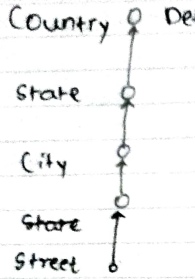


Def^n → Concept Hierachy defines a sequence of mapping from low level concepts to higher level. The mapping forms a concept hierachy for the dimension loc^n mapping a set of low level concepts (Cities) to higher level concepts (Countries)

★ OLAP Operations in multidimensional data model

10-12 marks

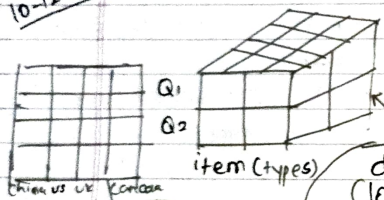
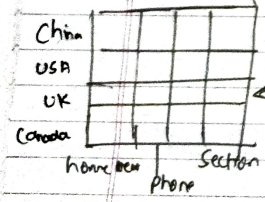


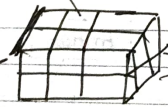
Figure: OLAP Operation on multidimensional data.

dice for (location = Toronto) and time = Q1 or Q2 and item = "home ent"

pivot

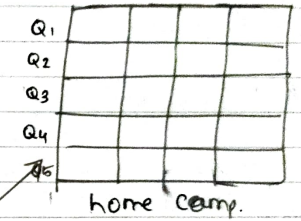


slice for time = Q1



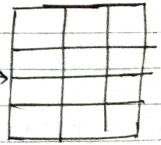
1) Roll up Operation: Climbing up a concept hierachy for a dimension
Eg: Street, City, State country.

2) Drill down Operation: Stepping down a concept hierachy for a dimension.
Eg: Time



roll up on location from cities to countries

drill-down on time from quarters to months



5) Slice and dice: Slice operation performs a selection on one dimension of given cube

4) Dice operation defines a sub cube by performing a selection of operation for 2 or more dimensions

5) Pivot: It provides an alternative presentation of data

★ Datawarehouse Architecture

→ Steps for a design of construction of DW
- What does DW provide for business analyst?
  i) Present relevant information?
  ii) Enhance business productivity
  iii) Customer relationship ~~regent~~ management
  iv) Cost Reduction.

The 4 different views regarding the design of datawarehouse

1) Top-down view: It allows the selection of relevant info necessary for datawarehouse
2) ~~Data~~ Data ~~warehouse~~ source view: The info captured, stored and managed by ~~systems~~ operational systems.
3) Datawarehouse view: Includes fact tables & dimensional table
4) Business query: This is the view of the end users.

★ Datawarehouse Design Process

Steps

1) Choose a business process to model
   Eg: Orders, Shipments, sales
2) Choose grain of the business process - Eg: Individual transaction
3) Choose the dimensions
4) Choose the measures for each fact table record
   Eg: Numerical measures such dollars exchanged

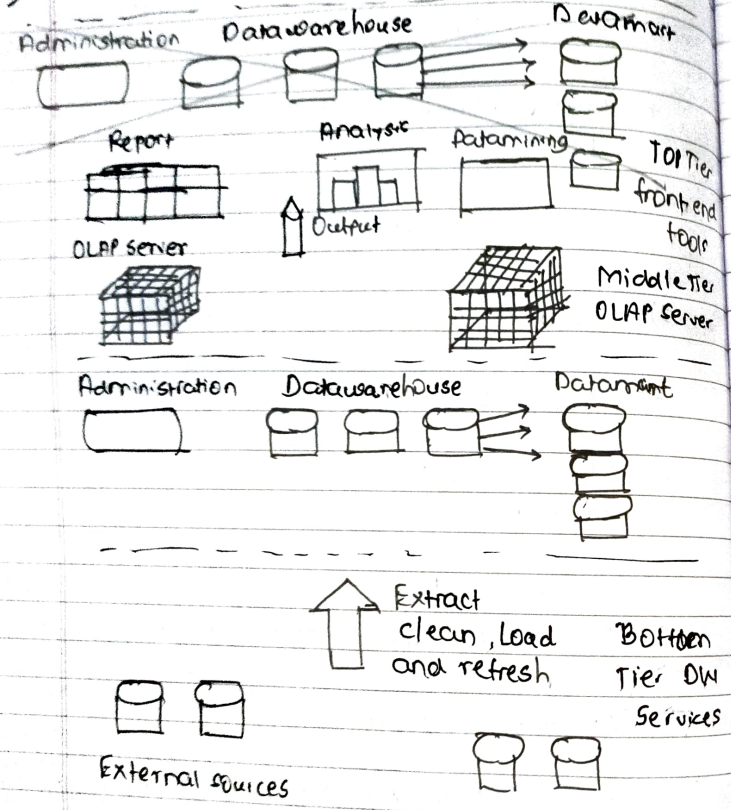
Process of Datawarehouse Design

→ Top down approach → Design and planning
→ Bottom up approach → Experiment and prototypes
→ Combined approach

Software Engineering → planning requirement study, problem analysis, warehouse design
Data integration and testing, deployment.

8-10m

Three-Tier Datawarehouse Architecture



1) Bottom Tier is almost a relational database where the data extracted from this tier in order create a data warehouse.

2) ~~Imple~~ Middle Tier: is implemented using relational OLAP and multidimensional OLAP.

3) Top Tier: Contains reporting tools, analysis tools and DM tools

4) Enterprise Warehouse and Data mart: It collects all of the information about subjects spanning entire organization.

Datamart: It contains information specific organization business unit.

Types of OLAP servers

1) Relational OLAP server
2) Multidimensional OLAP server.
3) Hybrid OLAP server.

1) Relational OLAP server: Applications based on relational DBMS. It can handle large amount of info, greater scalability, performance is low

2) Multidimensional OLAP server: Applications based on multidimensional DBMS, it has fast info retreival, complex calculations is performed and limited info it can handle.

3) Hybrid OLAP Server:

Datawarehouse Implementation

Compute cube Operator and it's implementation

→ Compute cube → Aggregates over all subsets of dimensions.

Eg:

3 dimensional - City, item, year and sales in dollar as measure

$2^3 = 8$ possible groupby - {City, item, year}, {City, item}, {City, year}, {item, year}, {City}, {item}, {year}

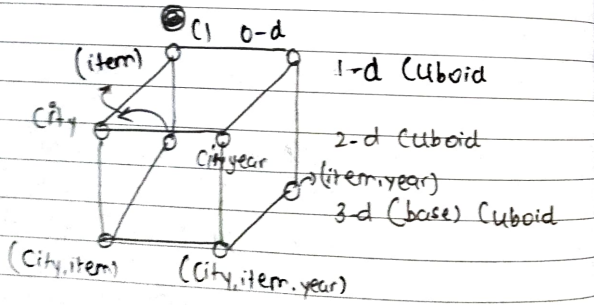


Figure: 3d data cube

★ Datawarehouse Back-end tools & Utilities

1) Data Extraction.
2) Data Cleaning
3) Data Transformation
4) Load
5) Refresh.

21-12-23

Unit-4 Clustering

Viva 8m-10m

★ **Cluster** It is a collection of Objects which is similar to one another and dissimilar to Objects in other clusters. It is widely used in pattern recorgnition, image processing and market research.

Clustering is an example of **Unsupervised learning**.

Unsupervised learning don't depend on predefined classes and class labelled training example.

★ **Requirements of Clustering**

1) Scalability
2) Ability to deal with different types of attributes.
3) Minimal requirements for domain knowledge to determine input parameter.
4) Ability to deal with noisy data
5) High dimensionality.
6) Constraint based clustering
7) Interpretability and Usability.

★ **Types of Data in Cluster Analysis**

1) Internal Scaled Variable → 200-300
2) Binary Variable → 0 or 1
3) Norminal Variable → Red, Blue, Green color
4) Ordinal Variable → Assistant, Associate Professor

Categories of clustering methods

1) Partitioning method

★ → DB of n0 of Objects, K portions of data
→ KEN
→ ~~Classification~~ classify data into k groups
  1) Each group must contains atleast 1 Object
  2) Each Object must belong to each group.

→ Uses iterative relocation technique
→ Methods
  1) K-means
  2) K-medoid

2) Hierachial Method
→ Hierarchial decomposition of given set of data Objects
→ Methods
  1) Agglomerative → Bottom up approach
  2) Divisive → Top down approach.

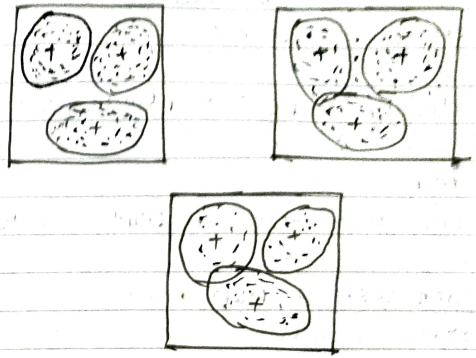
* Density based method: It is based on density, it continues growing given cluster as long as the density exceeds some threshold.

* Greed based method: It quantizes the object space into finite ~~space~~ number of a cells that form a grid structure. All clustering operations performed on grid structure

* Partitioned based method:
  → K≤n

* K-means algorithm (Centroid based Techniques)
  I/p: No of clusters K, n objects
  O/p: Set of K clusters
  Method:
  1) Arbitarily choose K objects as initial cluster centres.
  2) Repeat
  3) (re) assign each of object to cluster to which the object is most similiar based on the value of objects in the cluster.
  4) Update the cluster
  5) Until no change

K=3



* How K-means algorithm works?

First it selects randomly select K of the objects each of which represents cluster mean and cluster centre. For each of the remaining objects asign ~~an object~~ to the cluster to which it is the most similiar based on the distance between the object and the cluster mean. It then computes new mean for each cluster. This process iterates until all the objects are inside a cluster.

Vimp. ★ K-medoid algorithm

I/p: Number of K clusters, n Objects

O/p: Set of K clusters

Method:

1) Arbitrarily choose K objects as initial medoid
2) Repeat
3) Assign each remaining object to Cluster with nearest medoid
4) Randomly select non medoid Objects O_{random}
5) Compute total cost, S of swapping O_j with O_{random}
6) If $S<0$ then swap to form new set of K-medoids
7) Until no change.

K-medoid algorithm.

$K_1 = m_1$, $K_2 = m_2$

	X	Y	K_1	K_2	cost
①	5	2	0	6	~~20~~
②	3	5	8	2	2
③	4	7	9	3	3
④	8	4	2	4	2
⑤	5	5	6	0	0
					7

→ Manhattan Distance

→ $|x_2 - x_1| + |y_2 - y_1|$

→ $|5-8| + |5-2| = 3+3 = 6$

→ $|3-8| + |5-2| = 5+3 = 8$

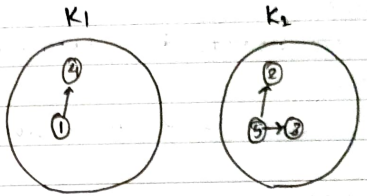
→ $|3-5| + |5-5| = 2$

→ $|4-8| + |7-2| = 4+5 = 9$

→ $|4-5| + |7-5| = 1+2 = 3$

→ $|8-8| + |4-2| = 0+2 = 2$

→ $|8-5| + |1-0| = 3+1 = 4$

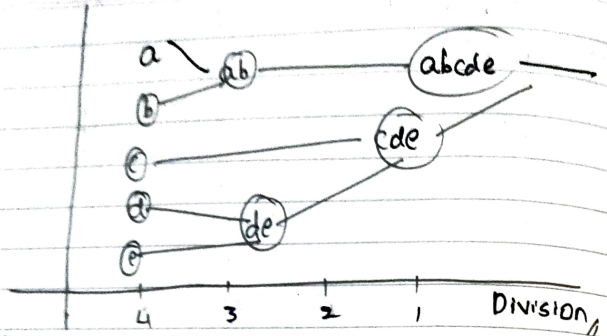


Hierachical Method

→ Works by grouping data Objects into tree of clusters

→ Dividend into → Agglomerative and Divisive Hieranchial ~~agglomeration~~. clusters.

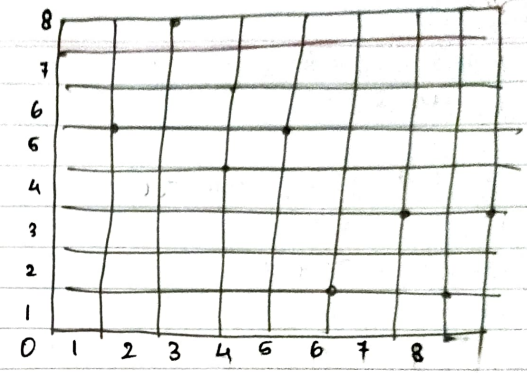
Agglomerative



1) This is top down approach it starts with all objects in one cluster.
2) ~~Starts with all objects~~ It sub divides cluster into smaller and smaller pieces until it each objects form cluster on its own.

* Balanced Iterator reducing and clustering using Hierarchies
→ It's scalable clustering
→ Works for very large data sets
→ Only one scan of data is necessary Featuring
→ Clustering is based on CF (Cluster ~~Forwarding~~
→ CF Tree → Stores the Cluster ~~forwarding~~ featuring
Cluster of data pts is represented by triple of numbers (N, LS, SS)
N = No of Iterations
LS = Linear sum of points
SS = Sum of squared of pts

Eg:



Eg: CF

CF = (N, LS, SS)

N = No of data pts

$LS = \sum_{i=1}^{N} X_i$

$SS = \sum_{i=1}^{N} X_i^2$

(2,6) , (3,8) , (4,4) (4,6) (5,6) (6,0) (8,4)

$LS = \sum_{1}^{N} X_i = 2+3+4+4+5+6 = 29$

$SS = \sum_{i=1}^{N} X_i^2 = 2^2 + 3^2 + 4^2 + 4^2 + 5^2 + 6^2 + 8^2$

$= 4 + 9 + 16 + 16 + 25 + 36 + 64$

$= 170.$

$LS = \sum_{i}^{N} Y_i = 6+8+4+6+6+0+4$

$= 34$

$SS = \sum_{i}^{N} Y_i^2 = 6^2 + 8^2 + 4^2 + 6^2 + 6^2 + 0^2 + 4^2$

$= 36 + 64 + 16 + 36 + 36 + 16$

$= 204.$

Basic Algorithm of BIRCH

Phase: ①

(a) Load the data into memory

Phase ②:

Condense Data - (B) Resize the data set by building smaller Cluster feature tree

Phase ③:

Global Clustering. It uses ~~vert~~ existing clustering algorithms on CF entries

Phase ④:

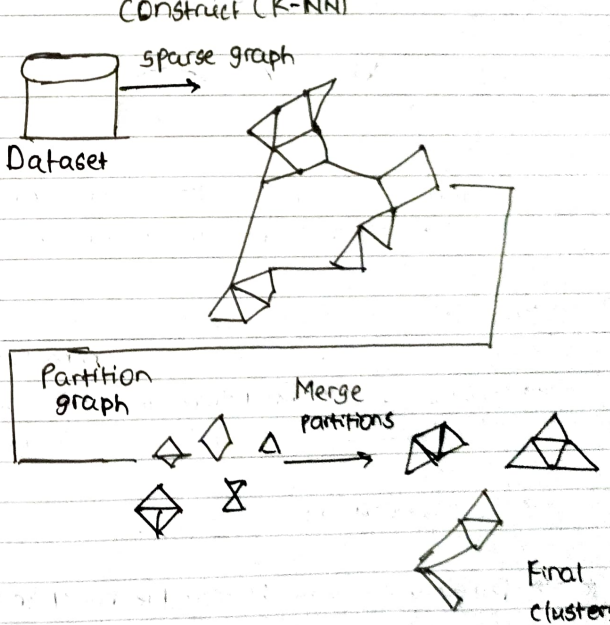
Cluster Refining.

★ Chameleon: Hierarchical Clustering using dynamic modelling.

→ Two Clusters merged → if interconnecting or Closeness between 2 Clusters are high

→ 2 parameters → Interconnectivity and Closeness.

→ Graph based and a two phase algorithm.
- Uses a graph partitioning algorithm
- Uses agglomerative hierarchial Cluster features.



★ DBSCAN

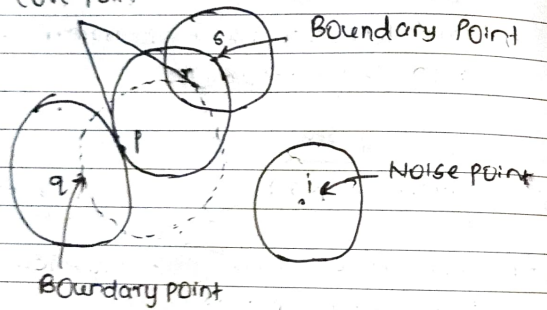
Time 1 - 3 marks

(Density based Spatial Clustering Of Application with noise)

→ Forms Clustering based on density

→ Eps, Minpts = 3 → 2 i/p parameters

Core point



→ p and r core points → satisfies Minpts = 3

→ s and Q → Boundary points → should be ~~neighbour~~ neighbour of any corepoint.

→ t is noise pt

★ ➡ Directly density Reacheable must be neighbour and core point so q is directly density reacheable from pt p.

5/1/24

Unit-3 Classification and Prediction

Classification: Data classification is 2 step process

1) A model is built ~~determining~~ describing a pre determined data classes or concepts

2) Test data are used to estimate accuracy of the classification rules.

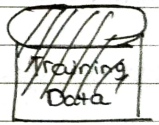
Classification

→ Training dataset

→ Testing dataset

→ Supervised Learning ⟶ labelled data

→ Unsupervised Learning ⟶ unlabelled data



★

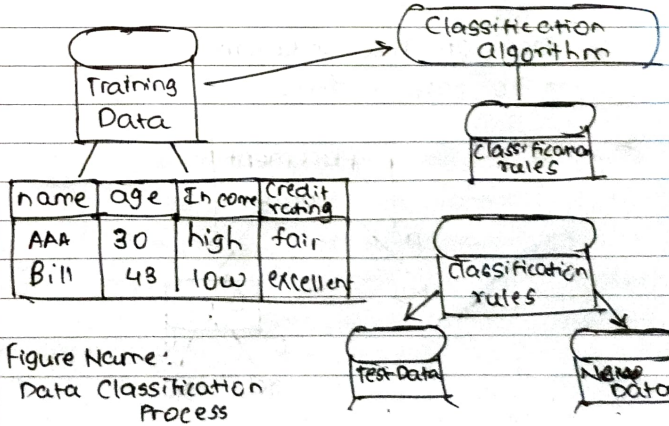
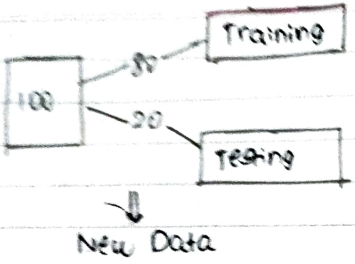


Figure Name:
Data Classification Process



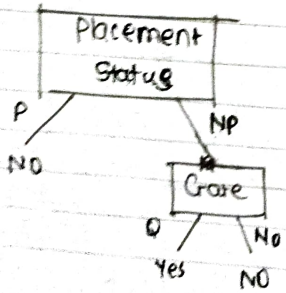
Issues regarding classification and prediction

1) Data Cleaning: Pre processing of data to reduce noise

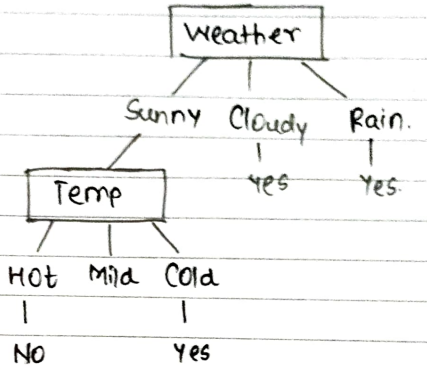
2) Relevance Analysis: Reducing irrelavant attributes

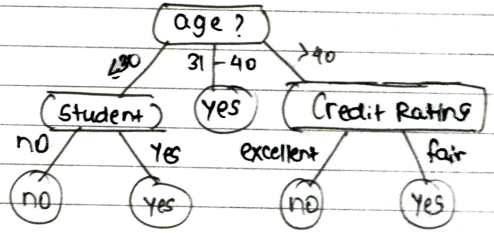
3) Data transformation: Data can be generalized to higher level concept

4) Decision Tree Induction
→ To take a decision.



Playing Football





★ Decision Tree is a flowchart like tree structure where each internal node denotes a test on an attribute, each branch ~~denotes~~ represents the outcome of test and leaf node represents class or class distribution. Topmost node of a tree is the root node.