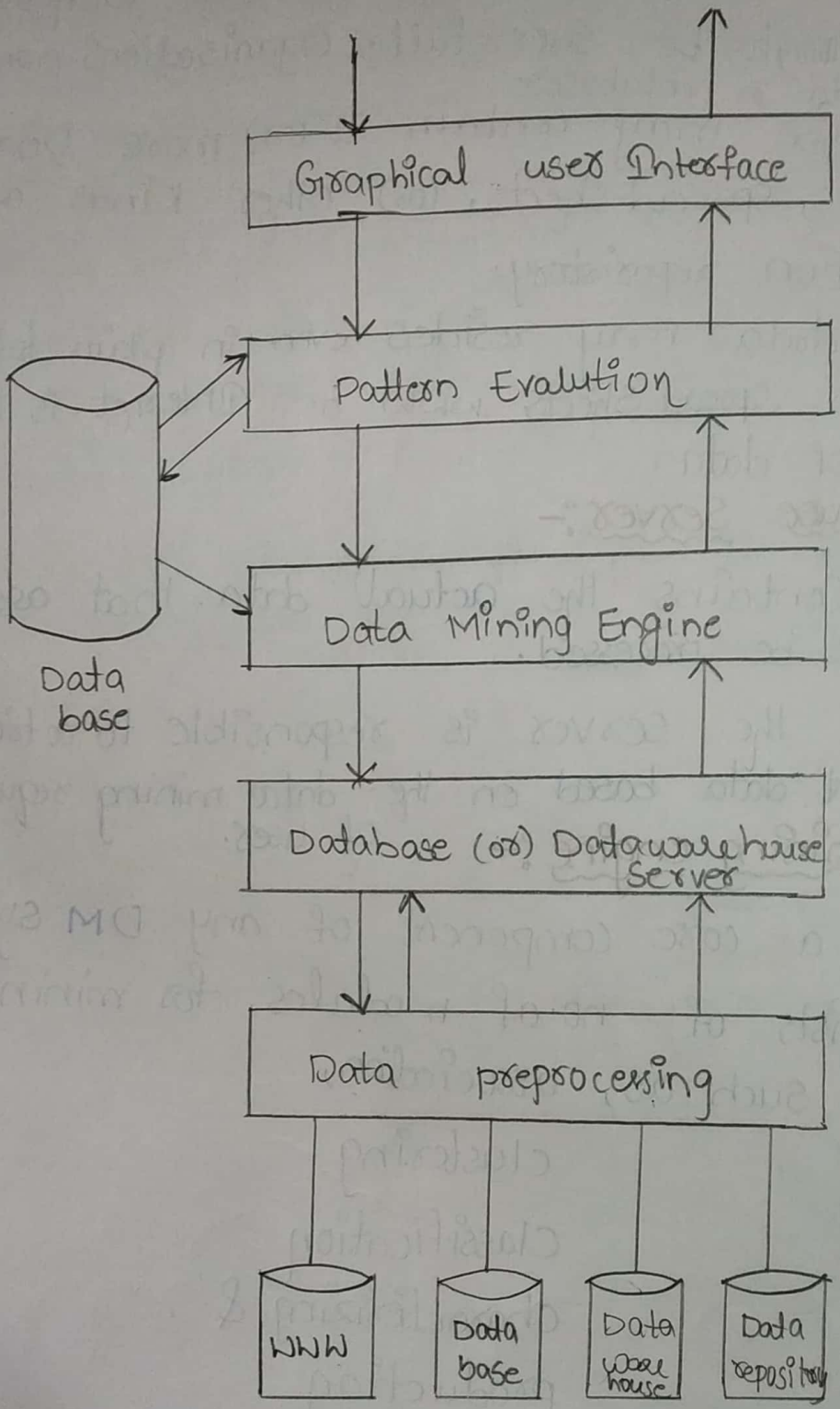
* Data Mining Architecture:-



Data Sources :-

Database, DW house & WWW & text files & other documents are actual sources

of data. You need
* Large volumes of historical data is needed for DMining to be sucessfull. organisations manually store data in databases.
* DW house may contain 1 (or) more DBases text files, spread sheets, (or) other kinds of information repository.
* some data may resides even in plain text files or spread sheets www or Internet is big source of data.
* DW house Server:-
* It contains the actual data. that are ready to be processed.
* Hence the server is responsible to retrieving the relevant data based on the data mining request of user.
* Data mining engine:-
* It is a core component of any DM system it consists of no. of modules for mining tasks. such as, association,
clustering
classification
characterizing &
production
* ~~Data mining Engine:-~~ Pattern Evaluation Modules:-
* It is mainly responsible for the measure of interestingness of the patterns by using the a

threshold value.

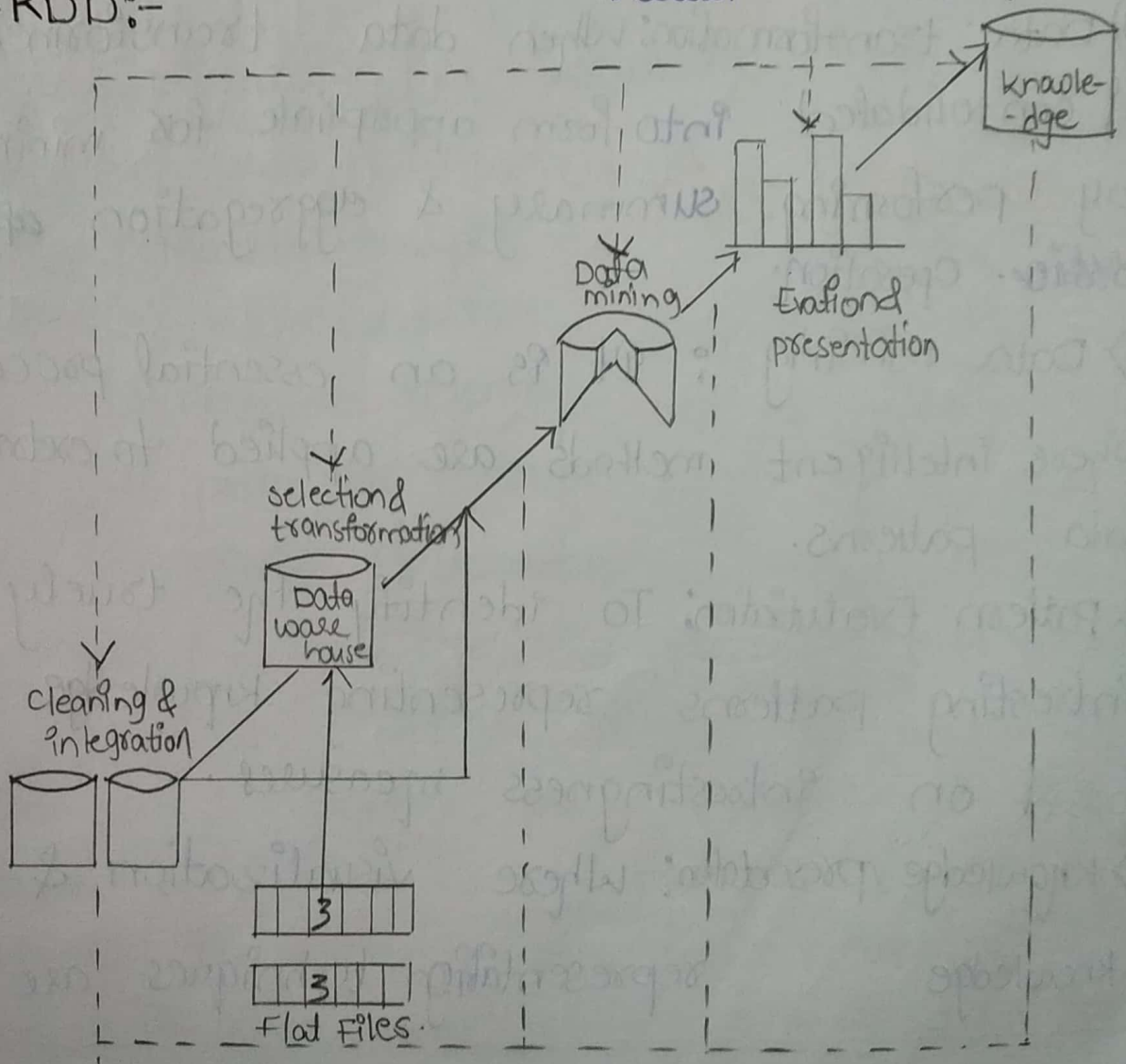
*It interacts with the DM Engine to focus the reach towards data mining engine to focus the reach towards intresting patterns.

*GUI :-

* GUI module communicates b/w the users & DM system. this module helps the users to use the system efficently without knowing the real complexity behind the Concepts. ↓ users.

*When the users specifys the query or a task, and display the result in an easily understandable manner.

*KDD:-



* Knowledge Discovery in Data
* KD process is shown in the fig as interactive steps.
* The steps involved in KDD as

1) Data cleaning: To remove noise and inconsistent data.
2) Data Integration: Where multiple sources may be combined
3) Data selection: When data relevant to the analysis ~~tasks~~ are retrieved from the DB
4) Data transformation: When data transformed & consolidated into form appropriate for mining by performing summary & aggregation ~~of~~ operation.
5) Data Mining: It is an essential process where intelligent methods are applied to extract data patterns.
6) Pattern Evaluation: To identify the truely intresting patterns representing knowledge based on intrestingness measures.
7) Knowledge presentation: Where visualization & knowledge representation techniques are

used to present mined knowledge to users.

* The preceding view shows data mining as one step in the knowledge discovery process,

* However, in Industry the term DM is used to refer KD process. ∴ We adopt a broad view of data mining functionality.

*What is data Mining?

Data mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted. So can be used for any of the applications.

1) Market Analysis.
2) fraud detection.
3) Customer Retention.
4) production control.
5) Science Exploration.

* What is the need for preprocessing of data?

In Complete noisy & inconsistent data are common place properties of large scale real world databases & data warehouses. In Complete data can occur for a no. of reasons. Attributes of interest may not

always be available, such as customer information for sales transition data. other data may not be included simply because it was not considered imp at the time of entry relevant data may not be recorded due to misunderstanding. Data that were inconsistent with other recorded data may have been overlooked Missing data, particularly for tuples with missing values for some attributes may need to be inferred.

* Data mining

DM is a process of extracting knldge from a volume of Data. It refers to away of finding significant & useful informatn from an organisation. DM is a tool to convert data into informatn.

* There are 3 reasons
  1) knldge discovery
  2) data visualization
  3) data correction.

* 1) It helps us to identify the invisible correlation ptns and trends all the data available in the DW house.

2) Data visualization:-

The main objective of this is to normalized large volumes of data to find the sensiable way to display the data.

3) Data correction:-

The main objective of the DC is to identify & correct the incomplete & inconsistent data by removing errors.

* DATA MINING TASKS:-

These are divided into 2 categories.

1) predictive tasks.
2) Descriptive tasks.

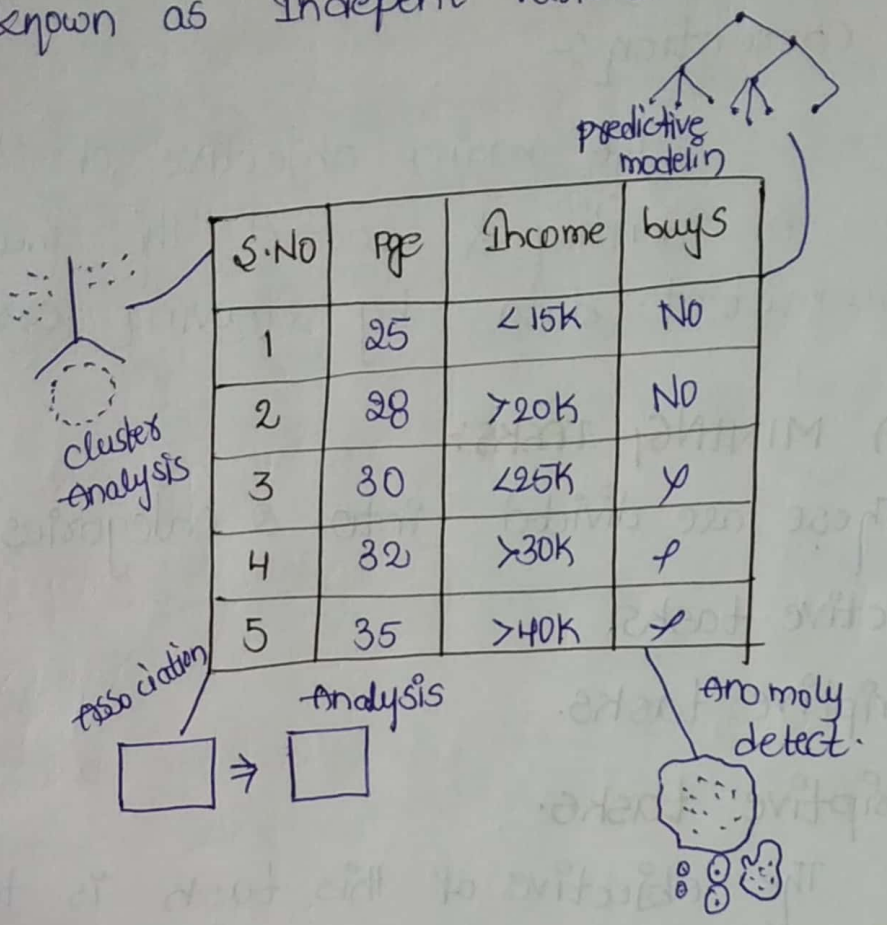
1) Descriptive tasks.

The objective of this task is to divide the pattern and [co-relation and clusters anamolies] thus summerizes the underlying relation ship in data. It req Post processing techniques to validate the results.

2) predictive Task:

The objective of this task is to predict the value of the a particular attribute based on the value of the other attributes.

• Attribute i.e., to be predicted is known as dependent (or) target variable and attributes used for making the prediction is known as Indepent variable.

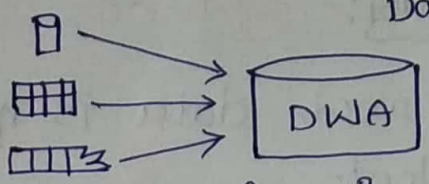


* Data Integration :-

When the data from the different multiple sources are combined from a ⊥ Coherent store.

Data Transforming :-

Where the data is converted from ⊥ form to the other data is normalized, aggregated and Generalized.



Data Transformation.

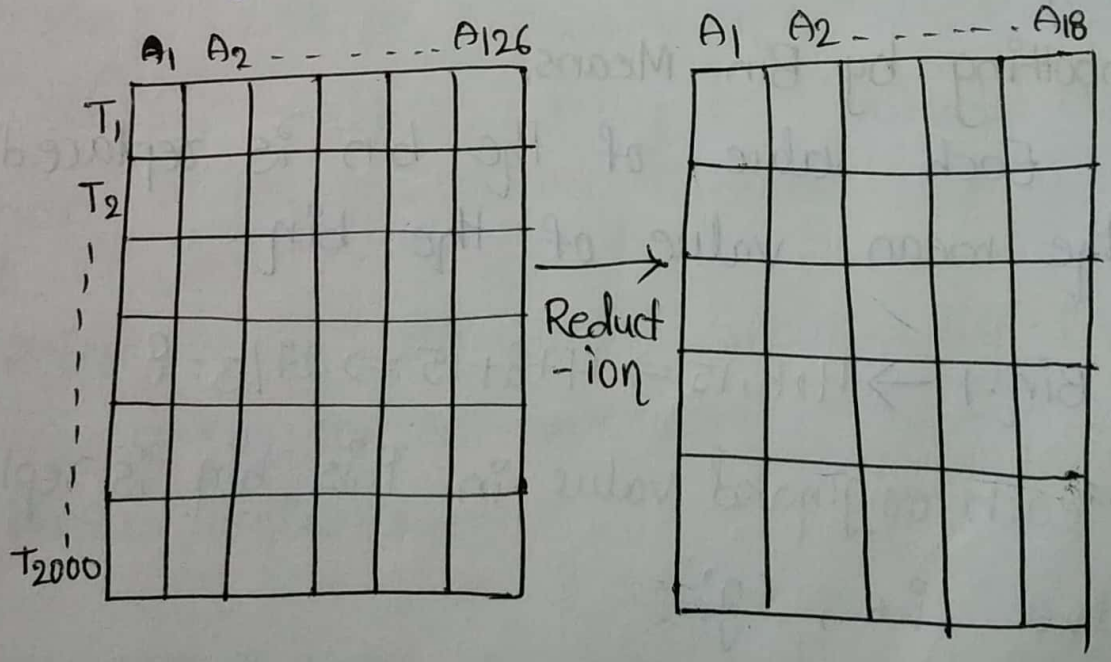
* Data Reduction:-

Which shows the reduced representation of the data in the Dwtl i.e., smaller in volume but produces some analytical results.

Data Transformation
-2, 32, 100, 59, 48 → 0.02,
0.32, 1.0, 0.59, 0.48

→ Data preprocessing techniques can improve the quality of data this by improve the assay's efficiency of mining process.

→ Data preprocessing is a important step in the KDD process, since quality divisions are based on quality on data.



DATA REDUCTION ATTRIBUTES.

* DATA CLEANING:

1) In the real word, data is noisy / incomplete and inconsistent.

2) Data cleaning is used to fill the missing values, smooth the Noisy data and convert the in consistencies.

1) MISSING VALUES:-

The following are the methods to fill the missing values of an attributes.

(a) Ignore the Tuple:-

→ When the class label is missing

→ This method is not effective, if tuple contains and several attributes with missing values.

(b) Smoothing by Bin Means:

Each value of the bin is replaced by the mean value of the bin.

Eg:- Bin. 1 → 4, 8, 15 ⇒ 4+8+15 ⇒ 27/3 = 8

i.e., each originated value in this bin is replaced by value i.e., '9'.

Smoothing by Bin means.

Bin 1 : 9, 9, 9

Bin 2 : 21, 21, 24 $\Rightarrow \frac{21+21+24}{3} = 66/3 = 22$

i.e., Bin 2 : 22, 22, 22

Bin 3 : 25, 28, 34 $\Rightarrow \frac{25+28+34}{3} = 87/3 = 29$

i.e., Bin 3 : 29, 29, 29

(C) Smoothing by Bin boundaries :-

→ In this min & max values of the bin are identified as Bin boundaries.

→ Each bin value is replaced by the closest boundary value.

(i) Bin 1 : 4 8 15 closest boundary values to 8 in bin

(min) (max)

1 is 4

Bin 1 : 4, 4, 15

(ii) Bin 2 : 21, 21, 24 → Closest value to 21 is same i.e., 21 Bin 2 : 21, 21

(iii) Bin 3 : 25, 28, 34 → closest value is 25, Bin 3 : 25, 25, 25

Clustering :- outliers may be detected by clustering, where similar values as organized into a group as clusters.

→ Regression :- Data can be smoothed by fitters the data to a function.

(a) Linear Regression :-

It involves 2 variables.

Finding the line, so that variable can be used to predict the other $Y = \alpha + \beta X$

* Data Reduction Techniques:-

These Techniques can be applied to obta-
-in a reduced representation of the dataset that is much smaller in volume, yet closely maintains the integrity of the original data. Data reduction includes,

* Data cube aggregation
* Dimension reduction
* Data Compression
* Discretization
* Numerosity Reduction
  - → Regression
  - → Histograms
  - → Clustering
  - → Sampling

→ Data cube aggregation where aggregation oper-
-ations are applied to the data in the construction of datacube.

→ Attribute subset selection, where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.

→ Dimensionality Reduction, where encoding mechanisms are used to reduce the data set size.

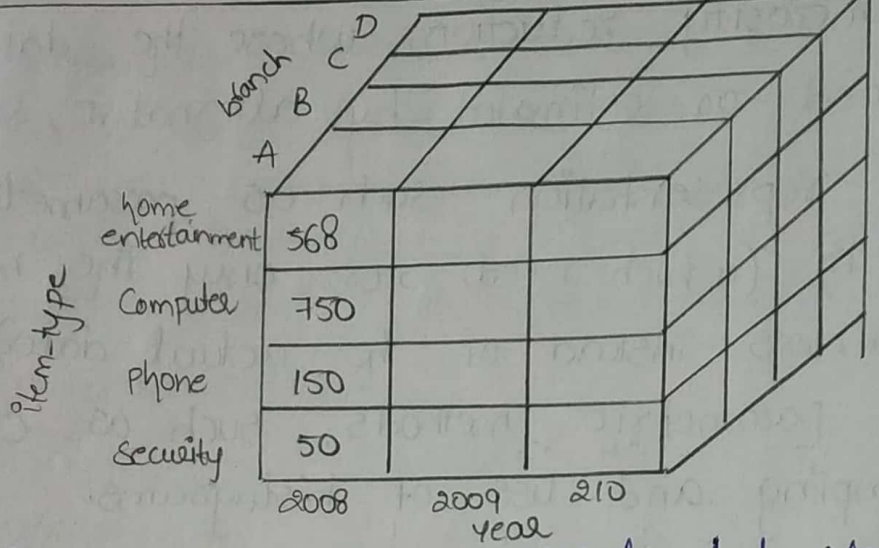
Examples :- Wavelet Transforms principal Components Analysis.

→ Numerosity reduction, where the data are replaced or estimated by alternative, smaller data representation such as parametric models (which need, store only the model Parameters instead of the actual data) or non parametric methods such as clustering, Sampling and use of histograms.

→ Discretization and concept hirearchy generation where raw data values for attributes are replaced by ranges or higher Conceptual levels. Data Discretization in the form of numerosity reduction that is very useful for the automatic generation of concept hirearchies.

* Data cube Aggregation:-

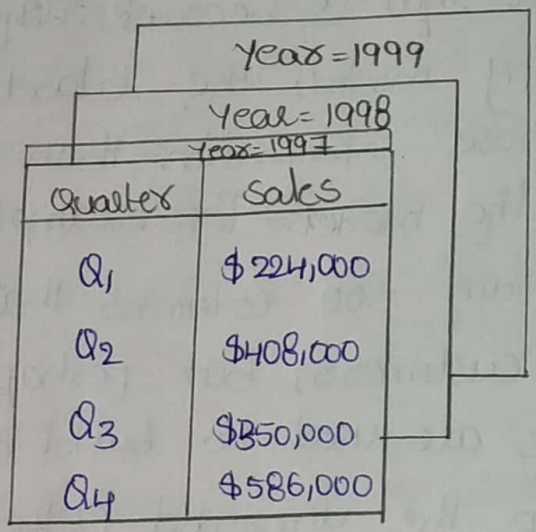
Reduce the data to the Concept level needed in the analysis. Queries regarding aggregated information. The following figure shows a data cube for multidimensional analysis of stately data with respect to annual sales per item type for each branch.

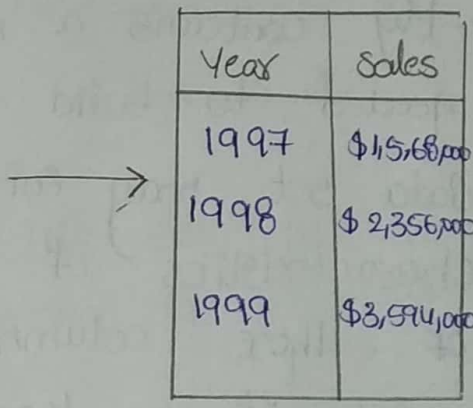


→ Each cells holds an aggregate data values corresponding to the data point in multidimensional space. Data cubes provide font access to pre computed, summarized data, thereby benefiting on-line analytical processing as well as data mining.

→ The cube created at the lowest level of abstraction is reffered to as the base cuboid. A cube for the highest level of abstraction is the apex cuboid. The lowest level of a data cube (base cuboid). Data cubes created for varying levels of abstraction are sometimes reffered to as cuboids, so that a "data cube" may instead after to a lattice of cuboids each higher levels of abstraction further reduces the resulting data size.

→ The following database consist of sales per quarter for the years 1997 - 1999





Year	Sales
1997	$1,568,000
1998	$2,356,000
1999	$3,594,000

Suppose, the analyzer interested in the annual sales rather than sales per quarter, the above data can be aggregated so that the resulting data summarizing the total sales per year instead of per quarter. The resulting data in smaller in volume without loss of information necessary for analysis task.

* Dimensionality Reduction :-

It reduces the data set size by removing irrelevant attributes. This is a method of attribute subset selection are applied. A heuristic method of attribute of subset selection is.

explained here:

Attribute sub selection/feature selection:

Feature selection is a must for any data mining product. That is because, when you build a data mining model, the dataset frequently contains a more information than is needed to build the model. For example, a data set may contain 500 columns that describe characteristics of customers, but perhaps 50 of those columns are used to build a particular model. if you keep the unneeded columns while building the model more cpu and memory are required during the training process, and more storage space is required for the completed model.

→ In which select a minimum set of features such that the probability of distribution of different classes gain the values for those features is as close as possible to the original distribution given the values of all features.

→ Basic heuristic methods of attribute subset selection include the following techniques

some of which are illustrated below.

01) step-wise foward selection :-

The procedure starts with an empty set of attributes. The best of the original attributes is determined and added to the set. At each subsequent iteration or step, the best of the remaining original attributes is added to the set.

2) step-wise backward elimination:

The procedure starts with full set of attributes. At each step, it removes the worst attribute remaining in the set.

3) Combination forward selection and backward elimination:

The step wise forward selection and backward elemination methods can be combined, where at each step select the best attribute and removes the worst from among the remaining attributes.

4) Decision tree induction:

It constructs a flow chart like structure where each internal (non leaf) node denotes a test on an attribute, each branch corresponds to overcome of the test, and each external (leaf) node denotes a class prediction

At each node, the algorithm chooses the "best" attribute to partition the data into individual classes. when decision tree induction is used for attribute subset selection, a tree is constructed from the given data. All appearing in the tree form the reduced subset of attributes.

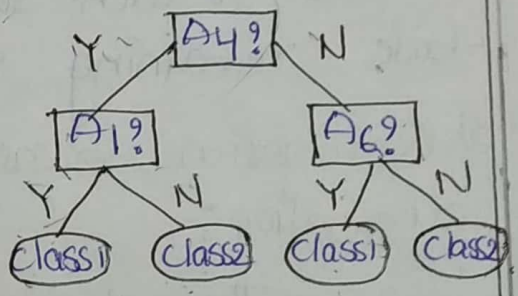
forward selection	Backward elimination	Decision tree Induction
Initial attribute set:	Initial attribute set:	Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$	$\{A_1, A_2, A_3, A_4, A_5, A_6\}$	$\{A_1, A_2, A_3, A_4, A_5, A_6\}$
Initial reduced set:	$\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$	
$\{\}$	$\Rightarrow \{A_1, A_4, A_5, A_6\}$	
$\Rightarrow \{A_1\}$	$\Rightarrow$ Reduced attribute set:	
$\Rightarrow \{A_1, A_4\}$	$\{A_1, A_4, A_6\}$	
$\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$		$\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$

Fig:- Greedy methods for attribute subset selections.

* Wrapper approach / filter approach:-

The mining algorithm itself is used to determine subset then it is called wrapper approach or filter approach. wrapper approach leads to greater accuracy since it optimizes the evaluation measure of algorithm while removing attributes.

* Data Compression:-

In Data Compression, data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of original data if the original data can be reconstructed from the compressed data without any loss of information, the data Compression technique used is called lossless. if instead, we can reconstructed only an approximation of the original data, then the data Compression technique is called lossey. effective methods of lossy data Compression.

* wavelet transforms
* principal Component analysis.

* Wavelet transforms:- (Discrete wavelet transformation).

It is a form of data Compression well suited for image Compression. The discrete wavelet transform DWT is a linear signal processing technique that, when applied to a data vector D, transforms it to a numerically different vector, D0, of wavelet coefficients.

→ The general algorithm for a discrete wavelet transform is as follows.

* The length L of the input data vector must be an integer power of two. This condition can be met by padding the data vector with zeros as necessary.
* Each transform involves applying two functions:
  - data smoothing
  - Calculating weighted difference.
* The two functions are applied to pairs of the i/p data, resulting in two sets of data of length L/2.
* The two functions are recursively applied to the sets of data obtained in the previous loop, until the resulting data sets obtained are of desired length.
* A selection of values from the data set obtained in the above iterations are designated the wavelet coefficients of transformed data.

→ If wavelength coefficient are larger that some as a specified threshold then it can be retained. The remaining are set too.

Haar2 and Daubechie are two popular wavelet transforms.

* principal Component Analysis (PCA):-

It is also called as Karhunen-Loeve (K-L) method.

Procedure:-

→ Given N data vectors from k dimensions, find CL=K orthogonal vectors that can be best used to represent data.

→ The original data set is reduced (projected) to one consisting of N data vectors on c principal Components (reduced dimensions)

→ Each data vector is a linear Combination of the c principal Component vectors.

→ Works for ordered and unordered attribute

→ Used when the no. of dimensions is large.

The principal Components new set of axees give important about variance. Using the strongest Components one can reconstruct a good approximation of the original signal.

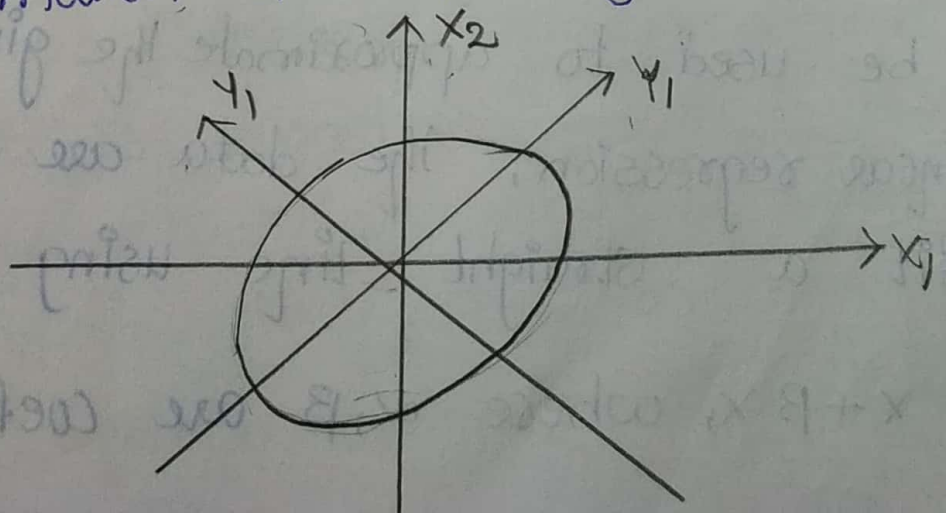


fig: principal Component analysis. Y_1 & Y_2 are the first two principal components for the given data.

*Numerosity Reduction:

Data volume can be reduced by choosing alternative smaller forms of data. This tech. can be

*parametric method

*Non parametric method.

*parametric method:-

Assume the data fites some model, then estimate model parameters and store only the parameters, instead of actual data.

Non parametric Method: In which histogram, clustering and sampling is used to store reduced from data.

Numerosity reduction techniques:

01) Regression and log linear model:

→ Can be used to approximate the given data.

→ In linear regression, the data are modeled to fit a straight line using

$Y = x + \beta X$, where α, β are coefficients

Multiple regression:

$$Y = b_0 + b_1 x_1 + b_2 x_2$$

Many non linear functions can be trans--formed into the above.

log linear model:

The multi way table of joint prob--ability is approximated by a product of lower order table.

probability: $P(a, b, c, d) = \alpha ab \beta ac \chi ad \delta bcd$

02) Histogram :-

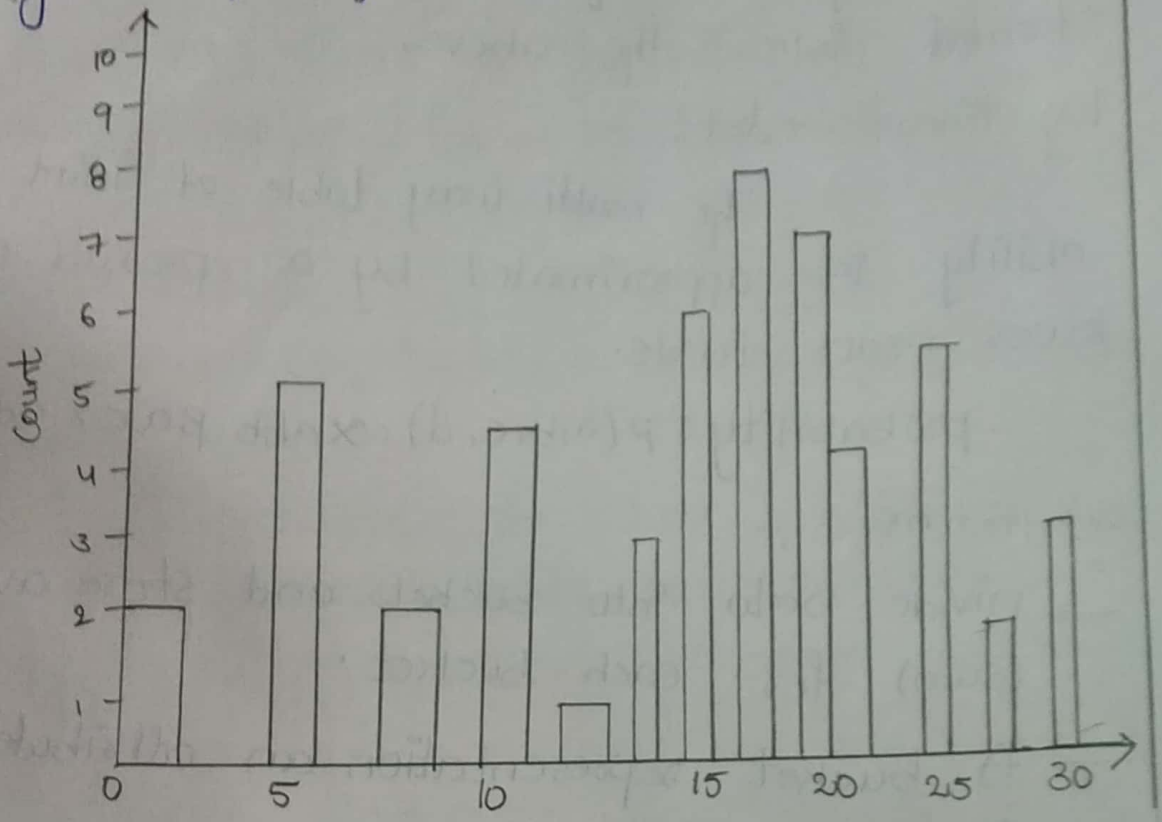
→ Divide data into buckets and store average (sum) for each bucket.

→ A bucket representation an attribute value/ frequency pair.

→ It can be constructed optimally in one dimension using dynamic programming.

→ It divides up the range of possible values in a data set into class or groups. for each group, a rectangle buck--et is constructed with a base length equal to the range of values in that specific group, and an area proportional to the no. of observation falling into that group.

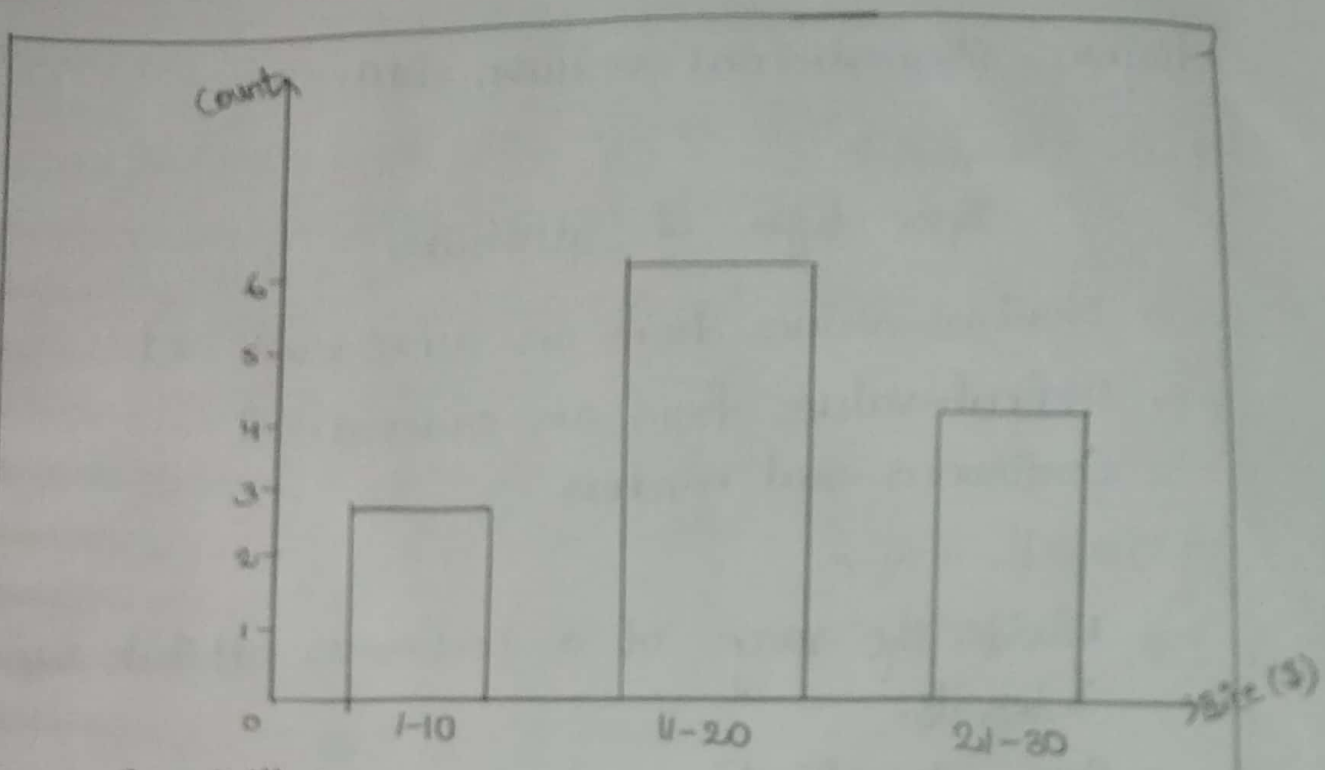
→ The buckets are displayed in a horizontal axis while height of a bucket represents the average frequency of the values.



Example :-

The following data are a list of prices of commodity of sold items. The number have been sorted. 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

The buckets can be determined based on the following partitioning rules, including the following.



1) Equi-width :-
Histogram with bars having the same width.

2) Equi-depth :-
Histogram with bars having the same height.

3) V-optional :-
Histogram with least variance □(count values).

4) Max diff : Bucket boundaries defined by user specified threshold.

V-optional and Max Diff histogram tend to be the most accurate and practical. Histograms are highly effective at approximating both pause and dense data, as well as

highly skewed, and uniform data.

*Discretization :-

Three types of attributes

→ Nominal - values from an unordered set

→ ordinal - values from an ordered set

→ Continuous - real numbers.

*Discretization :-

→ Divide the range of a continuous attribute into intervals.

→ Some classification algorithms only accept categorical attributes.

→ Reduce data size by discretization

→ prepare for further analysis.

Discretization and Concept hierarchy.

Discretization

→ Reduce the number of values for a given continuous attribute by dividing the range of attribute into intervals. Interval labels can then be used to Replace actual data values.

→ Concept hierarchies.

→ Reduce the data by collecting and replacing

low level Concepts (such as nemeric values for the attribute age). by higher level Concepts (such as young, middle age or senior).

Discretization and Concept hierarchy generation for numeric data.

→ Binning (see sections before)

→ Histogram analysis (see sections before)

→ clustering analysis (See sections before.)

→ Entropy based discretization

→ Segmentation by natural partitioning

Entropy based Discretization :-

→ Given a set of samples S, if S is partition-ed into two intervals S_1 & S_3 using boundry T, the entropy after partitioning is

$$E(S,T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

→ The boundry that minimizes the entropy function overall possible boundaries is selected as binary discretization.

→ The process is recursively applied to partitions obtained until some stopping criterion

is met, e.g.

$$Ent(S) - E(T,S) > \delta$$

→ Experiments show that it may reduce data size and improve classification accuracy.

* Linear Regression:

Data are modeled to fit a straight line

→ Often uses the least square method to fit the line

$$Y = \alpha + \beta X$$

→ Two parameters, α and β specify the line and all to be estimated by using the data at hand.

→ Using the least squares criteria to the known values of $Y_1, Y_2 ----, X_1, X_2 -----$

* Multiple Regression:-

It allows a response variable Y to be modeled on a linear function of multi dim-ensional feature vectors.

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

→ Many non linear functions can be transformed into the above.

* log - linear model :-

Approximates discretes multi dimen-sional probability distribution.

→ The multi way tables of joint probabilities is approximated by a product of lower order tables

Probability : $p(a,b,c,d) = \alpha ab \beta ac$ ad & bcd

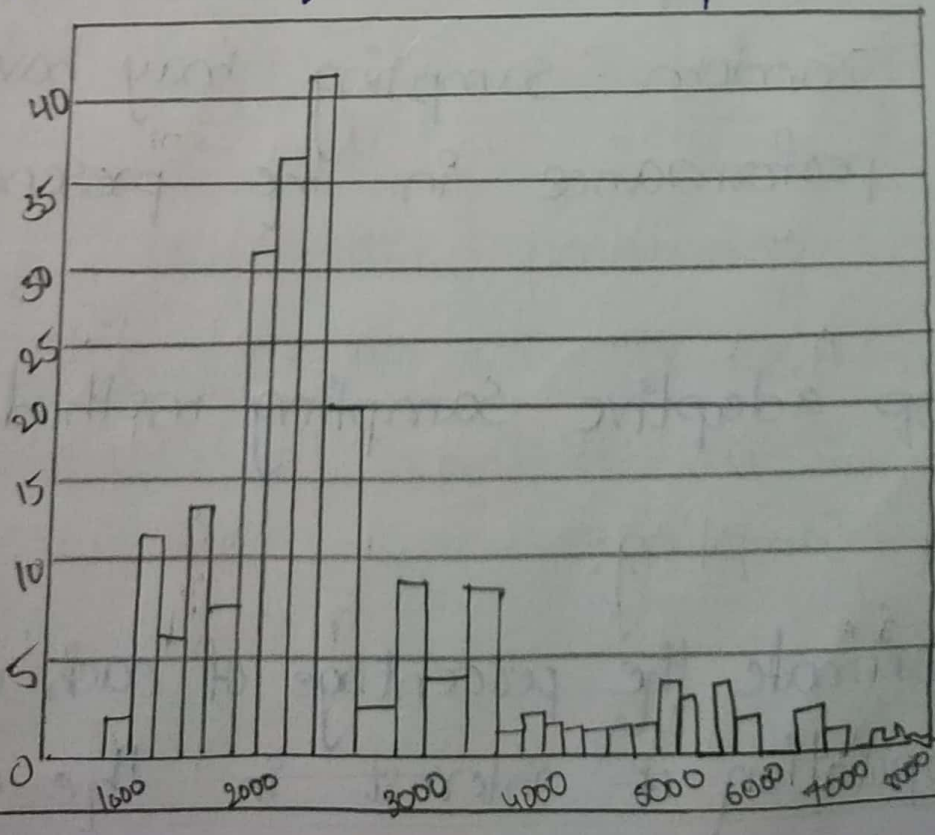
* Histograms :

→ A popular data reduction technique

→ Divide data into buckets and store average sum for each bucket.

→ Can be constructed optimally in one dimen-sion using dynamic programming.

→ Related to quantization problem.



* clustering :-

→ partition data set into clusters, and one can store cluster representation only.

→ can be very effective if data is clustered but not if data is "smeared".

→ Can have hierarical clustering and be stored in multi dimensional index tree structure.

→ There are many choices of clustering definitions and clustering algorithms, further detailed in chapter 8.

* Sampling :-

→ Allow a mining algorithm to run in complexity that is partionally sub line to the size of data.

→ Choose a representation subset of the data

→ Simple random sampling may have very poor performance in the presence of skew.

Develop adaptive sampling methods.

Stratified Sampling:

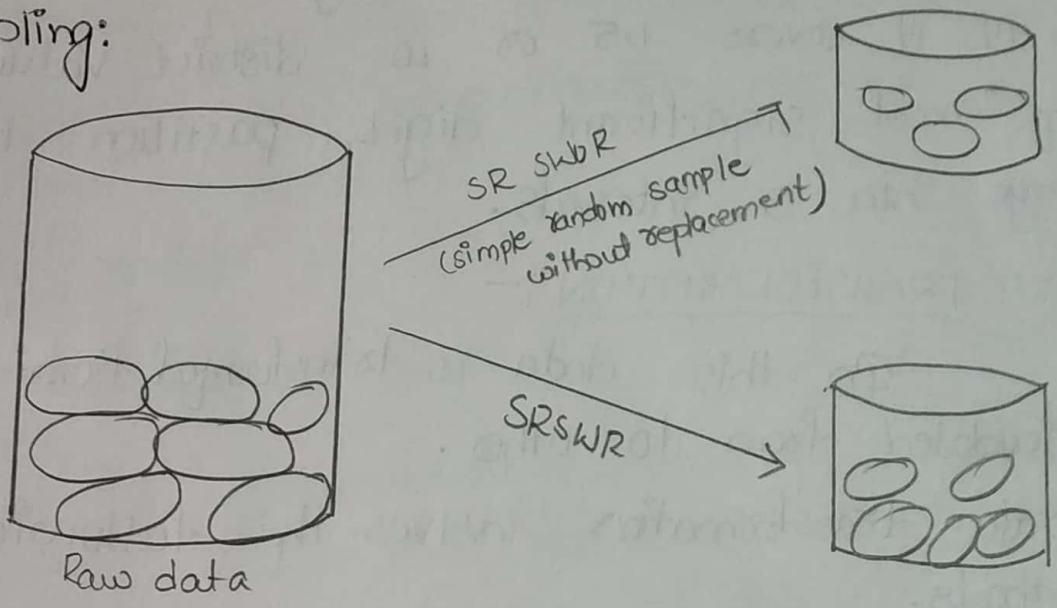
→ Approximate the percentage of each class/ sub population of interest in the overall

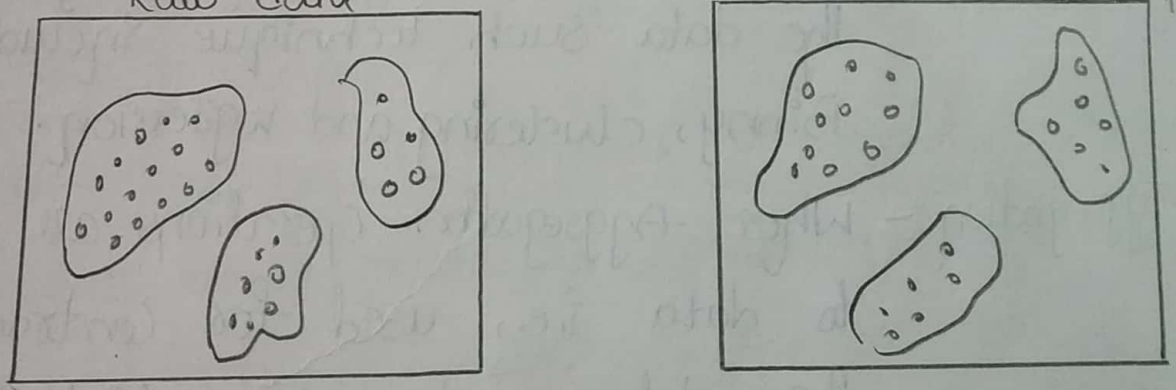
database.

→ used in conjunction with skewed data.

→ Sampling may not reduce database I/O's (page at a time)

Sampling:





segmentation by natural partitioning:

3-4-5 rule can be used to segment numeric data into relatively uniform "natural" intervals.

→ If an interval covers 3, 6, 7, 9 distinct values

at the most significant digit, partition the range into 3 equiwidth intervals.

→ If covers 2,4,8 distrct values at the most significant digit, partion range into 4 intervals

→ If it covers 1,5 or 10 distrct values at the most significant digit, partition the range into 5 intervals.

DATA TRANSFORMATION :-

In this data is transformed from 1 consolidated from to other.

→ Data transformation involves the following methods.

1) Smoothing :- It is used to remove the noisy from the data such technique include Binny, clustering and regression.

2) Aggregation :- Where Aggregation operations are applied to data i.e., used for constructing the data cubes at multiple levels.

3) Generalization :- Where the low level data are replaced by high level concept using concept hirearchies.

4) Normalization :- Where the attribute data falls with in a small specific range

such as 0.0 to 1.0.

5) Attribute construction:-

→ Where the new attributes are constructed and added given from the given set of attributes.

→ Smoothing is a form of data cleaning.

→ Aggregation and Generalization are the forms of Data Reduction.

→ This method is useful

* Normalization :-

The normalization strategy is also known as standization helps us to avoid dependence on the choice of measurement units. As the measurement unit can effect the data analysis.

→ When we try to express an attribute in smaller units (meters to inches) kg - pound etc..,). It will lead to larger range for that trending to give attribute greater effect or weight.

→ Normalization involves transforming the data to fall within a smaller (or) common range

Ex:- [-1, 1] [0.0 - 1.1].

Methods for Normalization :-

1) Minimum Maximum normalization
2) z-score (or) mean normalization
3) Decimal scaling (or) decimal normalization:

* Min - Max normalization :-

Let us assume an attribute 'A' with minimum A, maximum A are maximum values of attribute respectively.

→ Let the attribute A have the values $V_1, V_2, V_3 -- V_i$

$$V_i' = \frac{V_i - Min_A}{Max_A - Min_A} (new\ MAX - new\ min A) + new\ Min\ A$$

* z-score normalization :-

When Min & Max values are not known.

$$V_i' = \frac{V_i - mean_A}{Std.\ dev^n A}$$

* decimal scaling (or) decimal normalization :-

We convert the values in decimals by

$$V_i' = \frac{V_i}{10^j}, \text{ where } j \geq 1$$

*Measures of similarity & disimilarity.

Object in cluster are similar to one another & disimilar to two objects in another cluster.

Similarity & Dismilarity measures are known measures of proximity.

Similarity measure for two objects i,j is equal to '0' then the objects are unlike. The higher is the value of similarity measure the greater is the similarity b/w 2 objects.

Similarity measure is equal to '0'. means the objects are like (similar).

We Commonly use two data by structure for representating similarity & disimilarity.

1) Data matrix
2) disimilarity matrix.

Data Matrix:-

Also known an object by attribute structure It stores 'n' data objects with p attributes in nxp matrix.

Example:-

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & & & & \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & & & & \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity Matrix :- It is object by object structure represented by $d(i,j)$ and similarity is called form dissimilarity as $s(i,j) = 1 - d(i,j)$

representation.

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(n,1) & d(n,2) & \cdots & 0 \end{bmatrix}$$

$d(i,j) \Rightarrow$ diff b/w obj i,j

$d(i,j)$ is non-negative number

$d(i,j) = 0$

$d(i,j) = d(j,i)$

Proximity measure for nominal attribute :-

Example :- young senior middle

dissimilarity $d(i,j)$ of nominal attributes

$$= \frac{P - m}{P}.$$

P = total no. of matches

m = no. of similar matches.

$$S(i,j) = \frac{1 - P - m}{P} \Rightarrow \frac{m}{P}$$

Proximity measure for Binary attributes:-

Binary data has values 0 or 1.

d(i,j) for binary data is represented in Contigency matrix.

Contigency Matrix:-

	j		
	1	0	sum
i	q	r	q+r
0	s	t	s+t
sum	q+s	r+t	P

q : i=1 j=1

r : i=1 j=0

s : i=0 j=1

t : i=0 j=0

total no. of attributes $p = q+r+s+t$

dissimilarity that is based on symmetric binary attributes is called Symmetric binary dissimilarity.

If the two states are equally important them they are termed as symmetric binary attributes.

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

$$\text{Similarity } s(i,j) = 1 - d(i,j) = \frac{q}{q+r+s}$$

↓

which is equal to Saccaard Co-efficient.

for assymetric binary attributes, the dissimilarities termed as asymetric binary similarity where 't' is ignored.

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

Dismilarity for Numeric Data:-

It is calculated by eckaadian, man hattan & minkowski distance. The most

Popular is euclidian distance where.

$i = x_{i1}, x_{i2}, x_{i3}, \ldots$

$j = x_{j1}, x_{j2}, x_{j3} \ldots$

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 \ldots}$$

Manhattan Distance :- $d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| \ldots |x_{ip} - x_{jp}|$

Minkowski Distance :- It is generalized form of Eucleadan & manhattan where $d(i,j) = h$.

$$h = \sqrt{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h \ldots |x_{ip} - x_{jp}|^h}$$

where $h \geq 1$

Eg real number.

* Short Answer Questions:-

1) What is data mining?

Data mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications.

1) Market Analysis
2) fraud detection
3) Customer Retention
4) production control
5) Science Exploration.

2) Difference b/w datamining & OLAP?

Datamining	OLAP
1) Bottom up, discovery-driven	1) Top-down, query-driven
2) Requires no assumptions	2) Repetitive testing of user originated theories.
3) No intensive human interaction with the db is required.	3) Requires a great deal of human interaction with the data base.

4) Runs mostly automati-cally	4) users must have a good idea about the informatn which she/he is looking for.
5) User interaction is limited to the selection of data mining alg & the selection of appropriate parameter.	5) user is in permanent interaction with the system.
6) Why is this happening? And. What might happen if--?	6) → Is this true?

3) What are the types & tasks that arrived during data mining?

4) → The set of task-relavant data to be minned.

→ The kind of knowledge to be mined.

→ The background knowledge to be used in discovery.

→ The intresting ness measures & thesholds for pattern evalution.

→ The expected representation of visualizing the discovered pattern.

4) What is the need for preprocessing of data?

A) In Complete noisy & in-consistent data are common place properties of large real world databases & data ware houses. Incomplete data can occur for a no. of reasons. Attributes of intrest may not always be available, such as customer information for sales transition data. other data may not be included simply because it was not considered imp at the time of entry relevant data may not be recorded due to misunderstanding. Data may have been overlooked. Missing data, particularly for tuples with missing values for same attributes, may need to be inferred.

5) What say knowledge discovery?

Some people treat data mining some as knowledge discovery while other view data mining as an essential step in the process of knowledge discovery. Here is the list of steps. involved in the knowledge, discovery process.

* Data cleaning
* Data integration
* Data transformation
* Data mining
* pattern Evaluation
* knowledge presentation.

6) What is dimensionality reduction?

In dimensionality reduction, data encoding or transformation are applied as to obtain a reduced or "compressed" represent of the original data. If the original data can be reconstructed from the compressed data without any loss of information, the data reduction is called lossless.

7) What are the data preprocessing tasks?

* Data cleaning
* Data Integration
* Data Transformation
* Data reduction.

What is data cleaning?

Data cleaning is a technique that is applied to remove the noisy data & correct the inconsistency in data. Data

Cleaning involves transformations to correct the wrong data. Data cleaning is performed on a data preprocessing step while preparing the data from a data warehouse.

9) Issues during data Integration?

A) The issues to be considered during data integration are

(i) Matching
(ii) Redundancy
(iii) Detection & resolution of data value-conflicts.

10) strategies of data reduction?

A) (i) Data cube aggregation
(ii) Attribute subset selection
(iii) Dimensionality reduction
(iv) Numerosity reduction
(v) Discretization & concept hireachy generation.

11) How do you clean the data? for missing values.

A) (i) Ignore the typle
(ii) fill in the missing value manually
(iii) use a global constant to fill in the missing value

(iv) Use the attribute means to fill in the missing value.
(v) Use the attribute mean for all samples belonging to same class as given tuple.
(vi) Use most probably value to fill in the missing value. for Noisy data:-

1) Binning
2) Regression
3) Clustering

12) What is concept hisarcy & how it is useful in data mining?

A) Θ concept hirearchy for a given numerical attribute defines a discretization of the attribute. Concept hirearchies can be used to reduce the data by collecting & replacing low-level concepts with higher level concepts. Although detail is lost by such data. generalization, the generalized data may be more meaningful & easier to interpret.

What is data generalization?

It is process that abstracts a large set of task relevant data in a database from a relatively low Conceptual levels

to higher conceptual levels 2 approaches for Generalization.

1) Data cube approach
2) Attribute oriented induction approach.

14) List the primitives for specifing a data mining table.

A) → The set of task-relevant data to be mined
→ The kind of knowledge to be mined
→ The background knowledge to be used in the discovery process.
→ The intrestingness measures & threshold for pattern evaluation.
→ The respected representation for visualizing the discovered pattern.

15) What are the steps involved in KDD process.

A) 1) Developing & understanding of
→ The application domain
→ The relevant prior knowledge
→ The goals of the end user.

2) Creating a target data set
3) Data cleaning & pre processing
4) Data reduction & projection

5) Choosing the data mining task
6) choosing the data mining alg.
7) Data mining
8) Interpreting mined patterns
9) Consolidating discovered knowledge.

16) Mention some of datamining techniques.

There are several major data mining techniques have been developing & using in data mining projects recently. including association, classification, clustering, prediction, Sequential patterns.

17) What is meta learning?

Meta learning is a sub field of Machine learning where automatic learning algorithms are applied on meta Data about machine learning experiments.

18) Define summarization?

) "Summarization is a key data mining Concept which involves techniques for finding a Compact discription of a data sets. ... i.e., data visualization of automated report generation. Clustering [13,23] is another data mining techniques that is often used to summarize

Large datasets".

19) Classification of datamining system?

Datamining is an interdisciplinary field, the conference of a set of disciplies, including database systems, statistics, machine learning, visualization & information science, Data mining system can be categorized a/c to various criteria.

20) Challenges in Data mining?

A) * Introduction.
* Noisy & Incomplete data
* Distributed data
* Complex data
* performance
* Incorporation of Background knowledge
* Data visualization.
* Data privacy & security.

21) Identify any 3 functionality of Data Mining?

The functionality of Data Mining are:

1) Concept/class Description: Data characterization & Data Discrimination

2) Mining freq patterns, Associations & Correlations.
3) Association analysis.

22) Interpret major issues of Data mining?

The major issues of Data mining are:-

(i) Limited and irrelevant information
(ii) Noisy and irrelevant information
(iii) Human interaction & knowledge
(iv) Large data sets and high dimensionality
(v) Uncertainty
(vi) Dynamic updates.

23) Discuss relational databases?

A relational databases (RDB) is a collective set of multiple data sets organized by tables, records and columns. RDBs establish a well-defined relationship b/w database tables. Tables communicate and share information, which facilitates data searchability, organization and reporting.

RDBs use Structured Query Language (SQL), which is a standard user application that provides an easy programming interface for database interaction.

24) State object-oriented databases?

An object-oriented database is a system

offering database management facilities in an object-oriented programming environment. Data is stored as objects and can be interpreted only using the methods specified by its class. An object class is a set of objects that share a common structure and a common behaviour.

An object-oriented database is fundamentally created around an object-oriented data model.

25) Explain spatial databases?

A spatial databases is a database that is optimised for storing and querying data that represents objects defined in a geometric space. Most spatial databases allow the representation of simple geometric objects such as points, lines and polygons. Some spatial databases handle more complex structures such as 3D objects, topological coverages & linear n/ws.

26) Explain the outlier Analysis?

A database may contain data objects that do not comply with the general behavior or model of the data objects, which are grossly different from or inconsistent

with the remaining set of data, are called outliers.

→ Most data mining methods discard outliers as noise or exceptions.

27) Express what is a decision tree?

→ A decision tree is a flow-chart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, & each leaf node holds a class label. The TOP most node in a tree is the root node.

28) List the types of data that can be mined?

→ database data
→ data warehouse data
→ transactional data
→ data streams.
→ Spatial data
→ engineering design data
→ hypertext and multi-media data.
→ graph and networked data.

29) Describe transactional data bases?

In general, each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user's click on a

web page. A transaction typically includes a unique transaction identity number (trans_ID) and a list of the items making up the transaction; such as the items purchased in the transaction. A trans-actional database may have additional tables, which contain other information related to the transactions, such as item description, information about the salesperson or the branch, and so on:

30) Define multidimensional data mining?

Multidimensional data mining is an app-roach to data mining that integrates OLAP-based data analysis with knowledge discovery techniques.

→ It is also known as exploratory multidimensional data mining and online analytical mining (OLAM).

→ It searches for intresting patterns by exploring the data in multidimensional space.

31) Define Data cube?

A data cube is defined as a collection of dimensions & facts. Data cube allows the data to be modeled & viewed in multi-dimensions.

Ex:- 2D view.

32) Define data characterization?

Data characterization is a summarization of the general characteristics or features of a target class of data.

→ The data corresponding to the user-specified class are typically collected by a query.

33) Contrast heterogeneous databases and legacy databases?

A) Heterogeneous databases:-

→ A heterogeneous database consists of a set of interco-nnected, autonomous component databases.

→ The components communicate in order to exchange information and answer queries.

→ Component DB may differ greatly, Hence difficult to assimilate the information.

Legacy databases:-

→ A legacy database is a group of heterogeneous databases that combines different kinds of data systems, such as relational (or) object-oriented databases, hirerarchical dBs, n/w DBs, spreadsheets, multimedia dBs or file systems.

34) Differentiate classification & prediction?

Classification	prediction.
→ Classification is the process of identifying to which category, a new observation belongs to on the basis of a training data set containing observations whose category membership is known.	→ prediction is the process of identifying the missing or unavailable numerical data for a new observation
→ In classification, the accuracy depends on finding the class label correctly.	→ In prediction the accuracy depends on how well a given predicator can guess the value of a predicated attribute for a new data.
→ In classification, the model can be known as the classifier.	→ In prediction, the model can be known as the predictor.
→ A model or the classifier is constructed to find the categorical labels.	→ A model or a predictor will be constructed that predicts a continuous-valued function or ordered value.