Unit - II

Data mining refers to extracting or "mining" knowledge from large amounts of data.

① Data cleaning : (to remove noise and inconsistent data)

② Data integration (where multiple data source may be combined)

③ Data selection (where data relevant to the analysis task are retrieved from the database)

④ Data Transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations)

⑤ Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

⑥ Pattern evaluation (to identify the truly interestingness patterns representing knowledge based on some interestingness measures)

⑦ Knowledge presentation (where visualization & knowledge representation techniques are used to present the mined knowledge to the user).

Step 1 to 4 are different forms of data preprocessing, where the data are prepared for mining.

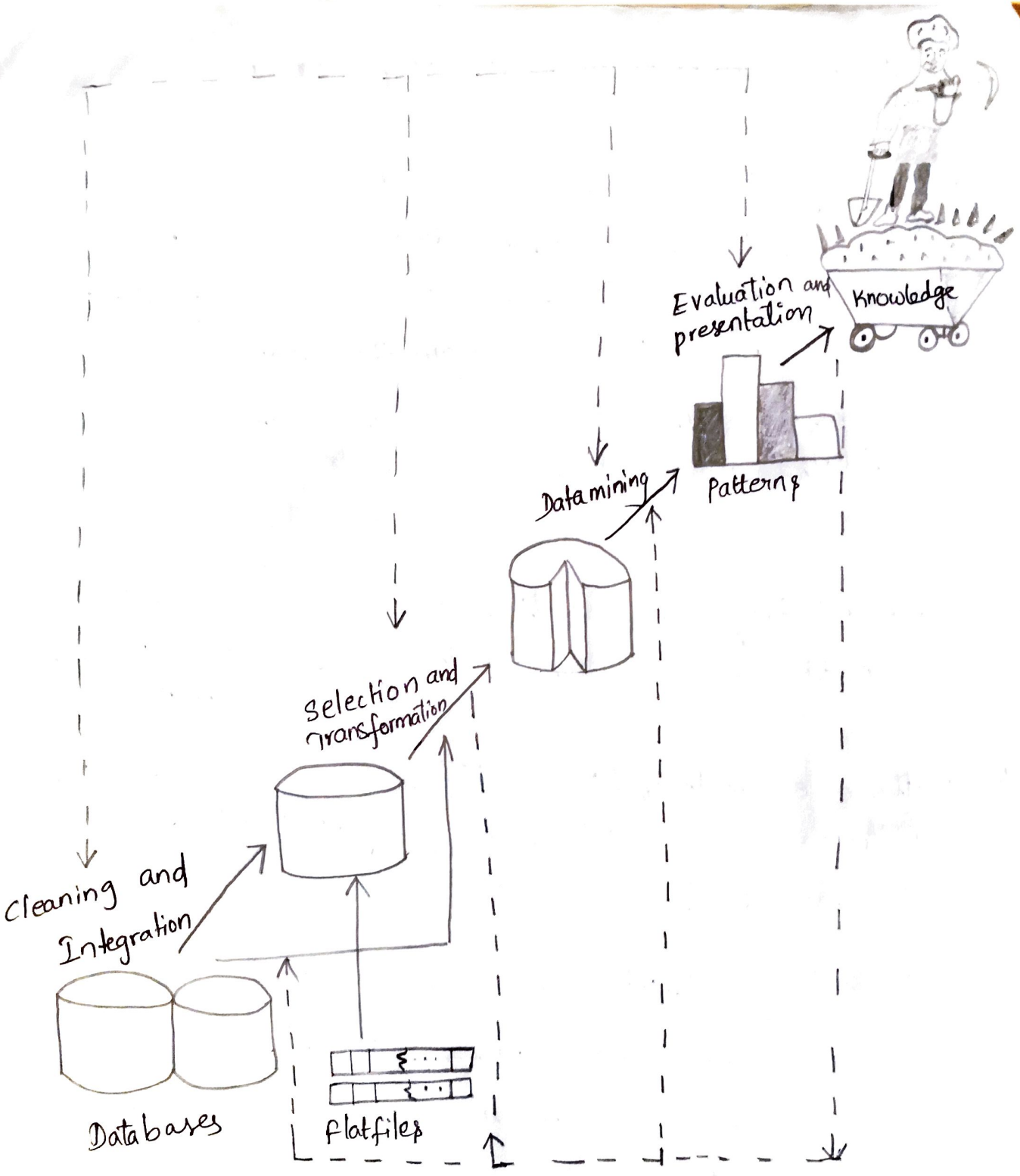


fig: Datamining as a step in the process of Knowledge discovery (KDD)

Database or datawarehouse server : The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request

Knowledge base : This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction.

Data mining engine : This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis and evolution analysis

Pattern evaluation module : This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns.

User Interface : This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a datamining query or task, providing information to help focus the search and performing exploratory data mining based on the intermediate data mining results.

Based on this view, the architecture of a typical datamining system may have the following major components.

(1) Database, datawarehouse, www, or other information repository: This is one or a set of databases, datawarehouses, spreadsheets or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

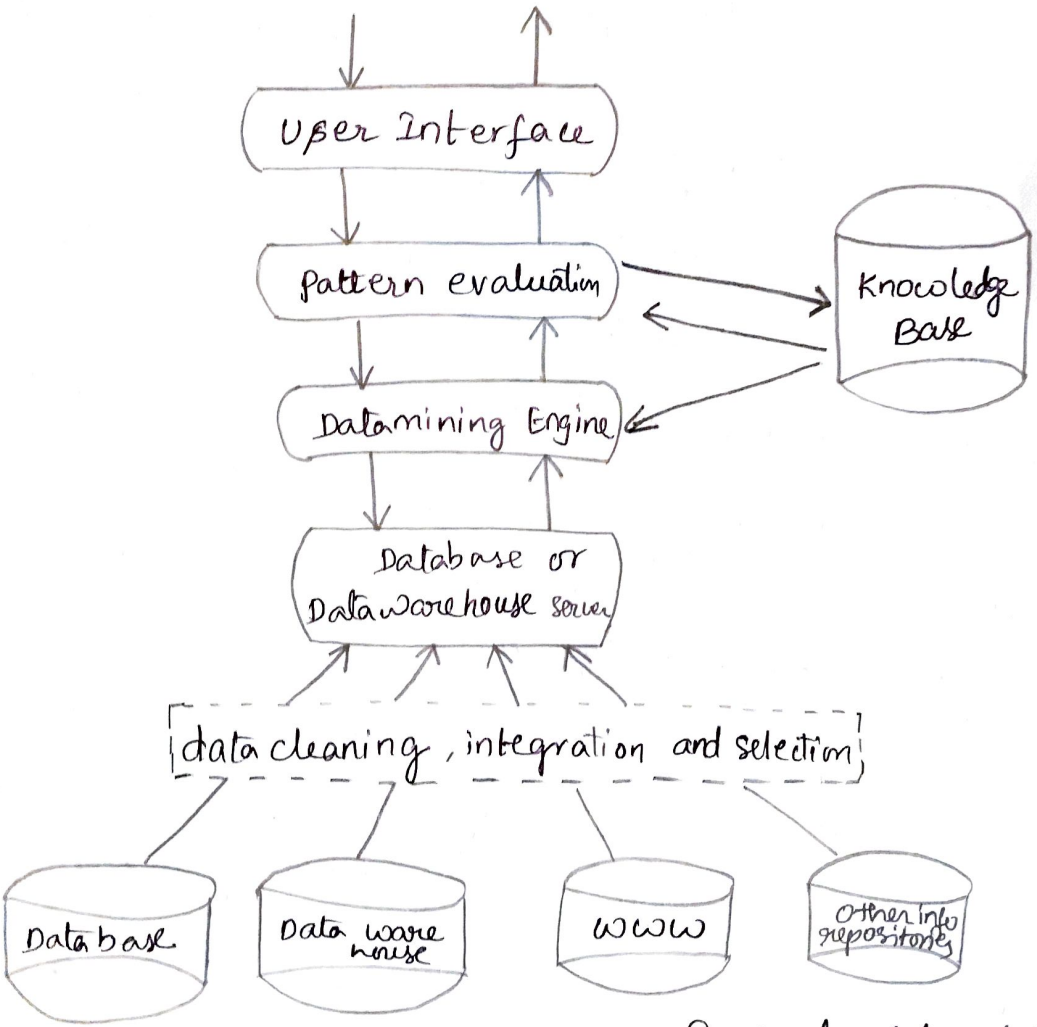


Fig: Architecture of a Typical data mining System

(2)

Data Mining - On what kind of Data?

1. Relational Databases
2. Data warehouses
3. Transactional Databases
4. Advanced data and information systems and advanced applications

Object-Relational Databases

Temporal Databases, Sequence Databases and Time-series DB's.

Spatial DB's and Spatiotemporal DB's

Text DB's and Multimedia DB's

Heterogeneous DB's and Legacy DB's.

Data streams.

The WWW

Relational DB's

A Relational DB is a collection of tables, each of which is assigned a unique name.

→ Each table consists of a set of attributes (col's or fields) and usually stores a large set of tuples (records or rows).

→ Each tuple in a relational table rep an object identified by a unique key and described by a set of attribute values.

→ An ER data model rep the DB as a set of entities and their relationships.

Customer

Cust-ID	name	address	age	income	credit-info	category	...
C1	Smith, Sandy	1223 Lake Ave, Chicago, IL	31	$78000	1	3	
...	...	...	...	...	...	...	...

item

item-IP	name	brand	category	type	price	place-made	supplier	
13	hi-res TV	Toshiba	high resolution	TV	$988.00	Japan	Nikox	$600
18	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.0
...	...	...	...	...	...	...	...	

employee (emp_ID, name, category, group, salary, commission)

Branch (branch_ID, name, address)

purchases (transID, cust ID, emp-ID, date, time, method_paid, amount)

items_sold (trans-ID, item-ID, qty)

works_at (emp-ID, branch-ID).

(2) DWH

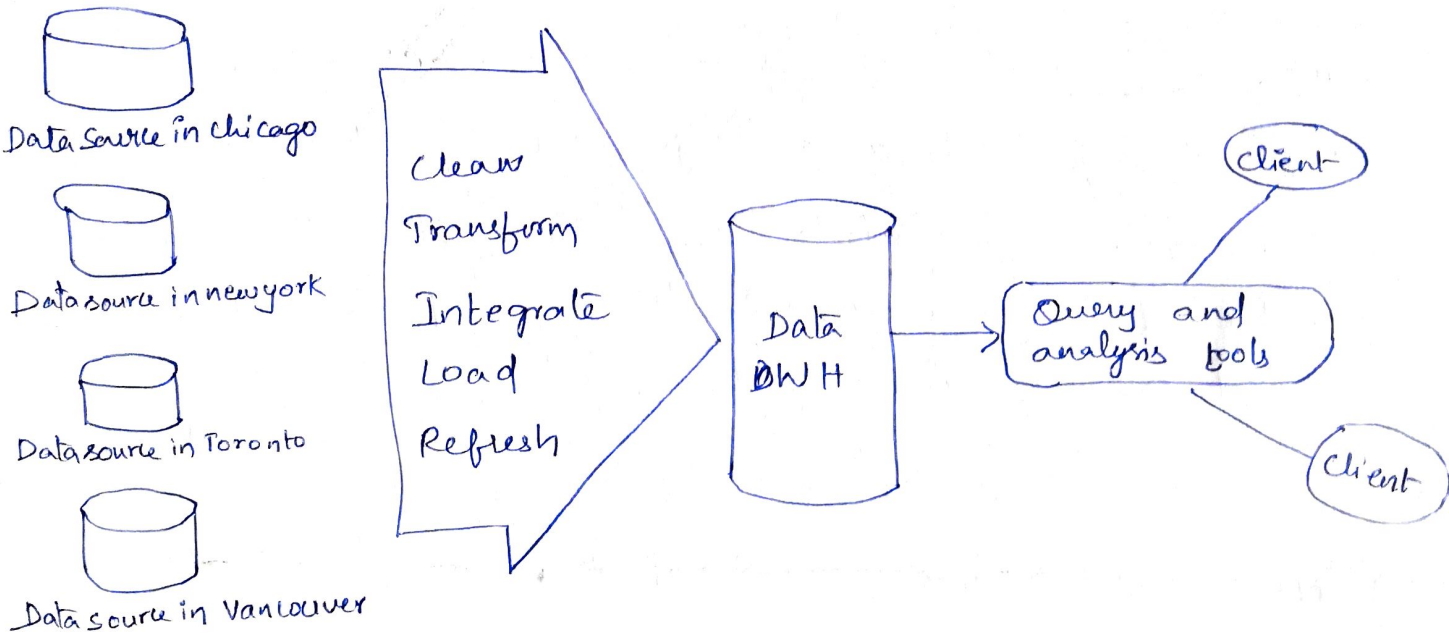


fig: Typical architecture of a datawarehouse for All Electronics

cost)
500.00
3.00

Transcational Databases

sales

trans-ID	list of item-IDs
T100	I1, I3, I8, I16
...	....

(5) Advanced DB systems and Advanced DB Applications

→ The new DB applications include handling spatial data (such as maps), engineering design data (such as the design of buildings, system components or integrated circuits), hypertext and multimedia data (including text, image, video and audio data), time related data (such as historical records or stock exchange data) and the www (a huge, widely distributed information repository made available by the Internet).

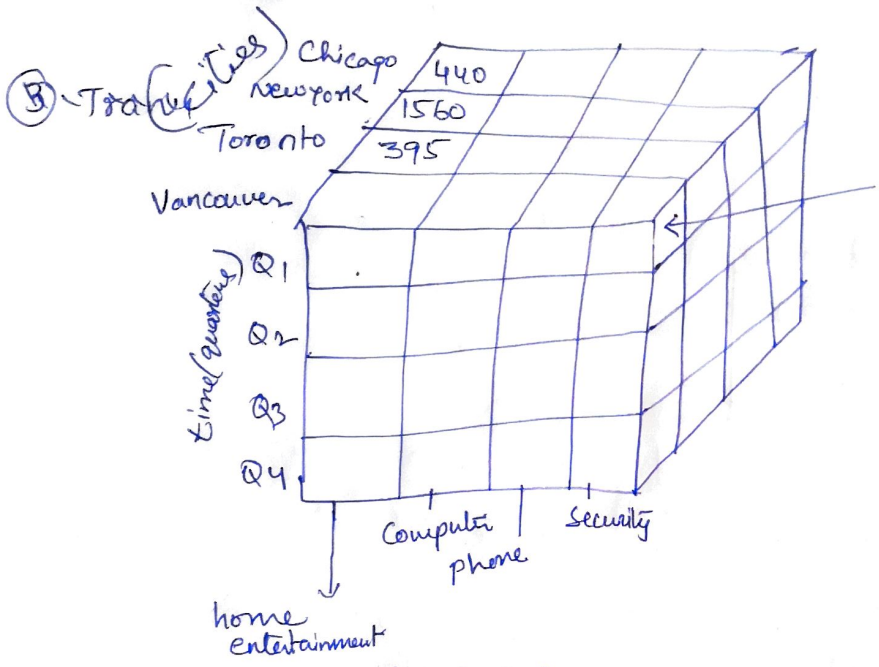
Object-Oriented DBs

→ Data and code relating to an object are encapsulated into a single unit.

→ Each object is associated with a set of variables, messages & methods.

Object-Relational DBs

→ The object-relational model extends the basic relational data model by adding the power to handle complex data ~~model~~ types, class hierarchies, and object inheritance.



(a) A multidimensional data cube, showing summarized data for ALLElectronics

(b)

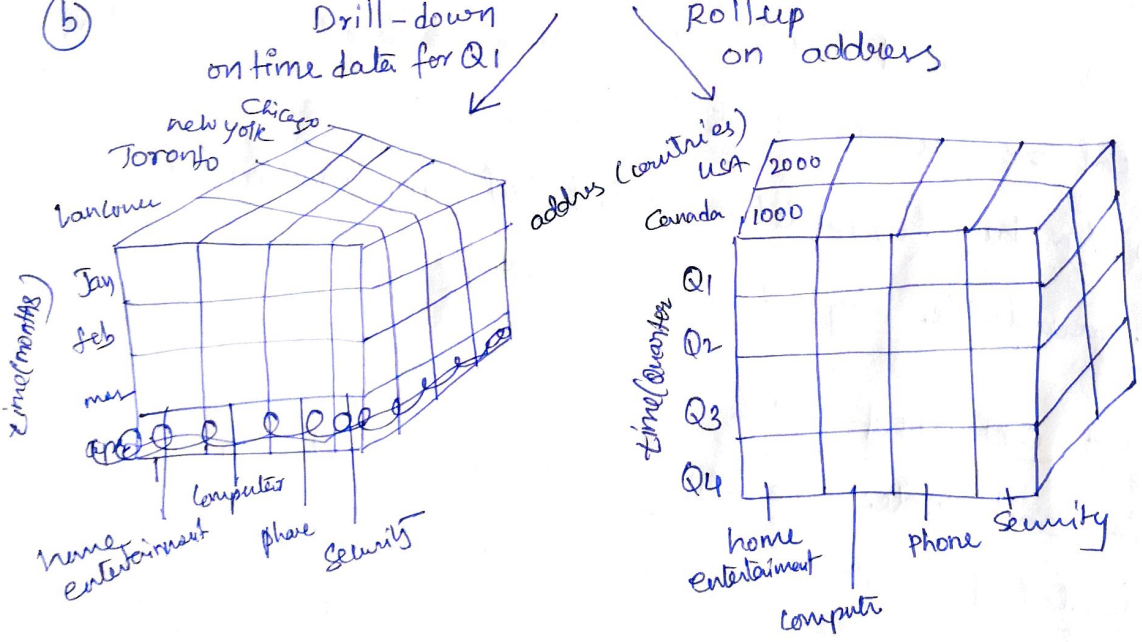


fig: showing summarized data resulting from drill-down and roll-up operations on the cube in.

Spatial DB's : contain spatial-related information.

→ such as DB's includes geographic (map) DB's, VLSI chip design DB's and medical and satellite image DB's.

Temporal DB's and Time-series DB's

→ Both store time-related data.

→ A time-series DB stores sequences of values that change with time, such as data collected regarding the stock exchange.

→ A temporal DB usually stores relational data that include time-related attributes (timestamps).

Text DB's and Multimedia DB's

→ Text DB's are DB's that contain word descriptions for objects. (Ex: keywords or paragraphs).

→ Multimedia DBs store image, audio and video data.

Ex: picture content-based retrieval, voicemail systems, video on-demand sys's, the www & speech-based user interfaces that recognize spoken commands.

Heterogeneous DB's & Legacy DB.

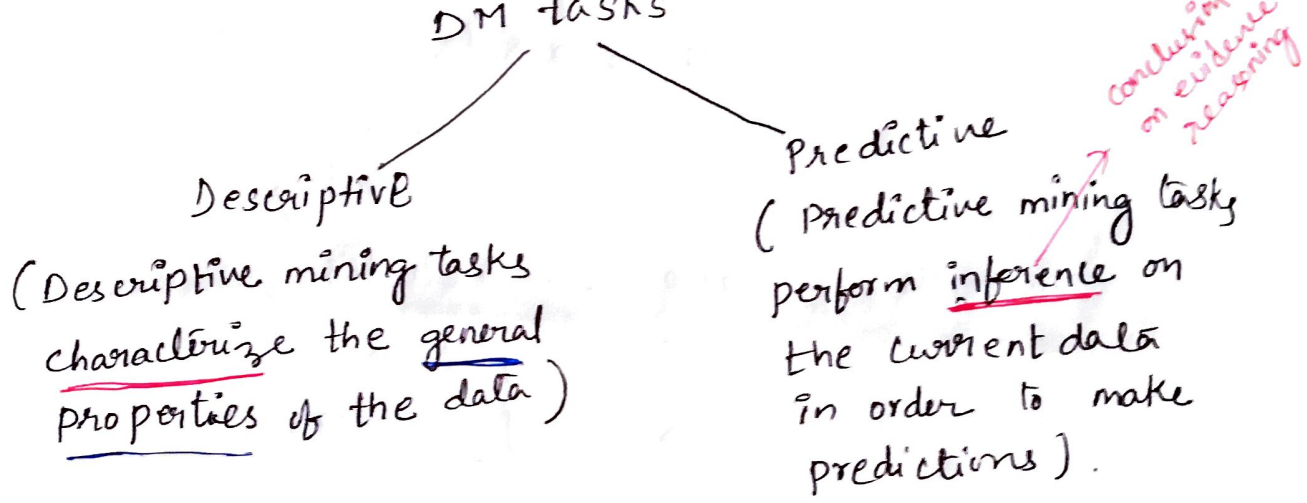
A legacy db is a group of heterogeneous db's that combines diff kinds of data systems.

The WWW:

Data Mining Functionalities - What kind of patterns can be mined

→ Data mining fun are used to specify the kind of patterns to be found in datamining tasks.

→ In general, DM tasks are classified into 2 categories



→ DM systems should be able to discover patterns at various granularities (i.e diff levels of abstraction).

DM fun and kinds of patterns they can discover are:

1. concept / class description: Characterization & Discrimination
2. Mining frequent patterns, correlations and Associations.
3. classification and prediction
4. Cluster ~~Analysis~~
5. Outlier "
6. Evolution "

(1) Concept / class description:

Characterization and Discrimination

→ Data can be associated with classes & concepts

Ex: AllElectronics store

Classes of items for sale include
- → computers
- → printers

Concepts of customers includes
- → Big spenders
- → budget spenders.

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

Ex: To study the characteristics of s/w products whose sales increased by 10% in the last year. O/P rep in pie charts, bar charts, curves, multidimensional data cubes and tables

→ Data discrimination is a comparison of the general features of target class data objects with the objects from one or a set of contrasting classes

Data characterization	Data discrimination
Generalization (common class to all customer) i.e A & B belongs to the same	comparative class i.e B is better than A

in y".

Ex: Given the AllElectronics relational database, a data mining system may find association rules like

$$age(X, \text{"20...29"}) \wedge income(X, \text{"20K..29K"}) \Rightarrow buys(X, \text{"CDplayer"}) \quad [S = 2\%, C = 60\%]$$

where X is a variable, rep a customer. The rule indicates that of the AllElectronics customers under study, 2% (support) are 20 to 29 years of age with an income of 20K to 29K and have purchased a CD player at AllElectronics.

→ there is a 60% probability (confidence, or certainty) that a customer in this age and income group will purchase a CD player.

NOTE :- That this is an association b/w more than one attribute, or predicate (i.e age, income and buys)

→ Adopting the terminology used in multidimensional databases, where each attribute is referred to as a dimension, the above rule can be referred to as a multidimensional association rule.

→ Suppose, as a marketing manager of AllElectronics, you would like to determine which items are frequently purchased together with in the same transcations.

→ An example of such a rule is

contains (T, "computer") ⇒ contains (T, "software")

[support = 1% confidence = 50%]

meaning that if a transcation, T, contains "computer", there is a 50% chance that it contains "software" as well and 1% of all of the transcation contain both.

→ This association rule involves a single attribute or predicate (i.e contains) that repeats.

→ Association rules that contain a single predicate are referred to as single-dimensional association rules.

→ Dropping the predicate notation, the above rule can be written simply as

"computer ⇒ software [1% 50%]

5) classification and prediction, is the process of finding set of models (or fun's) that describes and distinguish data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

→ the derived model is based on the analysis of a set of training data (i.e data objects whose class label is known)

→ the derived model may be rep in various form
- classification (IF-THEN) rules
- decision Trees
- mathematical formulae
- neural networks

→ A decision tree is a flowchart like tree structure where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves rep classes or class distributions.

→ Decision trees can be easily converted to classification rules.

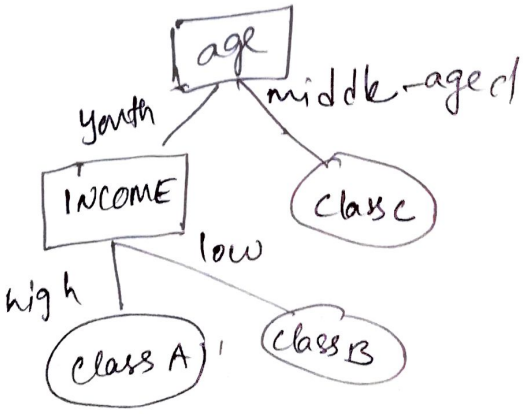
→ A neural n/w's when used for classification, is typically a collection of neuron-like processing units with weighted connections b/w the units.

→ Classification can be used for predicting the class label of data objects.

→ Users may wish to predict some missing or unavailable data values rather than class labels.

→ This is usually the case when the predicted values are numerical data and is often specifically referred to as prediction.

Ex: age (x, 29.. 35) and INCOME (x, high) → class(x



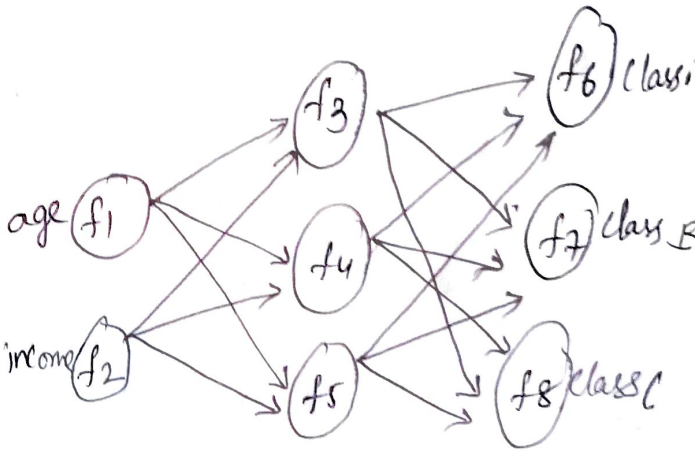


fig: A classification model can be represented in various forms such as (a) IF-THEN rules, (b) a decision tree, or a (c) neural network.

(4) Cluster Analysis: Unlike classification and prediction which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label.

→ In general, the class labels are not present in the training data simply because they are not known to begin with

→ clustering can be used to generate such labels.

→ The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.

→ clustering can also facilitate taxonomy formation that is, the organization of observations into a hierarchy of classes that group similar events together.

Ex: cluster analysis can be performed on AllElectronics customer data in order to identify homogeneous subpopulations of customers.

→ these clusters may rep individual target groups for marketing.

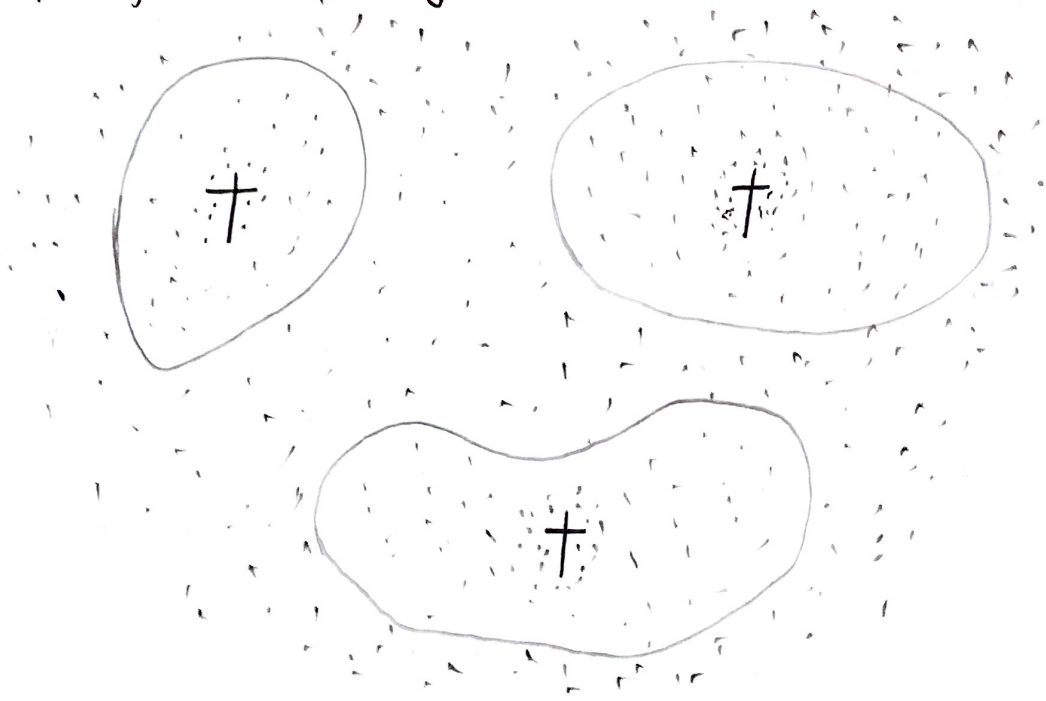


fig: A 2-D plot of customer data with respect to customer locations in a city, showing 3 data clusters. Each cluster "center" is marked with a "+"

(5) Outlier Analysis (Fraud Detection)

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers.

→ Most datamining methods discard outliers as noise or exceptions.

→ The analysis of outliers data is referred to as outlier mining.

→ outliers may be detected using statistical tests that assume a distribution or probability model for the data.

(6) Evolution Analysis

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association and correlation Analysis, classification, prediction or clustering of time related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

Classification of Datamining System

→ Because of the diversity of disciplines contributing to DM, Data mining research is expected to generate a large variety of datamining systems.

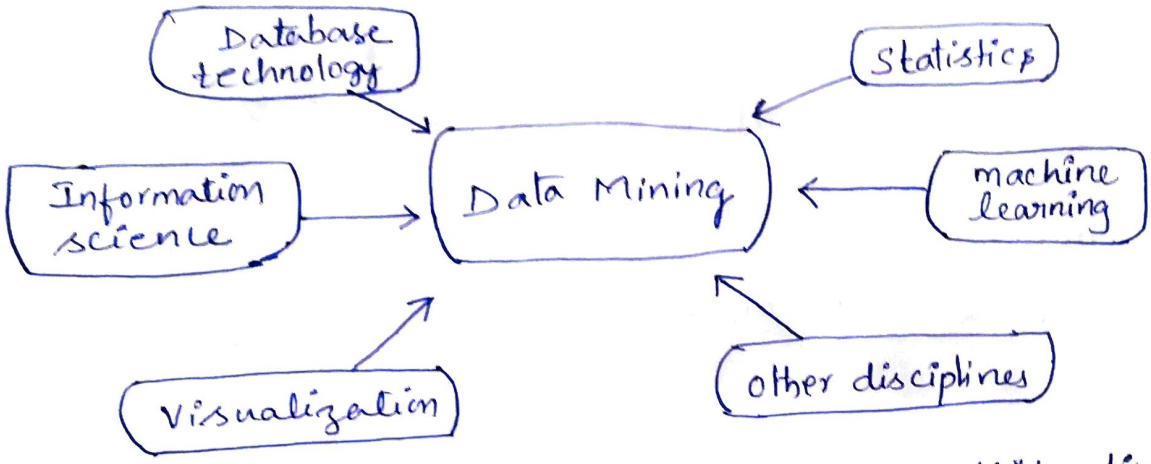


fig: Data mining as a confluence of multiple disciplines.

① classification according to the kinds of DB's mined.

if classifying according to Datamodels → Relational, transactional, object-oriented, object-relational or datawarehouse mining system.

if classifying acc. to the special types → spatial, time-series, text, or multimedia data mining system or a www mining system.

② classification acc to the kinds of knowledge mined

→ Based on DM functionalities, such as characterization, discrimination, association, classification, clustering, outlier analysis and evolution analysis.

③ Classification according to the kinds of techniques utilized

→ Database-oriented or datawarehouse-oriented techniques, machine learning, statistics, visualization, pattern recognition, neural networks and so on.

④ classification according to the applications adapted

→ Finance; telecommunications, DNA, stock markets, e-mail and so on.

Major Issues in Data Mining → challenges

① Mining Methodology and user interaction issues

→ Mining diff kinds of knowledge in DB's.

→ Interactive mining of knowledge at multiple levels of abstraction.

→ Incorporation of background knowledge (info regarding the domain under study)

→ DM query languages and adhoc data mining → SQL, DMQL

→ Presentation and visualization of DM results → easily understood by humans

→ Handling noisy or incomplete data

→ Pattern evaluation – the interestingness problem

② Performance issues → must be efficient and scalable " " predictable & acceptable in large DB's.

→ Efficiency and scalability of DM algorithms.

→ Parallel, distributed and incremental mining algorithms → the huge size of many DB's, wide distribution of data.

③ Issues relating to the diversity of database types: → diff things, types of data.

→ Handling of relational and complex

→ Mining information from heterogeneous DB's and global info systems.

Data Mining task Primitives

(1) Task-relevant data: what is the data set that I want to mine?

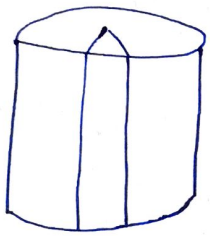
(2) What kind of knowledge do I want to mine?

(3) what background knowledge could be useful here?

(4) what measures can be used to estimate pattern interestingness?

(5) How do I want the discovered patterns to be presented?

(1) The set of task-relavant data to be mined



Task-relavant data
Database or Datawarehouse name
Database tables or data warehouse cubes
Conditions for data selection
Relevant attributes or dimensions
Data grouping criteria.

Knowledge type to be mined
characterization
Discrimination
Association
classification / Prediction
clustering

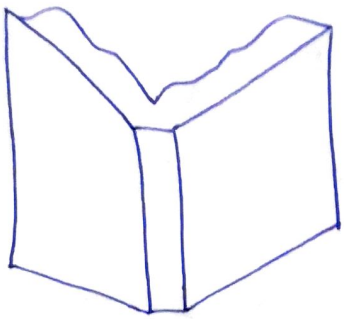
A user studying the ~~buying~~ buying habits of AllElec customers may choose to mine association rules of the form

$P(X: \text{customer}, W) \wedge Q(X, Y) \Rightarrow buys(X, Z)$

P, Q -> Predicate var's
W, Y, Z are obj var's

$age(X, "30..39") \wedge income(X, "40K..49K") \Rightarrow buys(X, "VCR")$

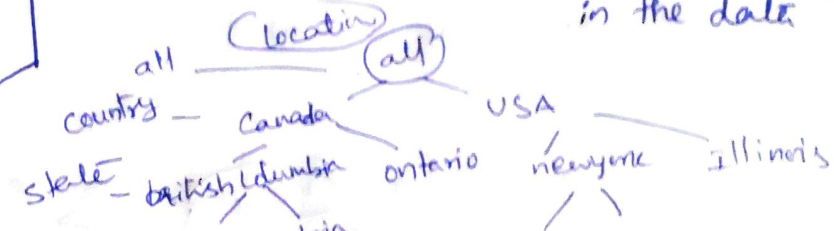
and $occupation(X, "Stud") \wedge age(X, "20..29") \Rightarrow buys(X, "comp")$ [2.2%, 60%] [1.4%, 70%]

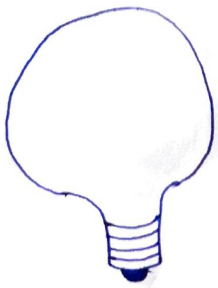


Background Knowledge

concept hierarchies → sets of mapping from a set of low-level concepts to higher-level

user beliefs about relationships in the data





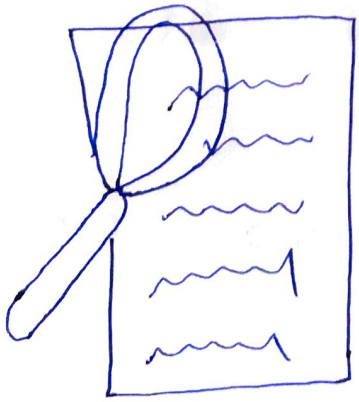
Pattern interestingness measures

simplicity

certainty (e.g., confidence)

utility (e.g., support)

Novelty [newness]



Visualization of discovered patterns

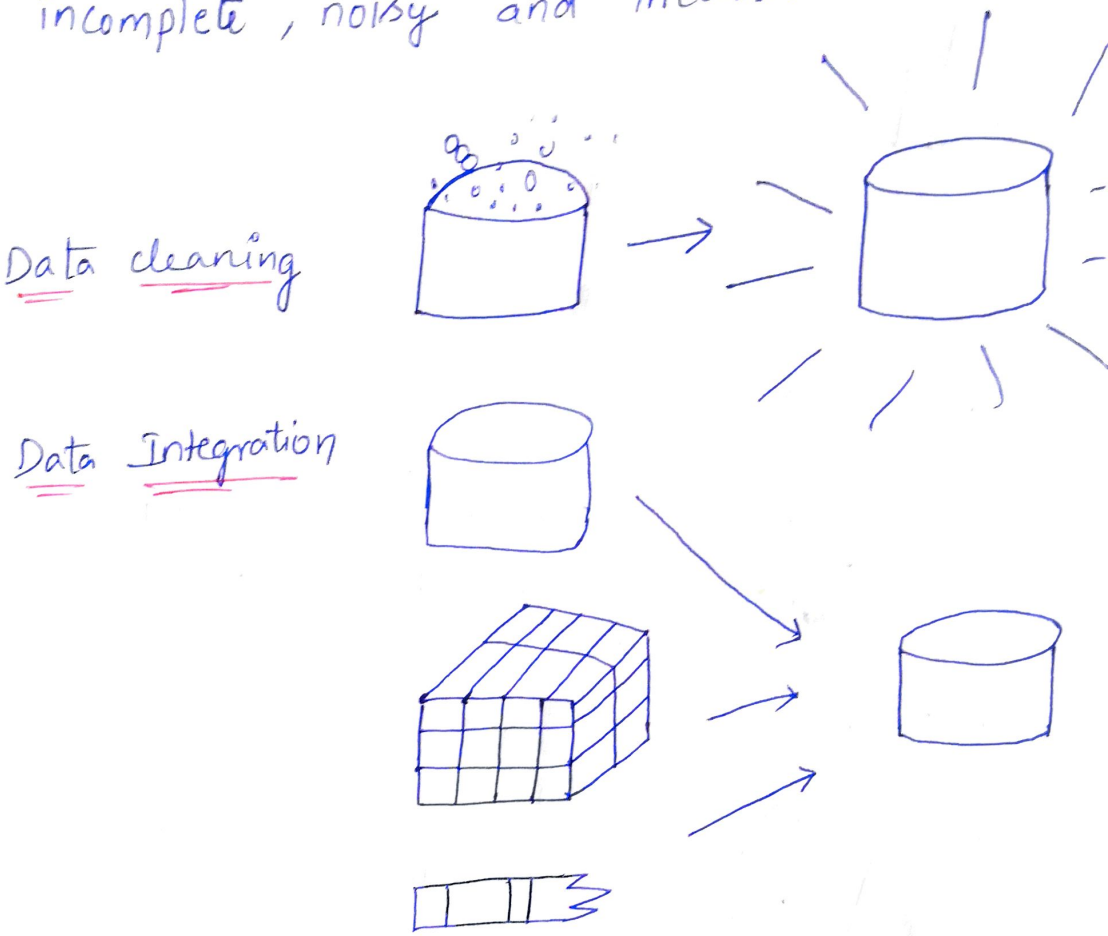
Rules, tables, reports, charts, graphs, decision trees and cubes

Drill-down and roll-up.

fig: Primitives for specifying a data mining task.

Data Preprocessing

→ Data Preprocessing is an imp issue in DW & DM as real-world data tend to be incomplete, noisy and inconsistent.



Data transformation −2, 32, 100, 59, 48 → −0.02, 0.32, 1.00,

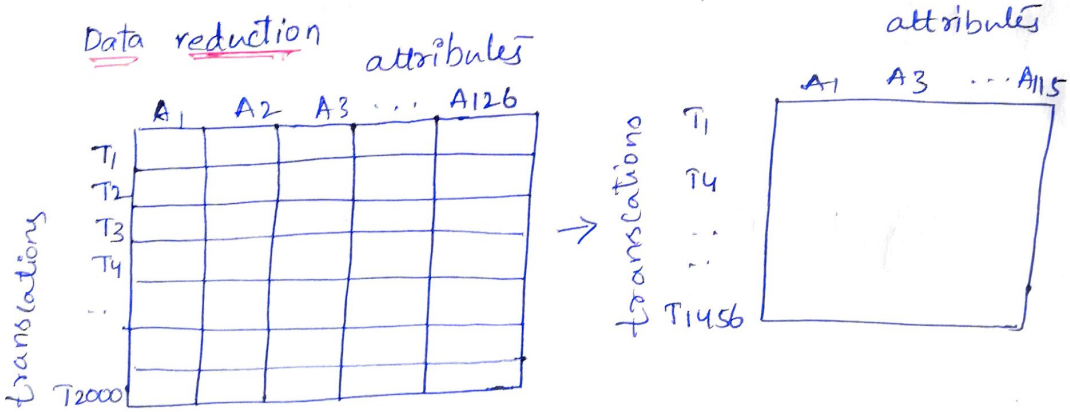


fig: Forms of data preprocessing

Descriptive Data summarization

(1) Measuring the central tendency.

The most common and most effective numerical measure of the "center" of a set of data is the (arithmetic) mean.

The mean of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

$$\text{Median} = L_1 + \left(\frac{N/2 - (\Sigma \text{ freq})l}{\text{freq}_{\text{median}}} \right) \text{width}$$

variance and standard deviation

the variance of N observations, $x_1, x_2 \ldots x_n$ is

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 = \frac{1}{N} \left[\Sigma x_i^2 - \frac{1}{N} (\Sigma x_i)^2 \right].$$

L_1 is the lower boundary of the median interval,

N " " no. of values in the entire data set,

$(\Sigma \text{ freq})l$ is the sum of the freq's of all of the intervales that are lower than the median interval,

$(\text{freq}_{\text{median}})$ is the freq of the median interval,

width is the width of the median interval.

Data cleaning

(1) Missing Values

→ Ignore the tuple :→ when class label is missing

→ Fill in the missing value manually :- Time consuming

→ Use a global constant to fill in the missing value → [unknown or ∞]

→ Use the attribute mean to fill in the missing value. Avg value filling

→ Use the attribute mean for all samples belonging to the same class as the given tuple

→ Use the most probable value to fill in the missing value
↳ determined with regression, inference based tools using a Bayesian formalism or decision tree induction

(2) Noisy Data

Noise is a random error or variance in a measured variable.

(1) Binning
(2) clustering — outliers may be detected by clustering.
(3) Combined computer and human inspection :- helps to identify outlier patterns. Ex: mislabeled characters
(4) Regression → Linear regression involves finding the "best" line to fit two variables, so that one variable can be used to predict the other.

Binning: Binning methods smooth a sorted data value by consulting its "neighborhood", that is, the values around it.

→ The sorted values are distributed into a no. of "buckets" or bins.

→ Becoz binning methods consult the neighborhood of values, they perform local smoothing.

Ex:

Sorted data for price (in dollars):

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

Partition into (equidepth) bins:

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

fig: Binning methods for data smoothing

(3) Inconsistent data: - Inconsistent in data recording for some transcations.

→ Some data inconsistencies may be corrected manually using external references.

Ex: Errors made at data entry may be corrected by performing a paper trace.

Data Integration and Transformation

→ The merging of data from multiple data stores.

→ Redundancy is an important issue.

⇒ Some redundancies can be detected by correlation Analysis.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{N}$$

Ex: A 2×2 contingency table for the data are Gender and Preferred – Reading Correlated?

	Male	female	Total
fiction	250 (90)	200 (360)	450
non-fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

$$e_{11} = \frac{300 \times 450}{1500} = 90.$$

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

$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

Issues

1. Schema integration & object matching can be tricky
2. Redundancy, can be detected by correlation analysis.
3. Data integration is the detection and resolution of data value conflicts.

Data Transformation

→ The data are transformed or consolidated into forms appropriate for mining.

1. smoothing — remove noise from the data. such techs include Binning, regression and clustering.
2. Aggregation — summary or aggregation operations are applied to the data. Ex: daily sales data may be aggregated so as to compute monthly & annual total amounts.
3. Generalization → Techs: Data cube for analysis of data at multiple granularities. Raw (primitive) data can be replaced by higher level concepts. Techs: concept hierarchies
4. Normalization → where the attribute data are scaled so as to fall within a small specified range, such as −1.0 to 1.0 or 0.0 to 1.0
5. Attribute construction. → (feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.

Min-max normalization performs a linear transformation on the original data.

→ Min-Max normalization maps a value, v of A to v' in the range $[new\_min_A, new\_max_A]$ by computing

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Suppose that the minimum & maximum values for the attribute income are \$12,000 & \$98,000, resp. we would like to map income to the range [0.0, 1.0]. By min-max normalization, a value of \$73,600 for income is transformed to.

$$\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$$

Z-score normalization or zero-mean normalization, the values for an attribute, A are normalized based on the mean and standard deviation of A. A value, V of A is normalized to V' by computing.

$$V' = \frac{V - \bar{A}}{\sigma_A}$$

where $\bar{A}$ & σ_A are the mean & SD.

Z-Score normalization suppose that the mean and standard deviation of the values for the attribute income are $54,000 & $16,000, respectively, with z-score normalization, a value of $73,600 for income is transformed to

$$\frac{73,600 - 54,000}{16,000} = 1.225$$

Data Reduction

→ Applied to obtain a reduced representation the data set, that is much smaller in volume, yet closely maintains the integrity of the original data.

(1) Data cube aggregation, where aggregation operations are applied to the data in the construction of a data cube.

(2) Attribute subset selection, where irrelavant, weakly relavant or redundant attributes or dimensions may be detected and removed.

(3) Dimensionality reduction, where encoding mechanisms are used to reduce the data set size. (compressed or encrypted without losing the identity.)

(4) Numerosity reduction, where the data are replaced or estimated by alternative, smaller data representations clustering, sampling & the use of histograms.

(5) Discretization and concept hierarchy generation is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies.

① Data cube aggregation

→ Data cubes store multidimensional aggregated information.

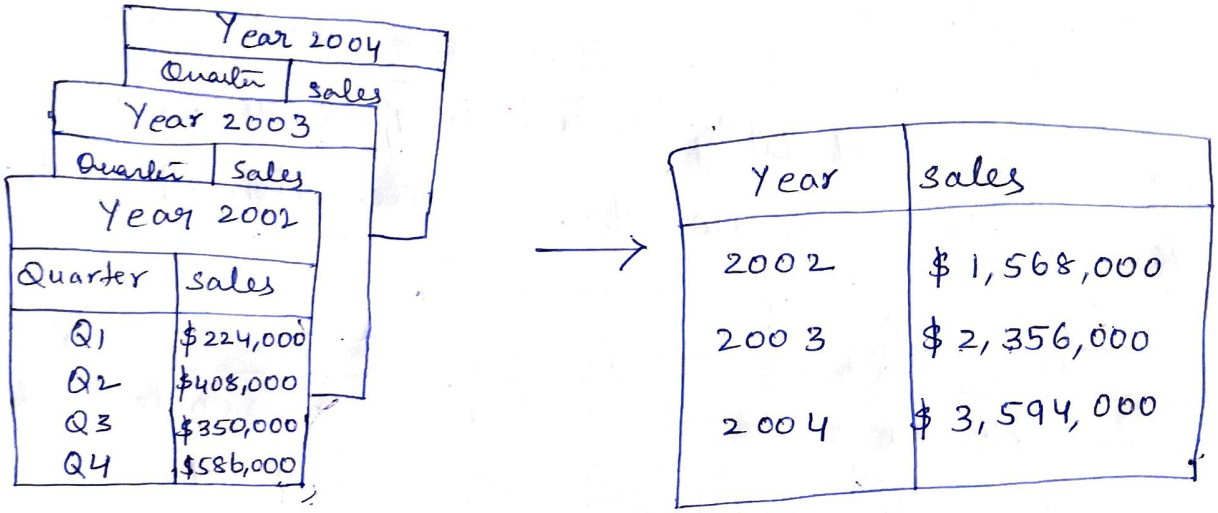


fig: AllElectronics for the years 2002 to 2004

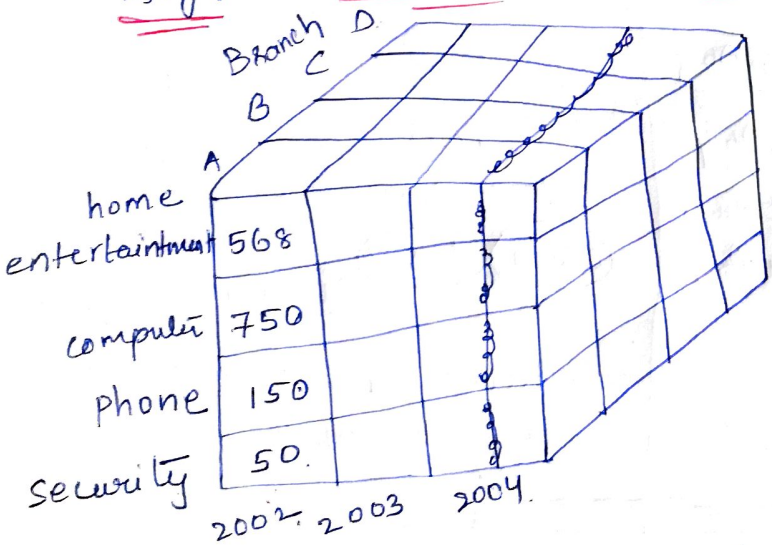


fig: A data cube for sales at AllElectronics

Attribute subset selection

→ Attribute subset selection reduces the dat
size by removing irrelevant or redundant att
or (dimensions).

→ The goal of attribute subset selection is to
a minimum set of attributes such that the
resulting probability distribution of the data clas
is as close as possible to the original distribu
obtained using all attributes.

Forward selection

Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

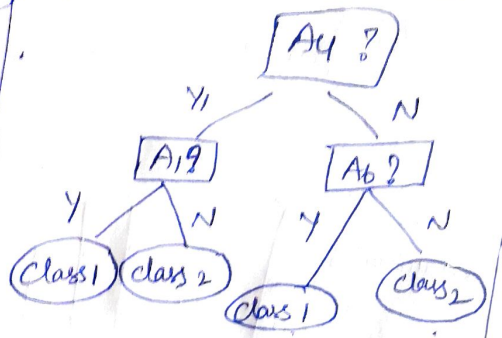
Initial reduced set:
$\{\}$
$\Rightarrow \{A_1\}$
$= \{A_1, A_4\}$
$\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$

Backward elimination

Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$
$\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$
$\Rightarrow \{A_1, A_4, A_5, A_6\}$
$\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$

Decision tree induction

Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$



$\Rightarrow$ Reduced attribute set:
$\{A_1, A_4, A_6\}$

(1) Stepwise forward selection: The procedure starts with an empty set of attributes as the reduced set.
→ The best of the original attributes is determined and added to the reduced set.
→ At each subsequent iteration or step, the best of the remaining original attributes is added to the set.

(2) stepwise backward elimination: The procedure starts with the full set of attributes. At each step, it removes the worst attribute remaining in the set.

(3) Combination of forward selection and backward elimination: The stepwise forward selection and backward elimination methods can be combined so that, at each step, the procedure selects the best attribute and removes the worst from among the remaining attributes.

(4) Decision tree induction: Decision tree algs, such as ID3, C4.5 and CART, were originally intended for classification. Decision tree induction constructs a flow chart, like structure where each internal (nonleaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external (leaf) node denotes a class prediction.

Dimensionality Reduction

→ In Dimensionality Reduction, data encoding or data transformations are applied so as to obtain a reduced or "compressed" rep 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.

→ If, instead, we can reconstruct only an approximation of the original data, then the data reduction is called lossy.

→ Two popular and effective methods of lossy dimensionality reduction:

1. wavelet transforms
2. Principal component analysis.

(1) wavelet transforms

The discrete wavelet transform (DWT) is a linear signal processing tech that, when applied to a data vector X, transforms it to a numerically diff vector x', of wavelet coefficients.

→ The two vectors are of the same length.

→ When applying this tech to data ~~do~~ reduction we consider each tuple as an n-dimensional data vector, that is $X = (x_1, x_2, \cdots x_n)$, depicting 'n' measurements made on the tuple from 'n' database attributes.

→ The usefulness lies in the fact that the wavelet transformed data can be truncated.

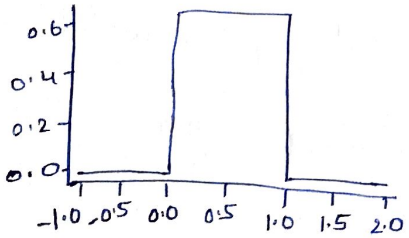
→ A compressed approximation of the data can be retained by storing only a small fraction of the strongest of the wavelet coefficients.

Ex:

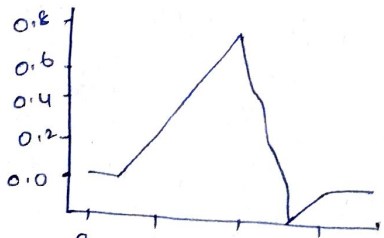
→ All wavelet coefficients larger than some user-specified threshold can be retained.

→ All other coefficients are set to 0.

→ The resulting data representation is therefore very sparse, so that operations that can take adv of data sparsity are computationally very fast if performed in wavelet space.



(a) Haar-2



(b) Daubechies-4.

Principal components Analysis

→ PCA as a method of dimensionality reduction.

Suppose that the data to be reduced consists of tuples or data vectors described by 'n' attributes or dimensions.

→ PCA also called Karhunen-Loeve, or K-L method), searches for 'K' n-dimensional orthogonal vectors that can best be used to represent the data, where $K \leq n$.

→ the original data are thus projected onto a much smaller space, resulting in dimensionality reduction.

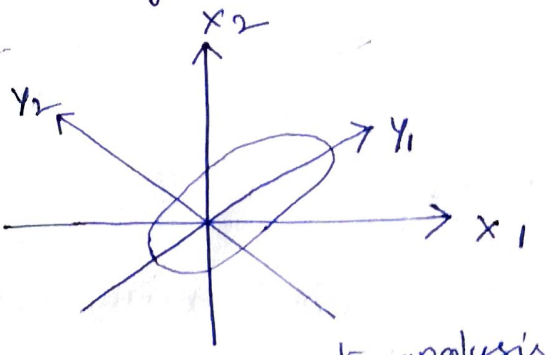


fig: principal components analysis Y_1 & Y_2 are the first two principal components for the given data.

Basic Procedure

① The i/p data are normalized, so that each attribute falls within the same range. This ~~ensure~~ step helps ensure that attributes with large domains will do not dominate attributes with small domains.

② PCA computes K orthonormal vectors that provide a basis for the normalized i/p data. These are unit vectors that each point in a direction perpendicular to the others.

These vectors are referred to as the Principal components. The i/p data are a linear combination of the principal components.

(3) The principal components are sorted in order of decreasing "significance" or strength. The principal components essentially serve as a new set of axes for the data, providing imp info about variance. That is, the sorted axes are such that the first axis shows the most variance among the data, the second axis shows the next highest variance and so on.

For ex: in the above fig show the first two principal components Y_1 and Y_2 for the given set of data originally mapped to the axes X_1 and X_2. This info helps identify groups or patterns within the data.

(4) Because the components are sorted according to decreasing order of "significance", the size of the data can be reduced by eliminating the weaker components, that is those with low variance.

Numerosity Reduction

Techniques of numerosity reduction are used to reduce the data volume by choosing alternative, "smaller" forms of data representation.
These techniques may be parametric or nonparametric.

(1) Parametric methods is used to estimate the data, so that typically only the data meters need to be stored; outliers may also be stred.
Ex: Log-linear models.
which estimate discrete multidimensional probability distributions, are an example.

(2) Non-parametric methods for storing reduced representations of the data include histograms, clustering and sampling.

Regression and log-linear models
→ Used to approximate the given data.
→ In linear regression, the data are modeled to fit a straight line.
Ex: A random variable, y (called a response variable), can be modeled as a linear fun of another random variable, x (called a predictor variable), with the eqn

$$y = wx + b$$

where the variance of y is assumed to be constant. In the context of datamining, x and y are numerical database attributes.
The coefficients, w and b (called regression coefficients).

specify the slope of the line and the Y-intercept, respectively.

→ These coefficients can be solved for by the method of least squares, which minimizes the error b/w the actual line separating the data and the estimate of the line.

→ Multiple linear regression is an extension of (simple) linear regression, which allows a response variable y, to be modeled as a linear fun of two or more predictor variables.

→ Log-linear models approximate discrete multidimensional probability distributions.

→ Given a set of tuples in n dimensions (e.g. described by n attributes), we can consider each tuple as a point in an n-dimensional space.

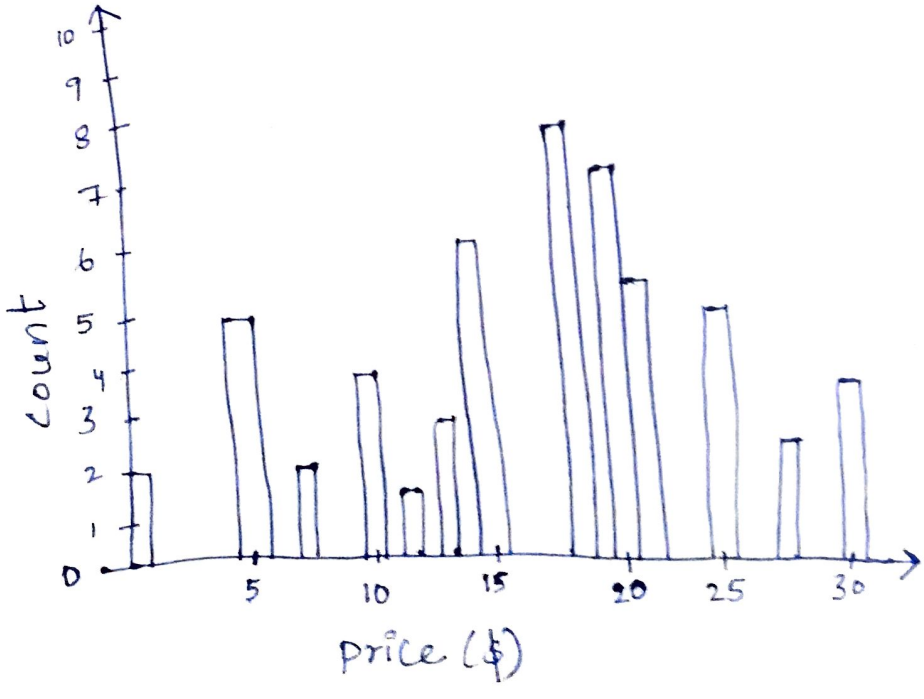
Histograms

→ Histograms use binning to approximate data distributions and are a popular form of data reduction.

→ A histogram for an attribute, A, partitions the data distribution of A into disjoint subsets, or buckets.

→ If each bucket represents only a single attribute-value / frequency pair, the buckets are called singleton buckets.

Ex:- The following data are a list of prices of commonly sold items at AllElectronics (rounded to the nearest dollar). The numbers 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, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.



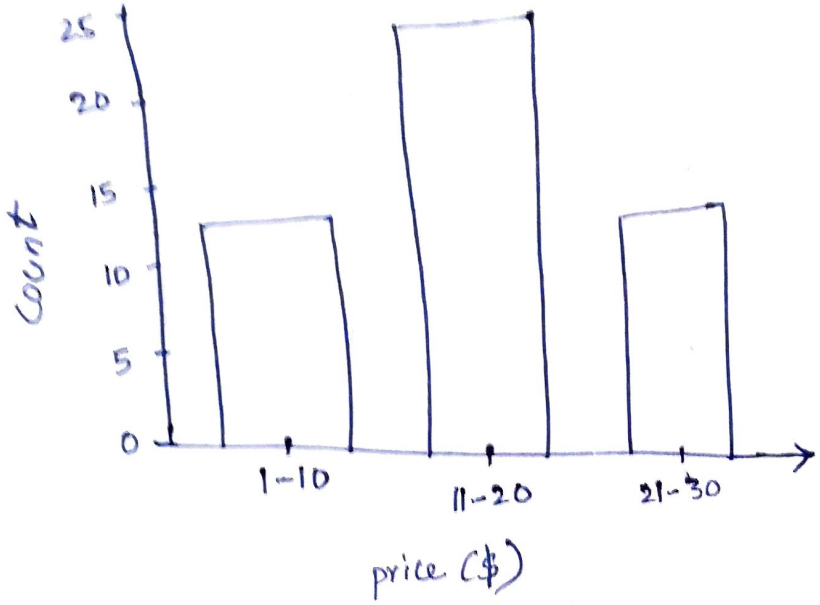


fig: An equal-width histogram for price, where values are aggregated so that each bucket has a uniform width of $10.

"How are the buckets determined and the attribute values partitioned?"

Partitioning rules:-

(1) Equal-width :- The width of each bucket range is uniform.

(2) Equal-frequency :- The buckets are created so that, roughly, the freq of each bucket is constant.

(3) V-Optimal :- If we consider all of the possible histograms for a given no. of buckets, the V-optimal histogram is the one with the least variance.

(4) MaxDiff :- We consider the diff b/w each pair of adjacent values. A bucket boundary is established b/w each pair for pairs having the β-1 largest differences, where β is the user-specified no. of buckets.

Clustering :- Clustering techniques consider data tuples as objects. They partition the objects to groups or clusters, so that objects within a cluster are "similar" to one another and "dissimilar" objects in other clusters.

⇒ The "quality" of a cluster may be represented by its diameter, the maximum distance b/w any two objects in the cluster.

→ centroid distance is an alternative measure of cluster quality and is defined as the avg distance of each cluster object from the cluster centroid.

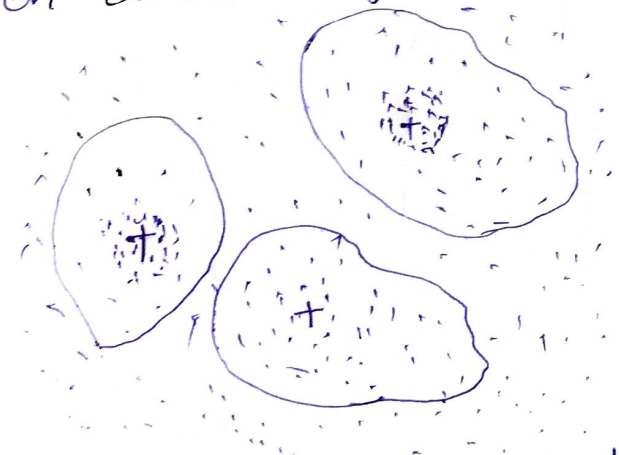


fig: A 2-D plot of customer data

→ A concept hierarchy for a given numerical attribute defines a discretization of the attribute.

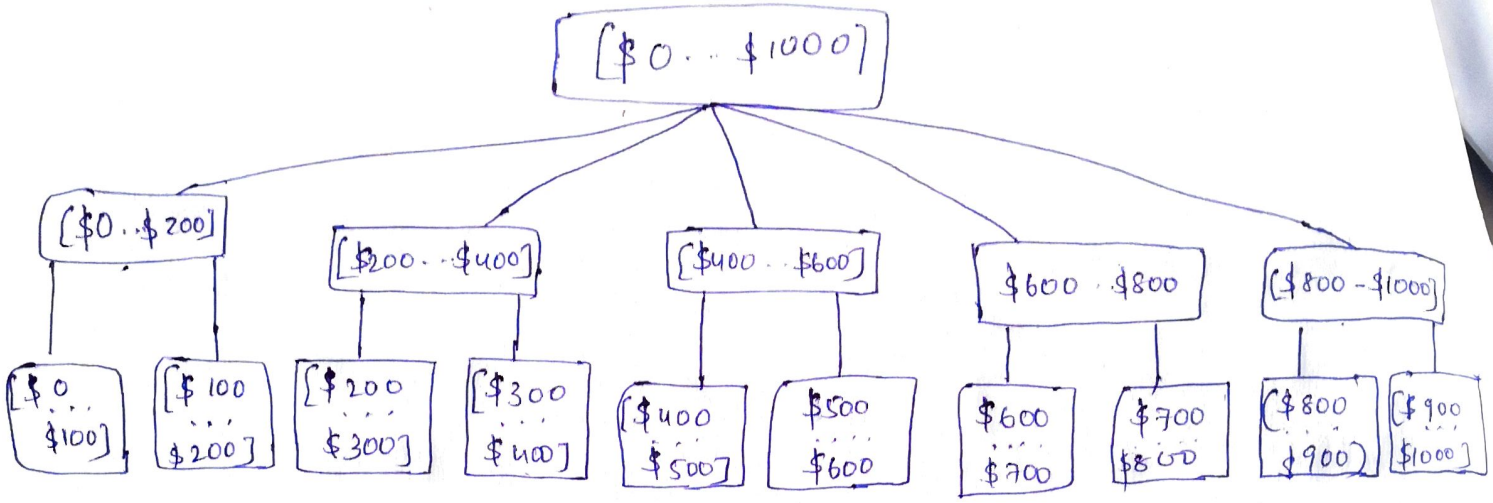


fig: A concept hierarchy for the attribute price, where an interval ($x...$y) denotes the range from $X (exclusive) to $Y (inclusive).

→ concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute age) with higher-level concepts (such as youth, middle-aged or senior).

Smoothing by Binning

Using the data for age

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 26, 25,
30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

Sol bin depth of 3

① sort the data

$N = 27$, $D = 3$ (depth)

② $\frac{N}{D} = \frac{27}{3} = 9$

③ mean or median (50th percentile) or boundaries (min & max val)

$$\text{mean} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$= \frac{1}{27}$$

Smoothing by bin means.

step 1: Sorted, by incrementing

step 2: $\frac{N}{D} = \frac{27}{3} = 9$ bins

Bin 1: 13, 15, 16, 16, 19, 20

Bin 2: 16, 19, 20

Bin 3: 20, 22, 22

" 4: 22, 25, 25

" 5: 26, 25, 30

" 6: 33, 33, 35

" 7: 35, 35, 35

" 8: 36, 40, 45

" 9: 46, 52, 70

step 3

Bin 1: $\frac{1}{3}(44) = 14.667 = 14\frac{2}{3}$

$B_2 = \frac{1}{3}(55) = 18.33 = 18\frac{1}{3}$

$B_3 = \frac{1}{3}(63) = 21$

$B_4 = \frac{1}{3}(72) = 24$

$B_5 = \frac{1}{3}(81) =$

$B_6 = \frac{1}{3}(101) = 33.667 = 33\frac{2}{3}$

$B_7 = \frac{1}{3}(105) = 35$

$B_8 = \frac{1}{3}(121) = 40.33 = 40\frac{1}{3}$

$B_9 = \frac{1}{3}(168) = 56$

step 4: Bin 1 = $14\frac{2}{3}, 14\frac{2}{3}$.

Ex:

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34.

Sol partition into three equal-depth bins

Step 1: Bins $= \frac{N}{4} = \frac{12}{4} = 3$

Bin 1 = 4, 8, 9, 15 $= \frac{36}{4} = 9$

Bin 2 = 21, 21, 24, 25 $= \frac{91}{4} = 22.75 = 23$

Bin 3 = 26, 28, 29, 34 $= \frac{117}{4} = 29.25 = 29$

Step 2: smoothing by bin means

Bin 1 : 9, 9, 9, 9

Bin 2 : 23, 23, 23, 23

Bin 3 : 29, 29, 29, 29

Smoothing by bin boundaries

Step 3 :

Bin 1 : 4, 4, 4, 15

Bin 2 : 21, 21, 25, 25

Bin 3 : 26, 26, 26, 34

bin 1 = 4, 8, 15

bin 1 = 4, 4, 15

Ex: Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows.

age	freq
1-5	200
5-15	450
15-20	300
20-50	1500
50-80	700
80-110	44

$N = 3194$

lower $(\Sigma \text{freq}) = 950$

median value $= 950$

width = width of median interval

$20 - 50 = 30$.

Compute an approximate median value for the data.

Sol

$$\text{median} = L_1 + \left(\frac{N/2 - (\Sigma \text{ freq})_l}{\text{freq}_{\text{median}}} \right) \text{width}$$

we have

$L_1 = 20$ lower boundary

$N = 3194$

$(\Sigma \text{ freq})_l = 950$

$\text{freq}_{\text{median}} = 1500$

$\text{width} = 30$

median =

1500 ⇒ 20-50

$$\text{median} = 20 + \left(\frac{\frac{3194}{2} - 950}{1500} \right) 30$$

$$= 20 + (0.431)\,30 = 20 + 12.94$$

$$= 32.94 \text{ years}.$$

Ex: Min-max normalization.

Suppose that the min and max values for the attribute income are $12000 and $98000, resp. we would like to map income to the range [0.0, 1.0]. By min-max normalization, a value of $73,600 for income is transformed to

$$v' = \frac{73,600 - 12000}{98000 - 12,000}(1.0 - 0) + 0$$

$$= 0.716$$

$v = 73,600$

$min_A = 12000$

$max_A = 98000$

$$\frac{v - min_A}{max_A - min_A}(new\,max_A - new\,min_A) + new\,min_A$$

Ex: Z-score normalization. Suppose that the mean and standard deviation of the values for the attribute income are $54000 and $16,000 resp. with Z-score normalization, a value of $73,600 for income is transformed to.

$$v' = \frac{v - \bar{A}}{\sigma A} = \frac{73600 - 54000}{16000} = 1.225.$$

Ex: suppose that the data for analysis includes the attribute age. The age value for the data tuples are (in increasing order)

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 30, 33, 33,

35, 35, 35, 35, 36, 40, 45, 46, 52, 70

(a) use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps comment on the effect of this technique for the given data.

(b) How might you determine outliers in the data?

(c) what other methods are there for data smoothing?

Sol Step 1: sort the data

step 2: partition the data into equal-freq bins of size 3.

Bin1: 13, 15, 16
Bin2: 16, 19, 20
Bin3: 20, 21, 22
Bin4: 22, 25, 25
Bin5: 25, 25, 30
Bin6: 33, 33, 35
Bin7: 35, 35, 35
Bin8: 36, 40, 45
Bin9: 46, 52, 70.

step 3: Calculate the arth. mean.

Discretization and concept Hierarchy generation for numerical data

Methods :-
1. Binning
2. Histogram analysis
3. Entropy-based discretization
4. χ^2-merging
5. Cluster analysis
6. discretization by intuitive partitioning.

Binning : Is a top-down splitting tech based on a specified no. of bins.

→ Binning methods are used for data smoothing.

Ex: attribute values can be discretized by applying equal-width or equal-freq binning, and then replacing each bin value by the bin mean or median, as in smoothing by bin means or smoothing by medians, resp.

Histogram Analysis

→ Like binning, histogram analysis is an unsupervised discretization tech bcoz it does not use class info.

→ Histograms partition the values for an attribute A, into disjoint ranges called **buckets**.

Entropy-Based Discretization

→ Entropy-based discretization is a supervised, top-down splitting tech.

→ To discretize a numerical attribute, A, the method selects the value of A that has the minimum entropy as a split-point, and recursively partitions the resulting intervals to arrive at a hierarchical discretization.

Interval Merging by x^2 Analysis

→ which employs a bottom-up approach by finding the best neighboring intervals and then merging these to form larger intervals, recursively. chimerge proceeds as follows. Initially, each distinct value of a numerical attribute A is considered to be one interval.

→ x^2 tests are performed for every pair of adjacent intervals.

→ Adjacent intervals with the least x^2 values are merged together, because low x^2 values for a pair indicate similar class distributions.

→ This merging process proceeds recursively until a predefined stopping criterion is met.

Cluster Analysis

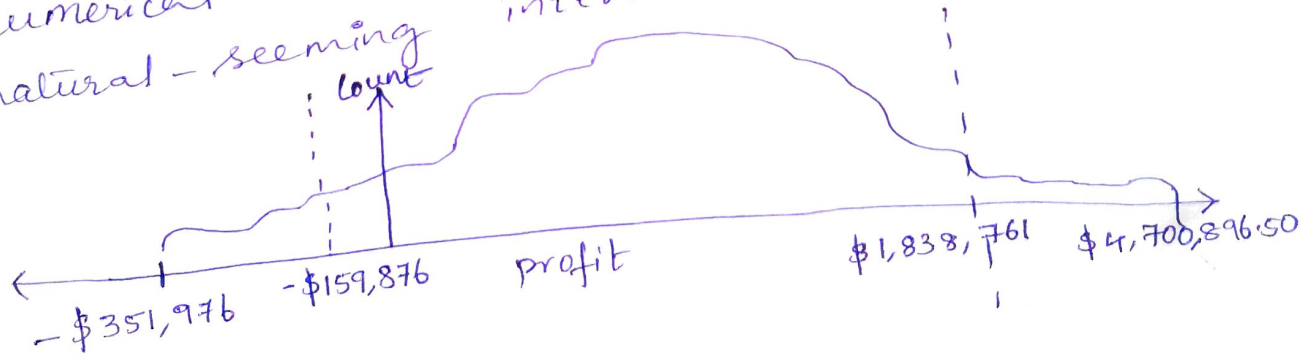
→ A clustering alg can be applied to discretize a numerical attribute, A, by partitioning the values of A into clusters or groups.

Discretization by Intuitive partitioning

→ Above discretization methods are useful in the generation of numerical hierarchies, many users would like to see numerical ranges partitioned into relatively uniform, easy-to-read intervals that appear intuitive or "natural".

Ex: Annual salaries broken into ranges like ($50,000, $60,000) are often more desirable than ranges like ($51,263.98, $60,872.34), obtained by, say, some sophisticated clustering analysis.

→ The 3-4-5 rule can be used to segment numerical data into relatively uniform, natural-seeming intervals.



Q: Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result.

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) calculate the mean, median and standard deviation of age and % fat.

(b) Draw the boxplots for age and % fat

(d) Normalize the two variables based on z score normalization.

Sol

(a) calculate the mean, median & standard deviation of age and % fat.

For the variable age the mean is 46.44

the median is 51

SD = 12.85