cluster analysis:

1) Data Matrix
- Represented as 'Object-by-variable'
- Person (n) , Age, Income (p)
- ∴ (n x p) matrix would be generated.

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

2) Dissimilarity matrix
- Represented as 'Object-by-Object'
- It shows the dissimilarity between two objects.

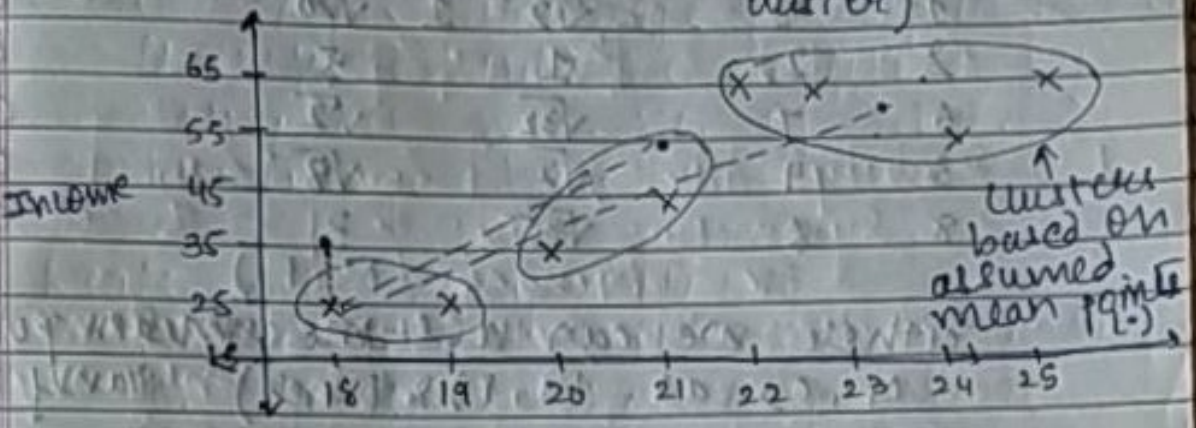
Same as $d(2,1)$

	1	2	3	...	n
1	0				
2	$d(2,1)$	0			
3	$d(3,1)$	$d(3,2)$			
⋮					
n	$d(n,1)$	...			0

Major clustering methods:
1) Partitioning method
2) Hierarchical method
3) Density-based method
4) Grid-based method
5) Model-based method

Partitioning Method.

- Distance-based algo. used.
  - i) K-mean // K = no. of clusters.
  - ii) K-medoid (mean = mean of particular cluster)



Assume k = 3

(a) Mark 3 (since k=3) random points on the graph

(b) Measure distance of every point (x) from the three random points (•) (Use Euclidean dist. formula)

(c) Cluster the points having roughly the same avg. dist. from the random point (•)

(d) These steps are iterated for various 3-random points until a min. dist. from the mean of cluster is obtained.
(Typically, iterated 21 times)

Eg: (2,3) (1,4) (3,4) k=2
(5,6) (2,2) (8,6)
(8,7) (6,7)

⇒

x	y	$C_1(2,3)$	$C_2(5,6)$	Cluster
2	3	0	$9\sqrt{2}$	C_1
5	6	$9\sqrt{2}$	0	C_2
8	7	$\sqrt{52}$	$\sqrt{10}$	C_2
1	4	$\sqrt{2}$	$\sqrt{20}$	C_1
2	2	1	5	C_1
6	7	$\sqrt{32}$	$\sqrt{2}$	C_2
3	4	$\sqrt{2}$	$\sqrt{8}$	C_1
8	6	$\sqrt{45}$	3	C_2

Either we can assume mean points C_1, C_2 as (1,1) & (3,4) (any random)

OR

we can assume C_1, C_2 as (2,3) & (5,6)

C1	C2
(2,3)	(5,6)
(1,4)	(8,7)
(2,2)	(6,7)
(3,4)	(8,6)

new $C_1 = \left(\frac{2+1+2+3}{4}, \frac{3+4+2+4}{4}\right)$ new $C_2 = \left(\frac{27}{4}, \frac{26}{4}\right)$

$= (2, 3.25)$ $= (6.75, 6.5)$

Iteration 2:

x	y	$C_1(2, 3.25)$	$C_2(6.75, 6.5)$	Cluster
2	3	0.25	5.9	C_1
5	6			
8	7			
1	4			
2	2			
6	7			

3 4

8 6

18/9/18

Types of attributes possible in data

i) Nominal attr. (Related to name/category)
ii) Binary attr.
iii) Ordinal
iv) Numerical
v) Mixed

Proximity measure = How the obj. are alike or unlike when compared to one another.

(i) ⇒ Let's assume two objects, i & j

$d(i,j) = \frac{P-m}{P}$ | $s(i,j) = 1 - d(i,j)$ [similarity measure]

P = Total no. of attr
m = Total no. of matched attr

Object ID	Color	Grade
1	Red	A
2	Yellow	B
3	Red	A
4	Blue	C

$d(1,2) = \frac{2-0}{2} = 1$ $s(1,2) = 0$.

$d(1,3) = \frac{2-2}{2} = 0$

$d(1,4) = \frac{2-0}{2} = 1$

$d(2,3) = \frac{2-0}{2} = 1$

$d(2,4) = \frac{2-0}{2} = 1$

$d(3,4) = \frac{2-0}{2} = 1$

(ii) Binary attr.

- Symmetric attr.

Ex - Grender (M, F)

Both represent same attr.

2 can take only these values

- Asymmetric attr.

Ex - Fever (+ve, -ve)

		object j	
		1	0
Object i	1	q	r
	0	s	t

Symmetric binary

$d(i,j) = \frac{r+s}{q+r+s+t}$

Asymmetric binary

$d(i,j) = \frac{r+s}{q+r+s}$

[Jaccard coefficient]

Name	Fever	Cough	T1	T2	T3	T4
Jack	Y	N	P	N	N	N
Mary	Y	N	P	N	P	N
Jim	Y	Y	N	N	N	N

All are asymmetric attr.

	Jack	
	2	1
Mary	0	3

$d(\text{Jack, Mary}) = \frac{1+0}{2+1+0}$

$= \frac{1}{3} = 33.3\%$

	Jack	
	1	1
Jim	1	3

$d(\text{Jack, Jim}) = \frac{1+1}{1+1+1} = \frac{2}{3}$

$= 66.6\%$.

	Mary	
	1	1
Jim	2	2

$d(\text{Mary, Jim}) = \frac{1+2}{1+1+2} = \frac{3}{4}$

$= 75\%$.

(iii) Ordinal attr.

Order is given as attr. value
Ex- low, medium, high or 1st, 2nd, 3rd

ID	Grade
1	High
2	Low
3	High
4	Med

Normalize ⇒

ID	Grade
1	1
2	0
3	1
4	0.5

$R_{if} = \{1, 2, 3, \ldots M_f\}$; M_f = Total stage (Here, 3)

$R_{if} = \{1, 2, 3\}$

High ↑ Med ↑ low

$Z_{if} = \frac{R_{if} - 1}{M_f - 1}$

∴ for high, $Z_{if} = \frac{1-1}{3-1} = 0$.

Using Manhattan dist.

$d(A,B) = |1| + |-2| + |-10| + |5|$

$= 1+2+10+5$

$= 18$

16/9/19

a) Interval scaled
- Linear measurement.

Temp

30°) it not half of 60°. Just
60°) represents a range.
0°

b) Ratio Scaled.

Ex:	Income	
	25K	
	50K	// Here, 50K is considered as double of 25K. (i.e. ratios are involved)
	75K	

Converting interval scaled values to standardized values.

1) Mean absolute deviation

$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$

s_f ↑ var. of attr.

m_f = mean value of f.

$= \frac{1}{n}(x_{1f} + x_{2f} + \dots x_{nf})$

2) Standardized measurement.

$z_{if} = \frac{x_{if} - m_f}{s_f}$ // calculated for each value of an attr.

PAGE NO.
DATE / /

Generalized eq. of Euclidean + Manhattan:

$$d(i,j) = \left[(|x_{i1} - x_{j1}|)^q + (|x_{i2} - x_{j2}|)^q + \dots + (|x_{ip} - x_{jp}|)^q \right]^{1/q}$$

$q = 1$: Manhattan dist.
$q = 2$: Euclidean dist.

Ex: Calculate dissimilarity matrix for:

A (7, 8, 3)
B (9, 9, 5)
C (7, 6, 1)

⇒ Total 6 distances would be generated.

$$d(A,B,C) = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 0 & 3 & \sqrt{8} \\ & 0 & \sqrt{29} \\ & & 0 \end{bmatrix} \end{array}$$

→ $d(i,j) \geq 0$
$d(i,i) = 0$
$d(i,j) = d(j,i)$
$d(i,j) \leq d(i,h) + d(j,h)$

$d(A,B) = \sqrt{2^2 + 1^2 + 2^2} = 3$

$d(B,C) = \sqrt{2^2 + 3^2 + 4^2} = \sqrt{29}$

$d(A,C) = \sqrt{0 + 2^2 + 2^2} = \sqrt{8}$

$d(A,A) = d(B,B) = d(C,C) = 0$

Ratio-scaled (3 methods)

i) Treat the ratio-scaled variable like the interval-scaled variable.

ii) Apply logarithm func for attr. value & then apply interval scaled technique.

iii) Convert attr. values to continuous ordinal values & then apply interval scaled techniques.

(v) Mixed Variable

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} \times d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

f: attr. value

where, $\delta_{ij}^{(f)} = 0$ if i) x_{if} or x_{jf} is missing

2) $x_{if} = x_{jf} = 0$ & f is asymm. binary

Otherwise, $\delta_{ij}^{(f)} = 1$.

$\Rightarrow$ i) f is binary or nominal

$d_{ij}^{(f)} = 0$, if $x_{if} = x_{jf}$

$d_{ij}^{(f)} = 1$, otherwise

2) f is interval scaled

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$$

3) f is ordinal or ratio scaled

x_{ij} & $y_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{x_j}}$

Conditions for partitioning:
i) Every object must belong to only one cluster.
ii) Every cluster must have atleast one object.

k-Medoids

The element having minimum dissimilarity with other ele. of the cluster.

	C_1	C_2	Cluster
A(8,7)	$\sqrt{2}$	$\sqrt{20}$	C_1
B(3,7)	$\sqrt{37}$	$\sqrt{5}$	C_2
C(4,9)	$\sqrt{34}$	4	C_2
C_1 D(9,6)	0	$\sqrt{26}$	C_1
E(8,5)	$\sqrt{4}$	4	C_1
F(5,3)	5	$\sqrt{5}$	C_2
G(7,8)	$\sqrt{13}$	$\sqrt{13}$	C_1 / C_2
H(8,4)	$\sqrt{5}$	$\sqrt{17}$	C_1
I(7,5)	$\sqrt{5}$	3	C_1
C_2 J(4,5)	$\sqrt{26}$	0	C_2

k=2

$K_1 = \{A, D, E, G, H, I\}$

$K_2 = \{B, C, F, J\}$

Cost = $\Sigma \{(DA), (DE), (DG), (DH), (DJ)\}$

+

$\Sigma \{(JB), (JC), (JF)\}$

$= (\sqrt{2} + \sqrt{2} + \sqrt{13} + \sqrt{5} + \sqrt{26}) + (\sqrt{5} + 4 + \sqrt{5})$

$D_1 = 19.37$

Repeat the steps by choosing a different C_1 or C_1 or both C_1 & C_2.

If D_2 comes out to be more than D_1, then it is considered as a bad decision and we stick with the earlier one.

18/9/19

PAM (Partitioning around medoid)

k clusters | n objects

k=2

Random selection: C_1, C_2

Next, two new non-medoid points are taken as medoids and the dist from each point calculated. If it comes out to be less than the previous medoids, the new ones are considered & the process reiterated.

(PAM)

Variations of k-medoid: k-median.
k-mode.

⇒ In presence of noise/outliers, k-medoid works better; else k-mean is mostly preferred.

Both the methods aren't sufficient or fulfilling when we have large amt. of data.

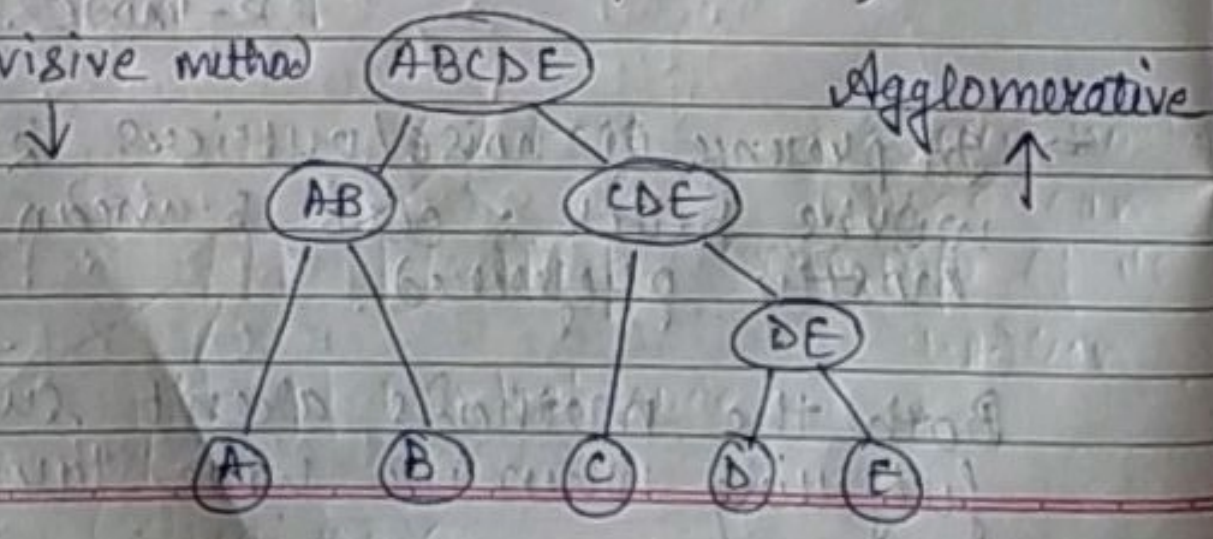
CLARA (Clustering Large Application)
Sampling + PAM

- The selection of samples is important
  1-50 - 51 - 100
  ↓ 15 samples ↓ 5 samples // not good solution. Should be balanced

$O(ks^2 + k(n-k))$ // complexity of CLARA
Where, s = sample size
k = no. of clusters
n = total no. of objects.

Improving efficiency of CLARA: Use CLARANS.
Here, new (random) samples are generated for every iteration
[In CLARA, samples are fixed for every iteration]

Hierarchical Clustering (method)
1) Agglomerative (Bottom-up)
2) Divisive (Top-down)



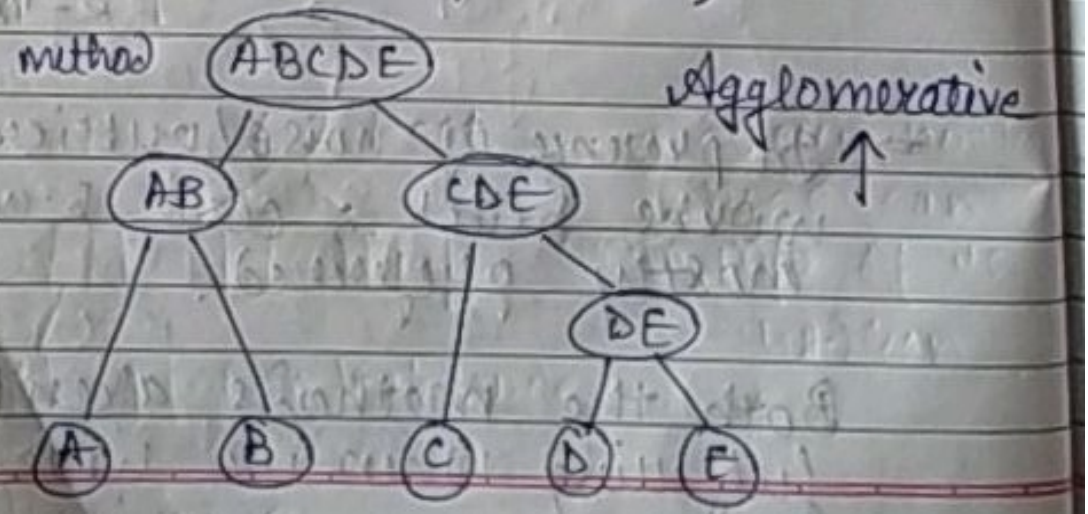
CLARA (Clustering Large Application)
Sampling + PAM

- The selection of sample is important
  1-50 - 51-100
  ↓ 15 samples ↓ 5 samples // not good solution. Should be balanced

$O(ks^2 + k(n-k))$ // complexity of CLARA
where, s = sample size
k = no. of clusters
n = total no. of objects.

Improving efficiency of CLARA: Use CLARANS.
Here, new (random) samples are generated for every iteration.
[In CLARA, samples are fixed for every iteration].

Hierarchical Clustering (method)
1) Agglomerative (Bottom-up)
2) Divisive (Top-down)



Algo. for Divisive: DIANA
(Divisive Analysis)
Algo. for Agg: AGNES
(Agglomerative Nesting)

No. of clusters is the termination point for both the methods.

Dist. b/w 2 clusters: $|P - P'|$ $P \in C_1$, $P' \in C_2$

Four dist. to be considered:

i) Min. distance

$$\min(C_i, C_j) = \min(|P - P'|)$$

ii) Max. dist.

$$\max(C_i, C_j) = \max(|P - P'|)$$

iii) Mean dist (mean of C_i)

$$\text{mean}(C_i, C_j) = |m_i - m_j|$$

iv) Avg. dist:

$$\text{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{P \in C_i} \sum_{P' \in C_j} |P - P'|$$

Limitations:

1) Selecting spliting & merging point. (cannot be undone)
2) Scalability
(Splitting & merging are costly operations)

Efficiency of hierarchical method can be improved by combining it with another method : BIRCH (Balanced Iterative Reducing and clustering using hierarchies)

- Integrated Hierarchical Method :
  i) clustering feature (CF)
  ii) clustering feature tree (CFT)

CF

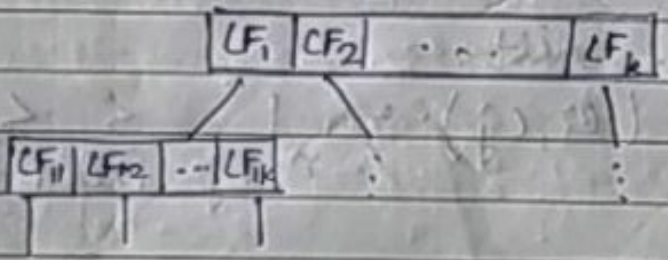
- Summarization triplets which include three factors :

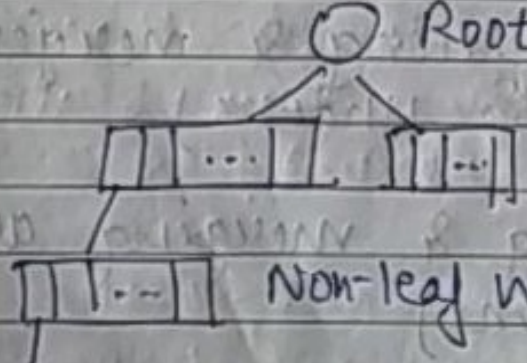
$(N, \vec{LS}, SS)$

N : No. of obj
$\vec{LS}$: linear sum
SS : sq sum.

A (3,2)
B (5,4)
} $CF = (2, (8,6), (34,20))$

CF-Tree (Height-balanced tree)





Every non-leaf node can contain CF info of its children.

Imp factors of BIRCH:
Branching factor (B): No. of children.
Threshold: Diameter of children.

⇒ BIRCH algo. phases:
Phase-1: Scan the database (dataset) & build the CF-tree. It is a dynamic process. (Data is added into the tree as it is read from the dataset). Complexity: O(n)

second scan can be used to remove noise.

Phase-2: Apply selected clustering Algo. Data is added in the nearest cluster (at leaf node). When diameter increases, we can further split the cluster.

CURE (Clustering Using REpresentative): Centroid can be used as a representative for every cluster.

Steps:
1) Draw random shape, S
2) Create partitions of the shape.
3) Cluster each partition (find centroid of each cluster).
4) Eliminate outliers by random sampling of the clusters. (Entire cluster is eliminated as outlier)
5) Cluster the partial cluster.
6) Mark appropriate cluster with req. obj (Shrink shape of cluster by finding mean centroid from centroids of sub cluster)

25/9/19

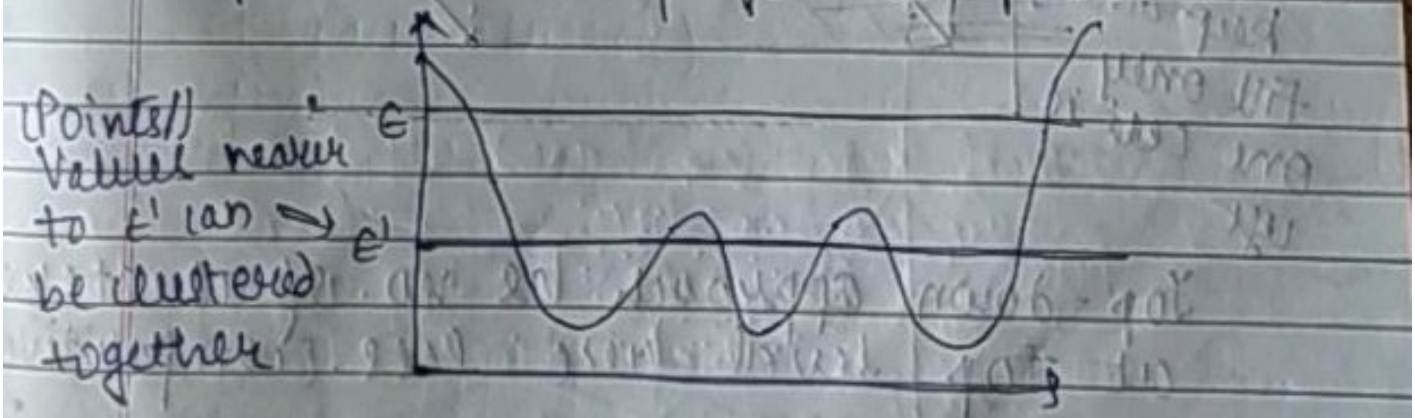
OPTICS :-

1) Core distance ϵ'
(Min. dist. to make P the core point)

P = core point
min. point

2) Reachability distance
(Max. value between (among) Euclidean dist. & core dist.)

After calculating for all points :

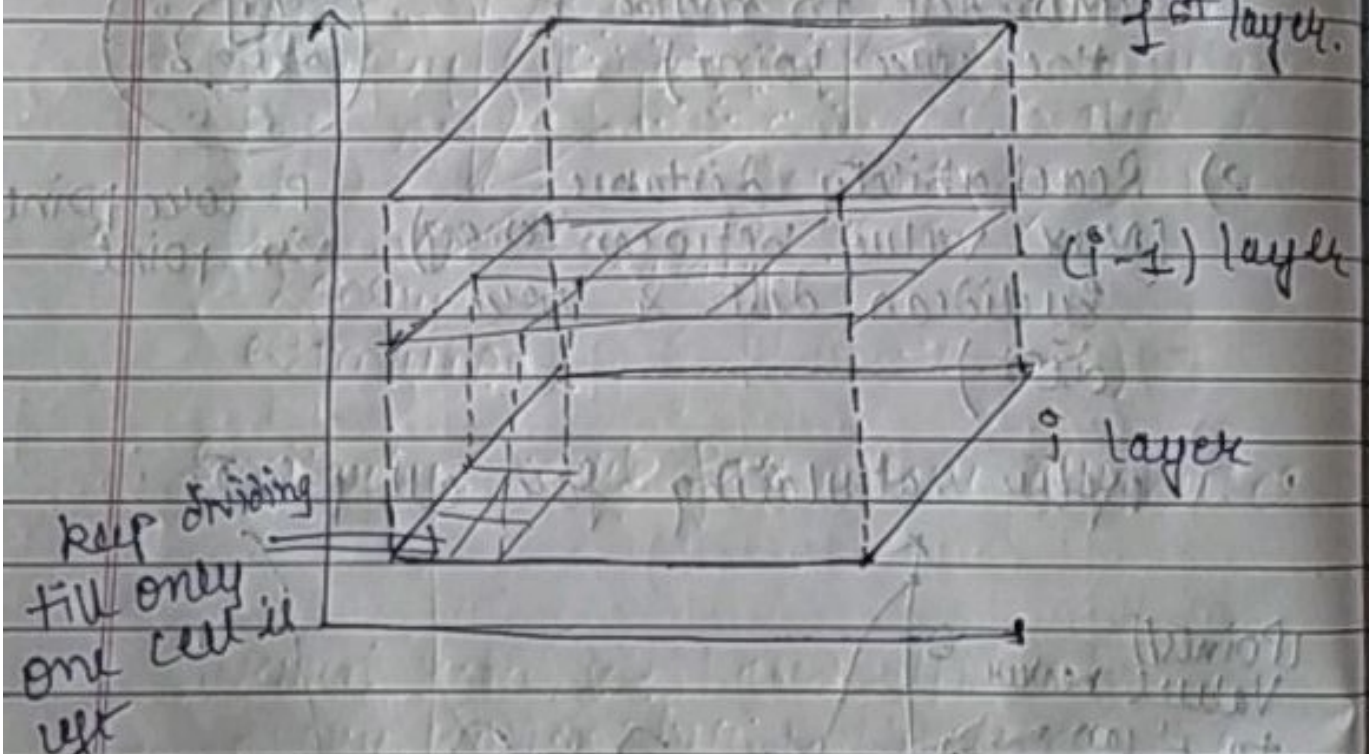


Use ordering algo. on any attr.
automatic / user-defined

Any random shape can be generated when density-based clustering is used.

4) Grid-based clustering
Square shaped clusters can be generated.

Methods:
STING (Statistical Information Grid)
Combination of hierarchical and grid based clustering.
Make use of statistical data such as min., max. and mean. Attr, SD.



Top-down approach as no. of clusters at top level are less (comparitively).

For each level, confidence interval for

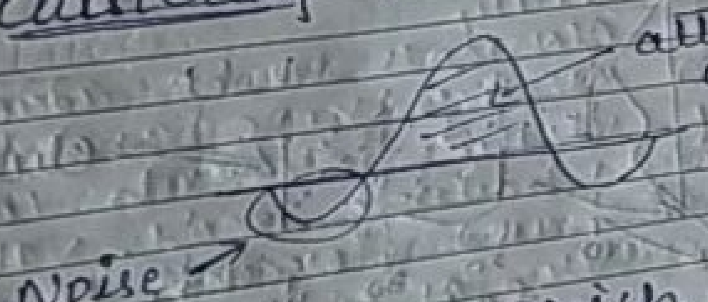
calculated

Eliminate irrelevant cells based on the confidence interval.

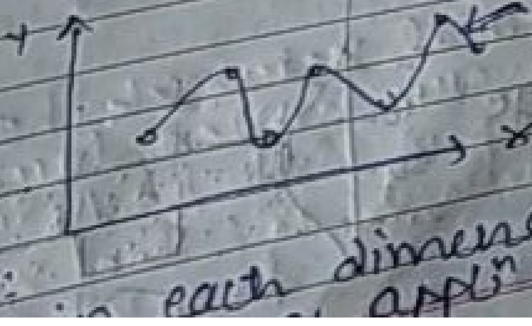
Every level needs to store the 4 statistical info. This increases the independence between levels.

Wavelet Transformation - Represent a wave as a sum of waves of various frequencies.

Clustering using wavelet trans.



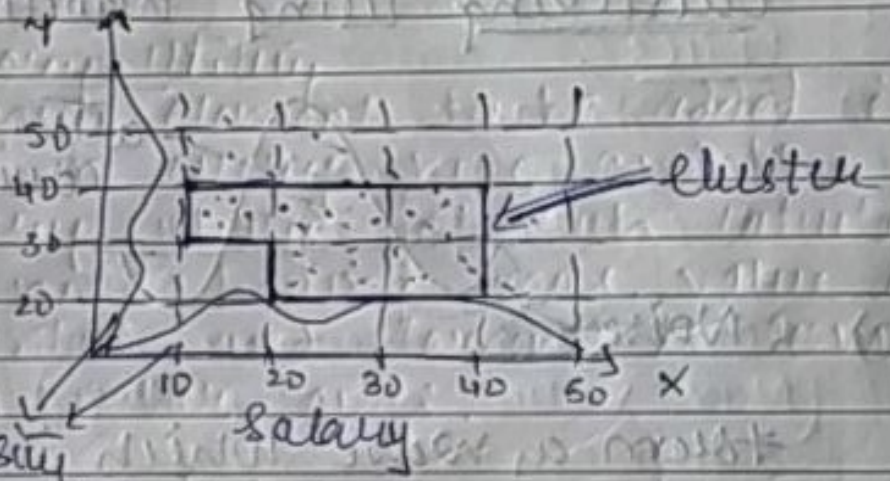
Form a wave which includes all the points in the dataset. Then apply wavelet transform and compare the results with the above figure.



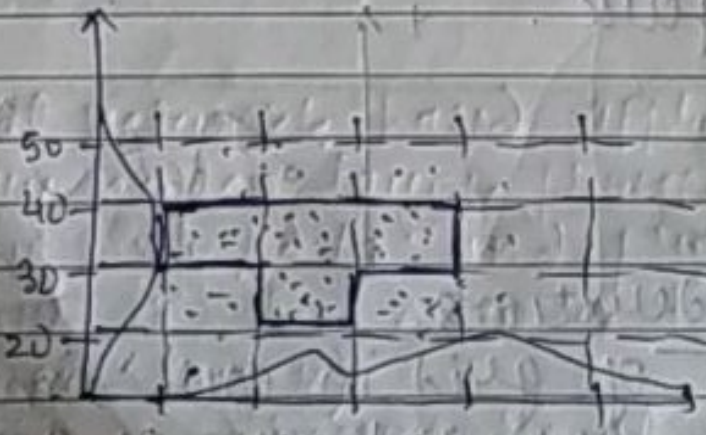
tions: in each dimension appl^n of wavelet

Arbitrary shape clusters can be generated.
Order insensitive algo. as the order does not matter (as the points are plot in a grid & the order of plotting does not matter)
[combination of grid + density based clustering]

CLIQUE (clustering high-dimensional space) used for high dimensional data
combination of grid + density based



Similarly, mark for other attributes



Finally take intersection of both the clusters.

5) Model-based clustering
Mixture of unsupervised & supervised learning
(The other four methods are fall under the category of unsupervised learning)

17/9/19 It tries to fit given data into some mathematical model.
The output of first step (unsupervised learning) are clusters which are then used to build a model (supervised learning).

i) Statistical method / approach
- Uses conceptual learning.
Clusters → concepts of the clusters
↑ extracting features of the data.

Conceptual clustering:
I/P: Unlabelled objects (clusters)
O/P: Classification model based on characteristics / features of obj.

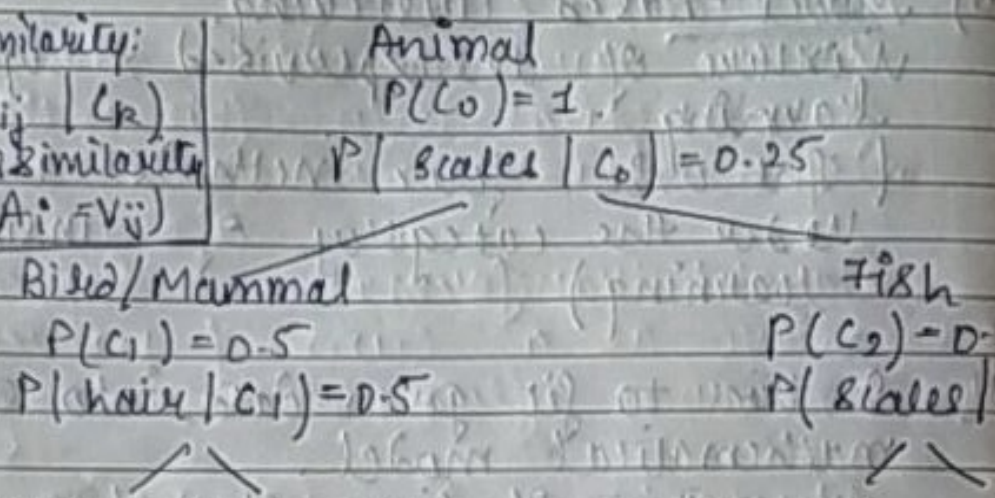
COBWEB

$P(A_i = V_{ij} \mid C_k)$

A_i ← Attribute (attr-value pair)
V_{ij} ← value
C_k ← concept class.

Classification, has two parameters:
i) Concept
ii) Probabilistic description of that concept

Intraclass similarity:
$P(A_i = V_{ij} \mid C_k)$
Interclass dissimilarity
$P(C_k \mid A_i = V_{ij})$



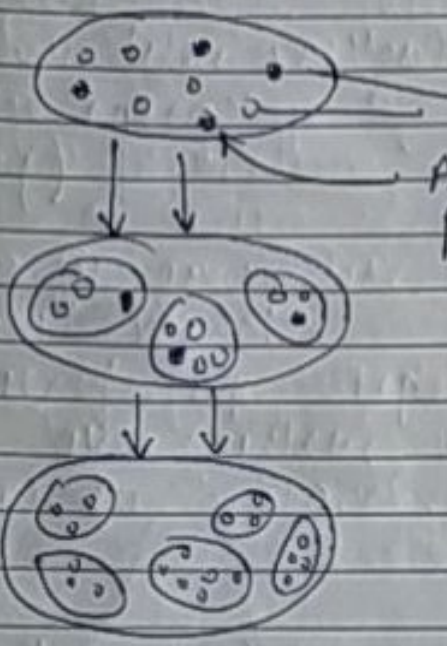
Category utility
Any new element to be added is added in every node and the category utility for each calculated. The node having max. value is kept intact and element is removed from other nodes.

$$\sum_{k=1}^{n} P(C_k) \left[\sum_i \sum_j P(A_i = V_{ij} \mid C_k)^2 - \sum_i \sum_j P(A_i = V_{ij})^2 \right]$$

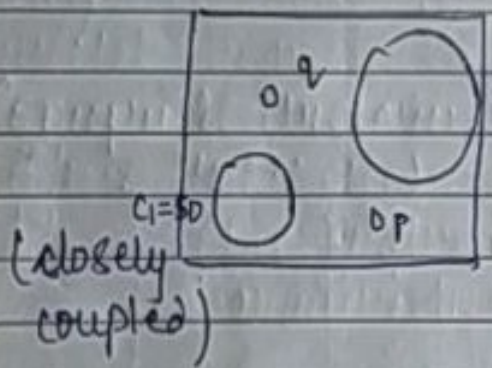
Neural network based approach
Every hidden layer processes the input info./data and propagates it further in the network (to next hidden layer).
It is a hierarchical approach

but, cannot be traversed bottom to top (i.e. from O/P layer to I/P layer)

Types of methods: a) competitive
b) SOM (Self-organizing



Outlier Analysis
a) Statistical methods
b) Distance based detection



universal threshold is ~~used~~ considered while using distance based. But, P may not be detected as an outlier as it is nearer to C2.