5/12/18

UNIT-2

Data Preprocessing

* Data preprocessing: overview
  - → Data quality
  - → Major Tasks in Data preprocessing
* Data cleaning
* Data Integration
* Data Reduction
* Data Transformation and Data Discretization
* Summary

Data Quality:-

Measures for data quality

Accuracy: correct or Wrong, accurate or not

Completeness: Not recorded, unavailable

consistency: some modified but some not.

Timeliness: timely update?

Believability: how trustable the data are correct?

Interpretability: how easily the data can be understood?

Major Tasks in preprocessing

Data cleaning - fill the missing values
- Smooth noisy data
- Identify (or) Remove outliers
- Resolve inconsistencies.

Data Integration - Integration of multiple databases
- data cubes (or) files

Data reduction - Dimensionality Reduction
- Numerosity Reduction.
- Data Compression

Transformation & Discretization - Normalization
- concept hierarchy generation.

* Data cleaning:-

Incomplete:- Lacking attribute values

eg: Occupation = " " (missing data)

noise:- containing noise, errors or outliers

eg: Salary = "-10" (error)

inconsistent:- containing discrepancies in codes (or) names eg: Age = "42", Birthday = "03/07/2010"

eg: discrepancy between duplicate records.

How to handle incomplete(missing) data:-

* Ignore the tuple
* fill the missing values manually
* fill the missing value by using a global constant "unknown", ∝.
* Use attribute mean for all samples belongs to the same class.
* Use most probable value

Noisy data:-

Noise:- Random error (or) variance in a measured variable.

* Incorrect attribute values may be faulty due to
  - → faulty data collection instruments
  - → data entry problems
  - → data transmission
  - → technology limitation
  - → inconsistency in naming convention

the following techniques are used to handle the noisy data

1. Biming
2. clustering
3. Regression
4. Combined computer and human inspection.

Binning:- This method is used to smoothout a sorted data by consulting the values around it, the sorted values are distributed into a no.of "buckets" (or) "bins" of equal Frequency.

i) **Smoothing by means Bin means:-**

Each value in a bin replaced with mean value of the bin.

ii) **Smoothing by Bin median:-**

Each value in a bin is replaced with median value of bin.

iii) **Smoothing by Bin boundaries:-**

The minimum and maximum values in a given bin are identified as bin boundaries, each bin value is replaced with closest boundary value.

Ex:- Sorted data for price (in dollars)

4, 8, 15, 21, 21, 24, 25, 28, 34

partition the given data into equal frequency of bins

Bin1 : 4, 8, 15

Bin2 : 21, 21, 24

Bin3 : 25, 28, 34.

Smoothing by bin mean:-

Bin1 : 9, 9, 9 $\quad \frac{4+8+15}{3}$

Bin2 : 22, 22, 22

Bin3 : 29, 29, 29

Smoothing by bin median:-

Bin1 : 8, 8, 8

Bin2 : 21, 21, 21

Bin3 : 28, 28, 28.

By Boundaries:-

Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

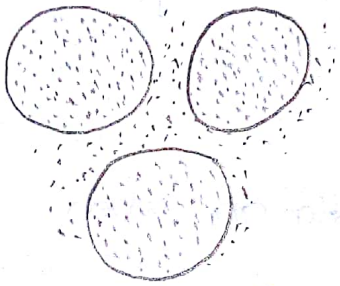
Minimum = 4

Maximum = 15

clustering:-

*Outliers may be detected by clustering, where similar values are organised into groups (or) clusters.

*The values that fall outside of the clusters, considered as outliers.



Regression:- Data can be smoothed by fitting the data to a function is known as regression.

Linear regression:- It involves finding the best line to fit two variables, so that one variable can be used to predict the other.

Multiple Regression:- Where more than two variables are involve to fit the data to a multi-dimensional surface.

combine Computer and Human inspection:-

detect suspicious values and check by human (eg: deal with possible outliers)

Inconsistent:-

*Data inconsistence is may be corrected manually, for example errors made at data entry, may be corrected by performing a paper trace.

Data Integration:-

* combines data from multiple sources into a coherent data store.

→ Schema structure Integration: eg:- A.custid = B.cust-#

* Integration metadata from different sources.

i) Entity identification problem:

* Identify real world entities from multiple data sources,

eg:- Bill Clinton = William Clinton

→ Detecting and resolving data value conflicts

* For the same real world entity, attribute values from different source are different.

ii) Redundancy:-

* Redundant data occur often on when integration of multiple databases.

Object identification:- The same attribute or object may have different names in different databases.

Derived Data:- One attribute may be a "derived" attribute in another table, eg:- annual revenue.

* Redundant attributes may be able to detected by correlation analysis and covariance analysis.

Correlation Analysis (Nominal Data):-

χ^2 (chi-square) test

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

eg:-

	play chess	Not play chess	Sum(row)
Like science fiction	250(90) expected	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col).	300	1200	1500

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

* It shows that like_science_fiction and play_chess are correlated in the group.

correlation Analysis (Numeric Data):-

* correlation coefficient also called pearson's product moment coefficient.

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

→ where n is the no.of tuples, $\bar{A}$ and $\bar{B}$ are the means of A and B.

* σ_A and σ_B are the standard deviation of A and B.

* $\Sigma(a_i b_i)$ is the sum of the AB cross product.

* If $r_{A,B} > 0$, A & B are positively correlated

* If $r_{A,B} = 0$ independent.

* If $r_{AB} < 0$ negatively correlated.

Covariance (Numeric Data):-

* Covariance is similar to correlation.

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}.$$

* If $Cov_{A,B} > 0$ then both are positively covariated.

* If $Cov_{A,B} < 0$ then both are Negatively covariated.

* If $Cov_{A,B} = 0$ independent.

* $Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$

ex: Suppose two stocks A and B have the following values in one week: (2,5), (3,8), (5,10), (4,11), (6,14).

Question: If the stocks are affected by same industry trend will their prices rise or fall together?

$$E(A) = \frac{2+3+5+4+6}{5} = \frac{20}{5} = 4$$

$$E(B) = \frac{5+8+10+11+14}{5} = \frac{48}{5} = 9.6$$

$$cov(A,B) = (2\times5 + 3\times8 + 5\times10 + 4\times11 + 6\times14)/5 - 4\times9.6 = 4$$

Thus A & B rise together since $cov(A,B) > 0$.

Data Reduction:-

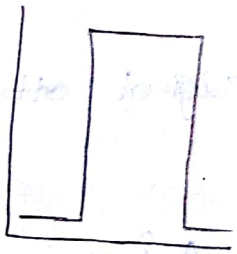
* Reduce representation of the data set that is much smaller in volume but yet produces the same analytical results.
* Why data Reduction? - A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
* Data reduction strategies.
* Dimensionality reduction, eg: remove unimportant attributes.
  - → wavelet transforms
  - → principal component analysis (PCA)
  - → Feature subset selection, feature creation.
* Numerosity reduction (some simply call it: Data Reduction)
  - → Regression and log-linear models.
  - → Histograms, clustering and sampling
  - → Data cube aggregation
* Data compression.

Dimensionality Reduction:-

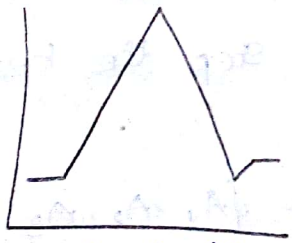
* Avoid the curse of dimensionality
* Help eliminate irrelevant features and reduce noise
* Reduce time and space required in data mining
* Allow easier visualization.

Wavelet Transforms:-

* Decomposes a signal into different frequency sub bands
* It is applicable to n dimensional signals.
* Data are transformed to preserve relative distance between objects at different levels of resolution
* Discrete wavelet Transform (DWT) for linear signal processing multi resolution analysis.
* It stores a strongest of wavelet co-efficients.



Haar2

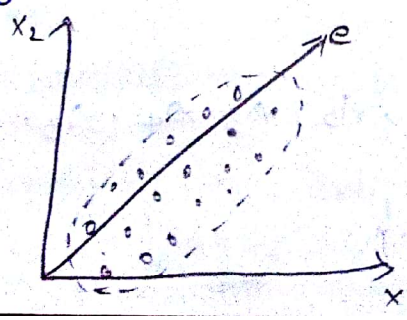


Daubechie4

* Emphasize region where points cluster.
* Suppress weaker information.

Principle Component Analysis (PCA):-

* the original Data are projected into a smaller space, resulting in dimensionality Reduction.
* We find Eigen vectors, these define the new space.



* Given N data vectors from n-dimensions, we find $k \leq n$ orthogonal vectors (principle components) that can be used to represent the data.

* Compute k orthonormal vectors i.e principle components.

* Each input vector is a combination of k principle component vectors.

* These components are sorted in decreasing order.

* The size of the data can be reduced by eliminating the weak components.

Attribute subset selection:-

* this includes the following techniques:

1. Forward Selection:- The procedure starts with an empty set of attributes, the best from original attribute is determine and added to the set.

* At each and every step the best of original attribute is added to the set.

eg:* Initial Data set = $\{A_1, A_2, A_3, A_4, A_5, A_6\}$

Step 1:- $\{\}$

Step 2: $\{A_1\}$

Step 3: $\{A_1, A_4\}$

Step 4: $\{A_1, A_4, A_6\}$

Backward Elimination:- The procedure starts with full set of attributes, at each and every step removes the worst attribute from the remaining set.

eg:- Initial Data set = $\{A_1, A_2, A_3, A_4, A_5, A_6\}$

$\{A_1, A_3, A_4, A_5, A_6\}$

{A1, A4, A5, A6}

{A1, A4, A6}

combine forward Selection & Backward Elimination:-

At each and every step selecting best attribute and removes worst attribute in the given data set.

Attribute creation (Feature Generation)

Creating a new attribute by using existing dimensions.

This is attribute specific

It can be used on multi-dimension space.

Reduction2 - Numerosity Reduction:-

* Reduce data volume by choosing alternative, smaller forms of data Representation.

* The following methods are used for numerosity reduction.

1. parametric methods (eg: regression)

* Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data

eg:-log-linear models

2. Non-parametric methods

* Do not assume models

* Major families, histogram, clustering, sampling.

Regression and log-linear models.

Linear regression:- Data modeled to fit a straight line.

Multiple regression:- Allows a response variable Y to be modeled as a linear function of multi-dimensional feature vector.

Log-linear model:-

* Approximates discrete multidimensional probability distributions.

* Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on

a smaller subset of dimensions.

* It is useful for dimensionality reduction and smoothing.

Histogram:-

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

Partitioning rules:

Equal-width: equal bucket range

Equal-frequency (or equal depth).

Sampling:-

Obtaining a small sample S to represent the whole data set

Techniques:- 1. Simple random sampling

2. Sampling without replacement

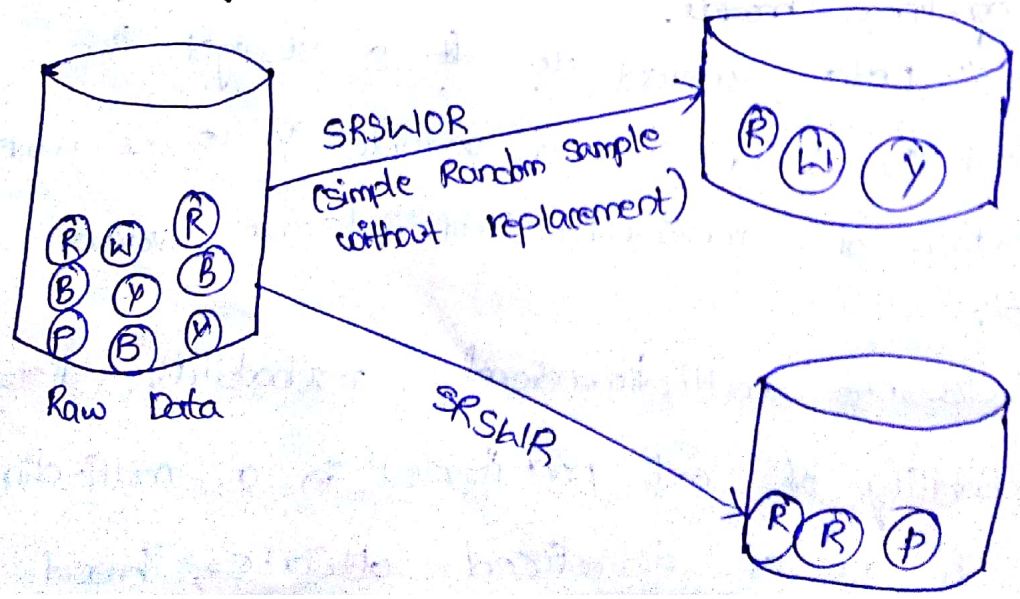
5. Sampling with replacement

4. Stratified sampling.

1. There is an equal probability of selecting any particular item
2. Once an object is selected, it is removed from the population
3. A selected object is not Removed from the population.
4. partition the data set, and draw samples from each partition

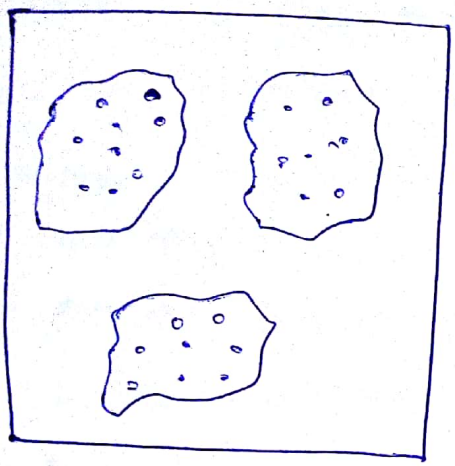
approximately the same percentage of the data.

Eg:- sampling: with or without replacement.

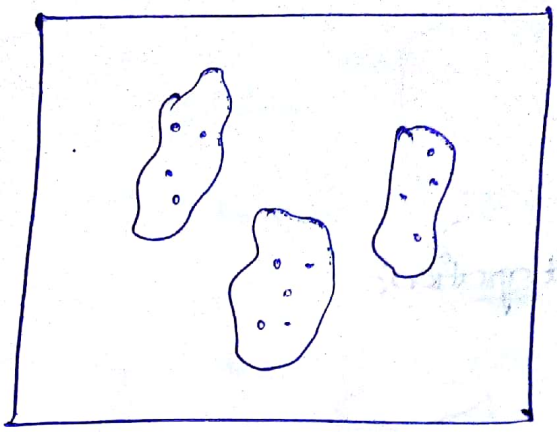


Stratified Sampling:-

Raw Data



cluster stratified sample.



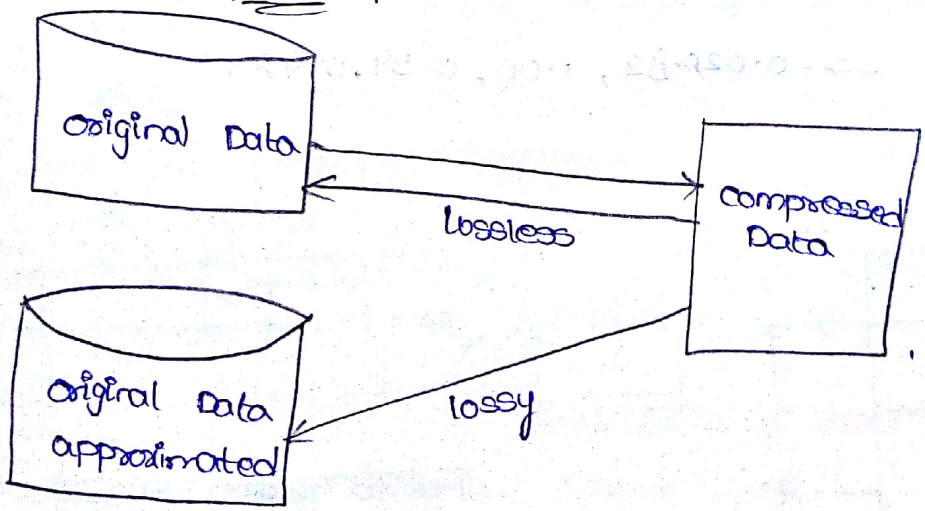
Data compression:-

* String compression:- lossless, limited manipulation is possible but only without expansion.

Audio/Video compression:-

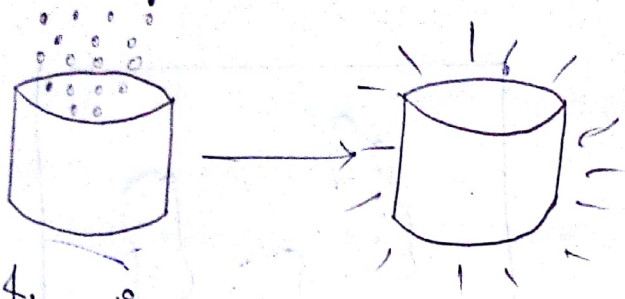
* lossy compression, with progressive refinement.
* sometimes small fragments of signal can be reconstructed without
* time sequence is not audio.

Data compression:-

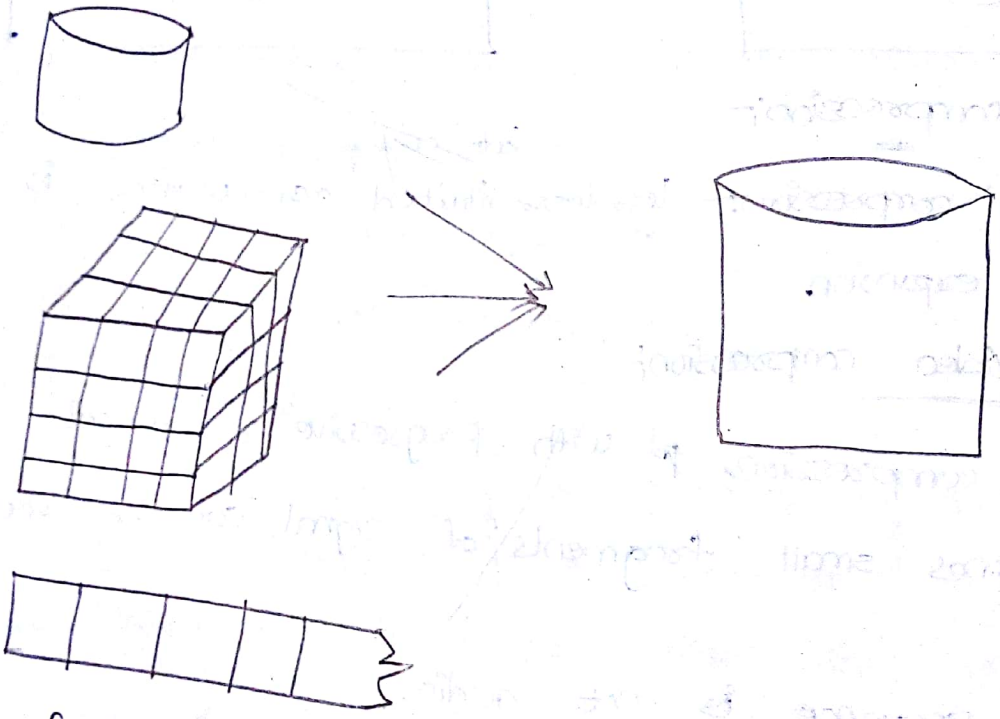


Forms of data preprocessing:-

Data cleaning:-



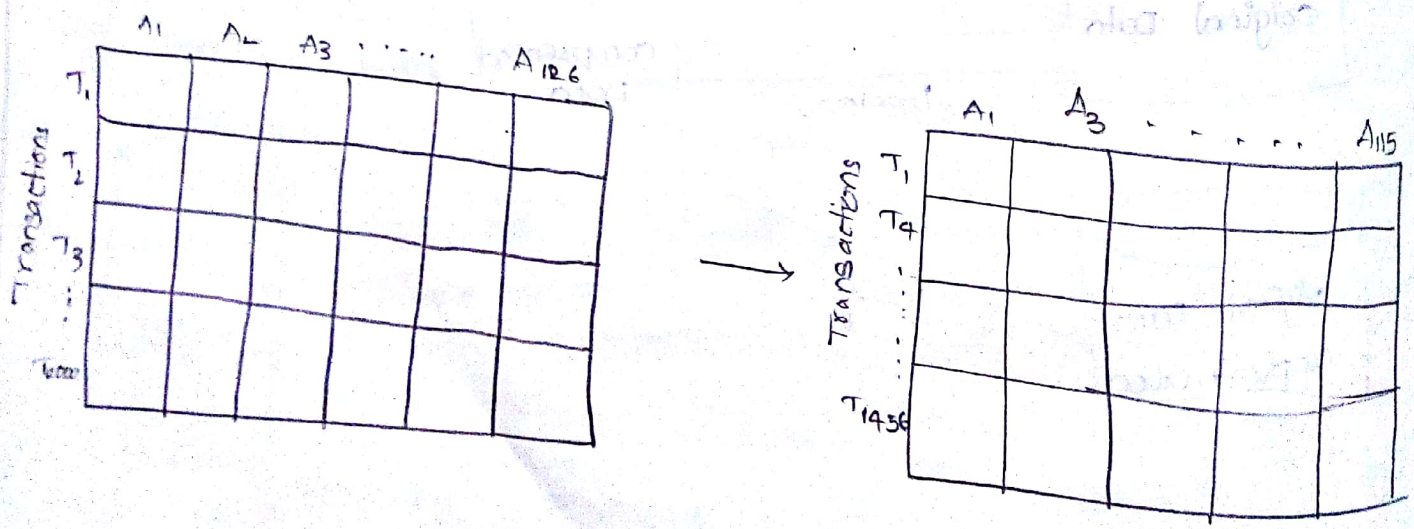
Data Integration:-



Data transformation:-

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48.

Data Reduction:-



Data transformation:-

* A function that maps the entire set of values of a given attribute to new set of replacement values i.e each old value can be identified with one of the new values.

Methods:-

* smoothing: Remove noise from the data.
* Attribute/feature construction
  - New attribute constructed from the given ones
* Aggregation: summarization, data cube construction.
* Normalization: Scaled to fall within a smaller, specified range
  - Min-Max normalization
  - Z-score normalization
  - Normalization by decimal scaling

Normalization:-

Min-Max normalization:-

to $[new_min_A, new_Max_A]$

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A.$$

min = $12,000

max = $98,000

new min = 0.1

new max = 1.0

v' = $73,000

$$= \frac{73.600 - 12,000}{98000 - 12000} (1.0 - 0) + 0 = 0.716$$

Z-score normalization:-

$$v' = \frac{v - \mu_A}{\sigma_A}$$

where μ: mean

σ: standard deviation

let $\mu = 54,000$, $\sigma = 16,000$

then. $\frac{73,600 - 54,000}{16000} = 1.225.$

Normalization by decimal scaling:-

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $Max(|v'|) < 1$

Discretization:- three types of attributes.

Nominal:- values from an unordered set eg:- color, profession.

ordinal:- values from an ordered set eg:- military (or) academic rank.

Numeric:- Real numbers eg:- integer (or) real numbers.

Discretization:-

* Divide the range of continuous attribute into intervals.
* Interval labels can then be used to replace actual data value
* Reduce data size by discretization
* supervised vs. unsupervised
* split (top down) vs. merge (bottom-up)
* Discretization can be performed recursively on an attribute
* prepare for further analysis eg:- classification.

Methods for discretization:-

Typical Methods:- All the methods can be applied recursively

→ Binning
  * Top-down split, unsupervised
→ histogram analysis
  * Top-down split, unsupervised
→ clustering analysis (unsupervised, top-down split or bottom-up merge)
→ Decision-tree analysis (supervised, top-down split)
→ Correlation (eg: χ^2) analysis (unsupervised, bottom-up merge)

concept Hierarchy generation:-

* organizes concepts hierarchically & is usually associated with each dimension in a data ware house.

* Recursively reduce the data by collecting and replacing low level concepts by higher level concepts.

eg:- Age: youth, adult or senior

* concept hierarchy can be automatically form for both numeric and nominal data.

concept hierarchy for nominal data:-

eg: set of attribute {street, city, state, country}

concept hierarchy is street < city < state < country

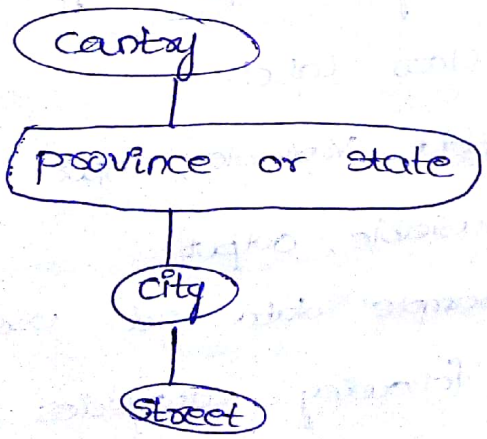
concept hierarchy for numeric data:-

eg:- set of attribute: {child age, young age, middle age old age}

child age < young age < middle age < old age.

Automatic concept hierarchy generation:-

The attribute with the most distinct values is placed at the lowest level of the hierarchy.



15 distinct values

365 distinct values

3334 distinct values

64-530 distinct values.