Data Mining & Warehousing.

3/2/2023

→ Data, Pattern, Attribute

→ discovery:- database → datamining tools → knowledge

→ Eg of data mining:- FD, M, DW, (A), MB.
↓
not structured

→ How data mining is used. (Point)

Point 2:- Excel to digital.

→ ~~Datami~~ Origin of data mining:-
datamining before machine learning / AI algo. * High dimention.

→ Data Mining tasks:- 1) Classification:- Fish goes to water (Eg) (New entity in then define one class)

2) Regression:- Not define to classification.
Eg:- Bird can fly or not (Yes or no only)

3) Clustering:- Eg:- 1000 data then find K=10 ~~to every~~ (Many) then find distance from first 10. & lastly small distace are take.

4) Dependancies.

5) ...

→ Data Mining Method:- (Not all required (Syll)

→ Why data preprocessing?

→ Data Cleaning.

- To classify we have to know the ~~typ~~ category of the items.

Classification technique or algo:-
Regression, distance, ...

classification Example:- Grading.

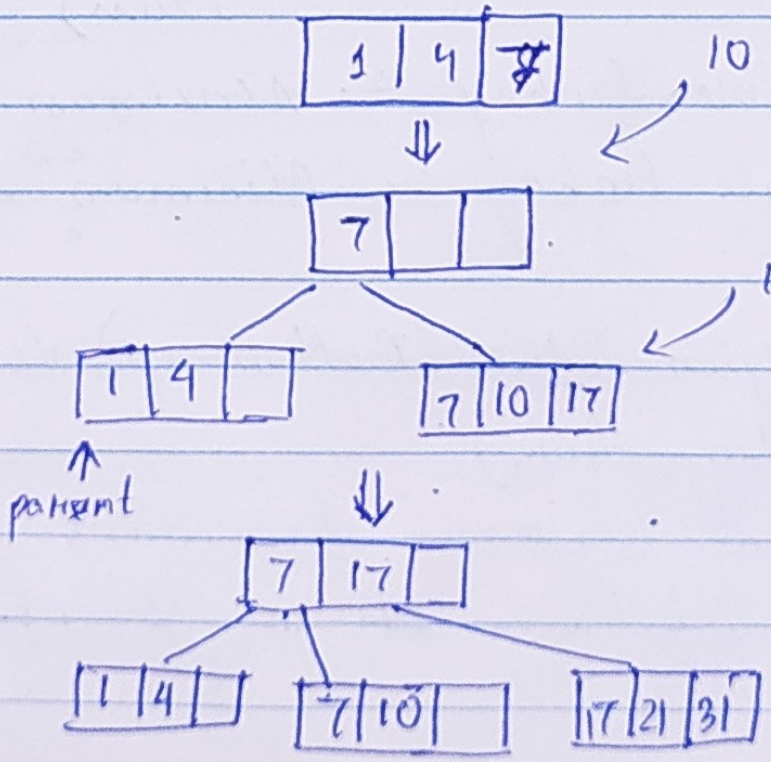
- Partitioning based classification:
  Distance based classification.
  Classification using Regression → Devision, Prediction
- ** KNN:- ~~K[illegible]~~ Algorithm

13/3/2023

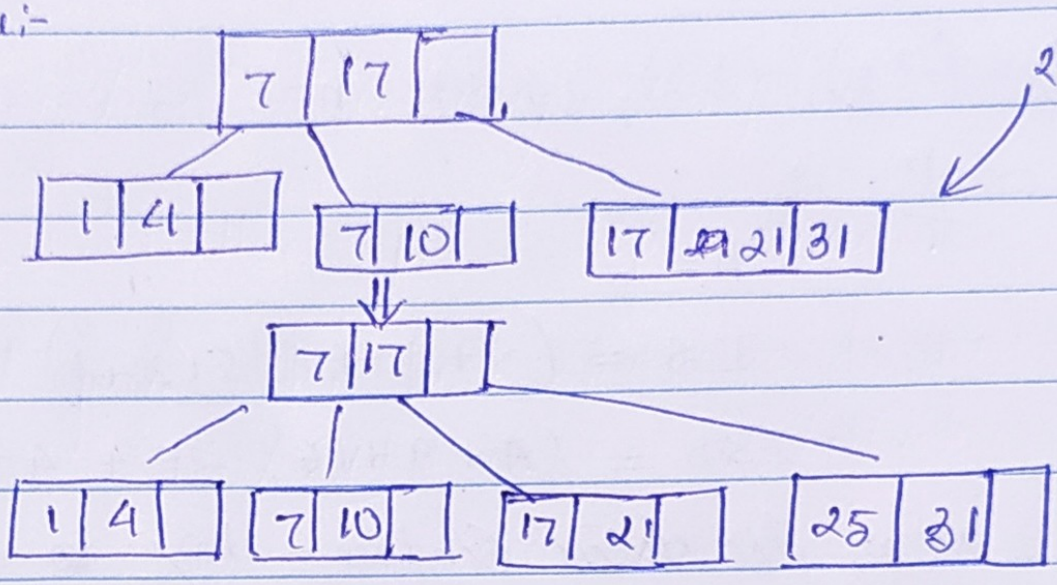
→ BIRCH Algorithm:- (Hierarchical Algorithm)

Eg:- (1, 4, 7, 10, 17, 21, 31, 25, 19, 28, 42)

Solr.



Continue:-



* Phases of BIRCH Algo:- (2-phases)

(a) Phase-1:- (a) Cluster feature tree.

(ii) Scan the database.

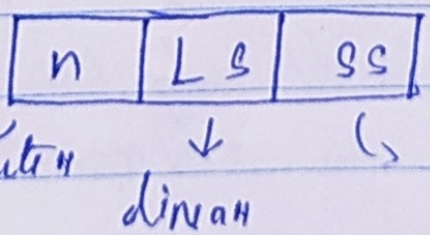
(iii) To build & initialize the cluster.

(b) Leaf node holds many small & tight clusters.

(b) Phase-2:- (a) BIRCH applies a clustering algo to clustered a leaf node of the CF tree.

↳ (cluster feature)

* Every node has three values:-



No. of clusters ↓ ↳ Squared distance

Linear distance

Q) (2,5), (3,2), (4,3) n=?, LS=2, SS=?

Solⁿ.

n = 3

LS = (2+3+4, 5+2+3) = (9, 10)

SS = (4+9+16, 25+4+9) = (29, 38)

Now, the clusters Centre = LS/n = x_0

$$x_0 = \frac{(9,10)}{3} = (3, 3.33)$$

* CF Tree:-

(a) A height balanced tree, it stores clustering features.

(b) ~~Long~~ Non leaf nodes stores sums of the CF's of their children.

(c) CF tree has two parameters branching factor B and threshold ~~for~~ value T.

where, B = maximum no. of ~~long~~ non leaf nodes.

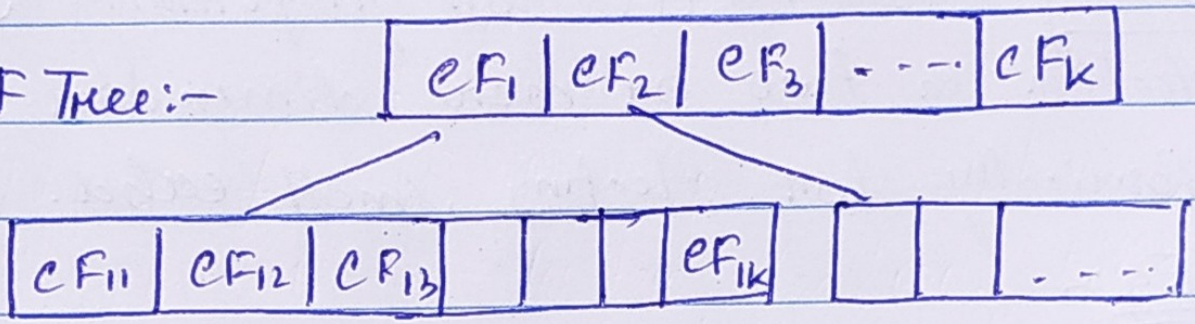
T = max^m diameter of subclusters at leaf node.

(7) R, D Refers to the typist of the cluster around the centroid. It is measured as :-

$$R = \sqrt{\frac{\sum_{i=1}^{n} (x_i - x_0)^2}{n}} = \sqrt{\frac{nSS - 2LS^2 + nLS}{n^2}}$$

$$D = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j)^2}{n(n-1)}} = \sqrt{\frac{2nSS - 2LS^2}{n(n-1)}}$$

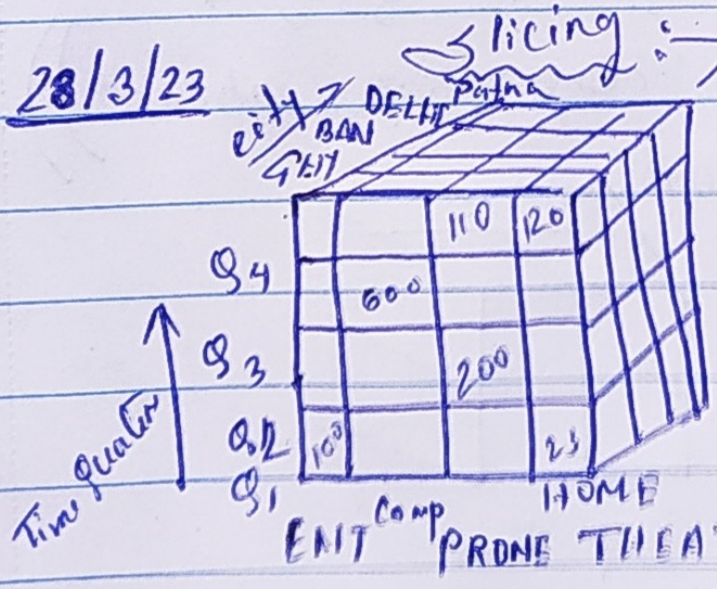
CF Tree:-



We can find R & D from this CF tree.

Data structures used in CF tree :- B+ tree.

28/3/23



Milling (up & down)

Pivoting.

Sale warehouse

Silicon Q_3

data cube

Slice C [quarter, city, item]

time = 'Q2'

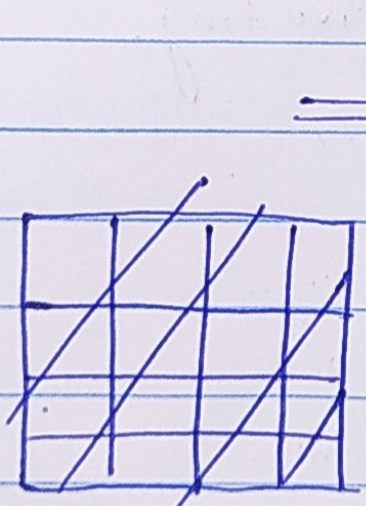
* Slicing reduces dimension by one.

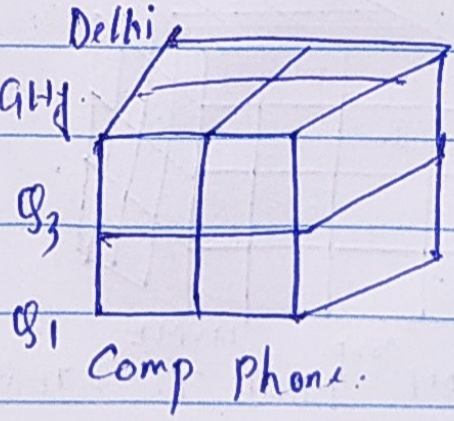
→ Dicing:- It is use for performing a selection on two or more dimensions. Basically for selecting small cubes.

Eg:-

DICE FOR ((City = "GHY" OR "DELHI") &
(Time = "Q1" OR "Q3") &
(ITEM = "comp" OR "phone"))

=





→ Drilling:- It means moving up & down along classification ~~hierarchical~~ levels.

Drilling up:- It performs a detail to an aggregated level.

Drilling down:- It is reverse of drill up, it navigates from less detail to more detail.

* Drill up / down => Roll up / down.

Eg:- (a) Roll-up location = "~~Cities~~ from Cities to ~~Countries~~ States"

C [ITEM, QUARTER, CITIES]

(b) Roll down

From Quarter to month

on

C [ITEM, QUARTER, CITIES]

(c)

Roll-up on location (num of phone items sold in 'Q2' in Delhi & GHY)

(d) Pivot is a rotate is a visualization operation which rotates the data axes to provide an alternative representation of the same data.

SLICE on Time = "Q1" ∈ [ITEM, QUARTER, CITIES]