Introduction to Data Mining

Introduction, challenges, Data mining tasks, Types of data Data Preprocessing, Measures of Similarity and Dissimilarity: Similarity measures for binary data Jaccard co-efficient & cosine similarity.

Data Mining:

Data mining is the process of automatically discovering useful information in large data repositaries (or)

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.

Data Warehouse:-

Data warehouse is a repositary of information collected from multiple sources, stored under a unified schema, and usually residing at a single sight (source).

DataBase:-

A database is a collection of related data with an implicit meaning stored together to serve multiple application.

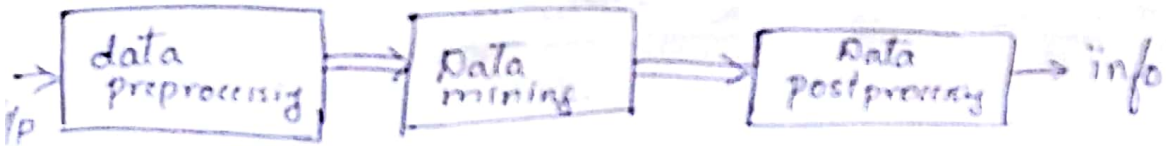
The overall purpose of database system is to record and maintain information.

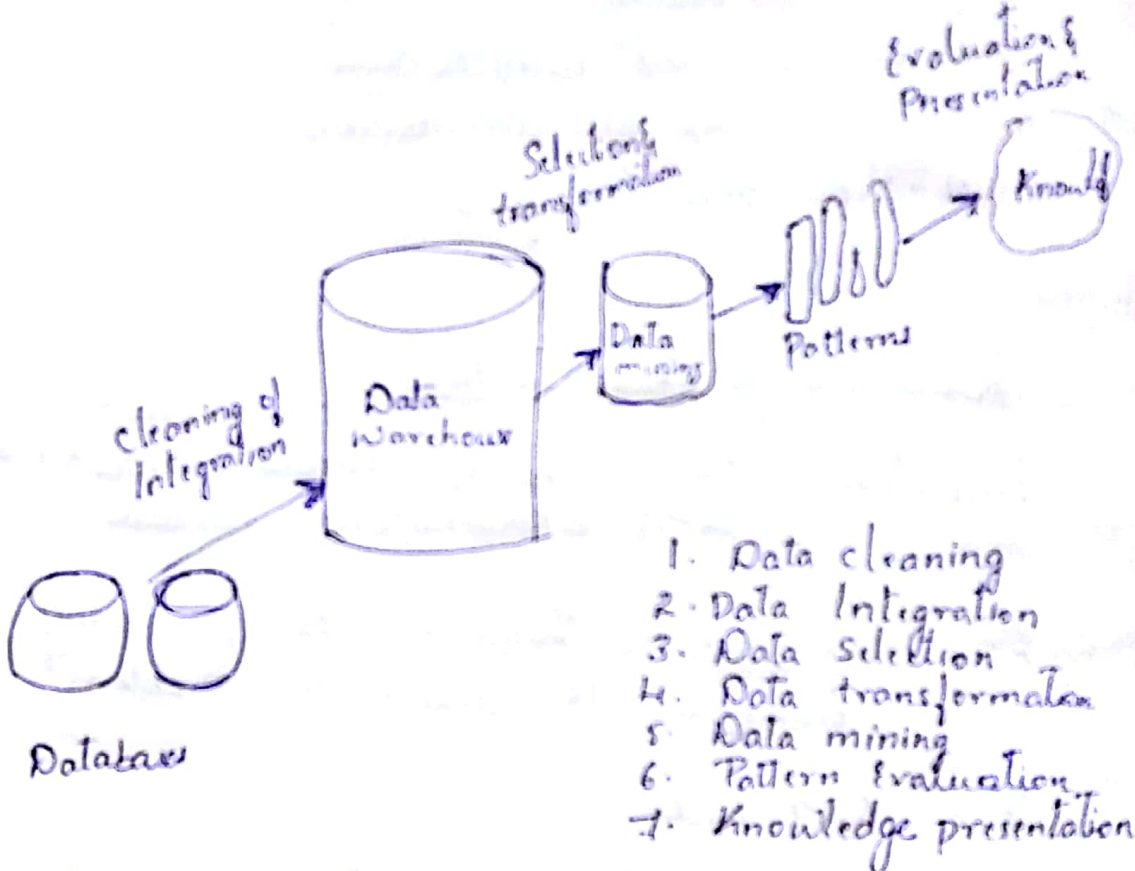
(Processed data - information)

The Purpose of data processing is to generate the information required to carry out the buisness activities.

(Data means - Raw information)

Knowledge data discovery (KDD)





1. Data cleaning
2. Data Integration
3. Data selection
4. Data transformation
5. Data mining
6. Pattern Evaluation
7. Knowledge presentation

1. Data cleaning :- To remove noise and inconsistant data
2. Data integration :- Where multiple data sources may be combined.
3. Data selection : Where data relevant to analysis task are retrived from the data base.
4. Data transformation :- Data are transformed and consolidated in forms appropriate for mining by performing some set of operation.
5. Data mining : Essential proces where intelligent method are applied To extract information.

6. Pattern evalulation: To identify the truly interesting patterns representing knowledge based on interesting measures methods.

7. Knowledge presentation: Where visualization and knowledge representation techniques are used to present mined knowledge to users.

22/1/18

Motivating Challenges of Data Mining

1. Scalability
2. High Dimensionality.
3. Heterogenous and complex data
4. Data ownership and distribution.
5. Non-traditional Analysis.

24/1/18

Data mining ~~Cause~~ Tasks

1. Predictive tasks. - finding targets (dependent) reliable by using Explanatory variable.

2. Descriptive task - Requires post processing technique to deduce the problem

Core DM tasks.

1. Predictive modelling.
   - Regression - continuous target (independent) variable
   - Classifications - discrete target variables
     - It is binary valued
2. Cluster Analysis - grouping of closely related
   - used to group set of related customers
3. Association analysis - used to decide patterns that describe strongly associated features in the data

4. Anomaly detection:- Significantly different from the rest of data.

Types of Data

Association Analysis can be applied to find items that are frequently brought together by customers.
ex:- bread-jam.
{coffee} → {sugar, milk}

Types of data

1. Data Set :- A data set refers to a collection of data objects and their attributes.
Other names for data object are record, transaction, vector, event, entity, sample, observation.

29/1/18

Attribute :- An attribute is a characteristic of an object that may vary either from one object to another object or from one time to another.

Properties of an attribute which classifies attributes

1. Distinctness : = ≠
2. Order : < >
3. Addition : + −
4. Multiplication : × \

i Nominal attributes → only distinctness. Ex:- Id, number, name
ii Ordinal attributes → Distinctness & order. Grade {FL, SC, FL, FC}
iii Interval attributes → Distinctness, Order, addition
Ex:- calendar & temper.
iv Ratio attributes → All four ex:- length, time, counts

Categorial / Qualitative Attribute

→ Nominal attribute
→ Ordinal attribute

Quantitative / numeric attributes

→ Interval attributes
→ Ratio attributes

31/1/18

Types of Data Sets

1. Record data
   → Transaction (Market based data)
   → Data matrix
   → Document data, sparse data matrix
2. Graph data
   → Data with relationship among objects.
   → Data with objects that are graph.
3. Ordered data
   → Sequential data
   → Sequence data
   → Time Series data
   → Spatial data

Ordered data

→ Collection of records.
Each record consist of set of attributes

Record Data

i) Transaction data
Each transaction consists of set of items.
Ex: Grocery store

ii) Data matrix
m x n
m → row (object)
n → column (attribute)

iii) Sparse data matrix
only non zero values are important

GRAPH based data

1) → data objects are mapped to the objects
→ relationships among the objects are captured by link properties such as direction & weight.

2) → Objects are reported as graphs.
Ex:- Molecular Structures.
→ Nodes are atoms
→ link b/w the nodes are chemical bonds

Ordered data

1 Sequential data
→ Extension of the record data
→ Each record has a time associated with it
Ex:- purchase history of the customer.

2 Sequence data
→ Individual entities such as sequence of letter & words.
Ex:- Genetic information

3 Time Series data
→ Measurement are taken over a time

4 Spatial data
→ attributes such as positions & area

1/2/18

Methods of Data Preprocessing Strategies and techniques.

Data preprocessing

1) Data Aggregation
2) Sampling
3) Dimensionality reduction
4) Feature Subset selection
5) Discretization and Binarization
6) Variable transformation

Aggregation

Combines Two or more attributes (object) into single attribute (object)

Motivations for aggregation

1) Data reduction
2) Provides the high level view of the data instead of low level view
3) Selecting the group of objects for decisions.

Sampling

Selecting a Subset of the data-objects to be analysed

Sampling Methods

1) Simple random Sampling
   - Sampling without replacement
   - Sampling with replacement

2) Stratified Sampling

- equal no of objects are drawn from each group irrespective of the size
- Samples are selected from each group is proportional to the size of the group.

~~3) Progressive Sampling~~

- Sample size is different to determine difficult in above methods.

3) Progressive Sampling.

Starts with small sample and than increase the Sample size until a sample of sufficient size has been obtained.

2/2/18

Dimensionality Reduction

DM algorithm work better when dimensionality is lower

Purpose:

- Eliminating irrelevant features (noise)
- Can lead to a more understandable model.
- Reduce amount of time and memory required by DM algorithm.
- avoide curse of dimensionality.

Data analysis becomes significantly harder as the dimensionality of the data increases.

* Feature Subset Selection

Use only a subset of feature

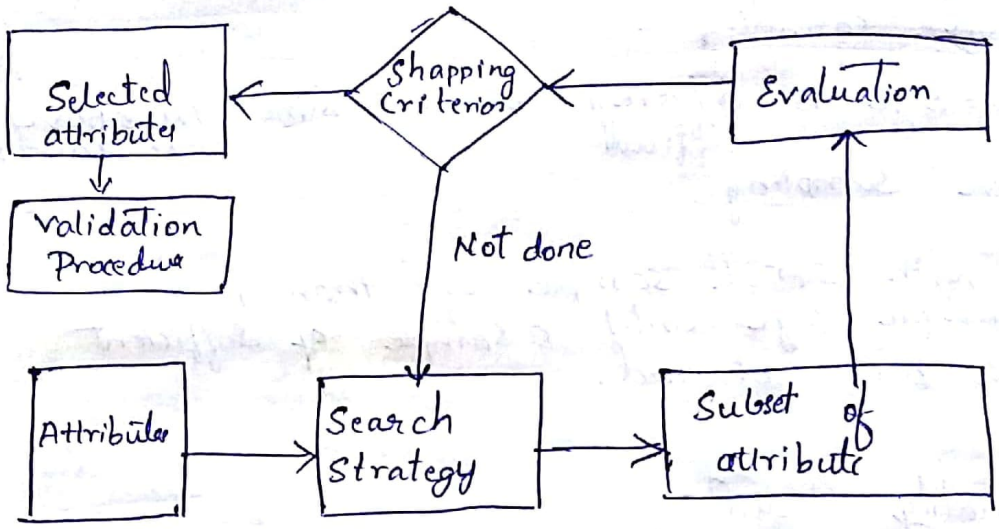
1. Redundant feature
2. Irrelevant features.

3. Techniques for feature selection.

1. Embedded approaches (F.S occur as a part of DM algorithm)
2. filter approaches → (feature on selected before DM algorithm is run)

3) Wrapper approach → Use black box to find the subset features

Architecture of FSS



Flowchart of feature subset selection process

Variable transformation

Converting one form of data into another form.

(i) Simple function.

(ii) Normalization or Standardization

Discretization and Binarization

Some DM algorithm requires that the data to be in the form of categorial attributes or Nominal (order (qualitative) attributes.

Transforming continuous attributes into a categorial attributes is called Discretization.

Ex:- Transforming continuous and Discrete attribute into Binary attributes is called Binarization.

7/2/18

* Measure of Similarity(s) and Dissimilarity(d)

* Proximity is used to refer to either S or D

Similarities are usually non-negative

0 (no Similarity) & 1 (complete similarity)

The term distance is used as a synonym for dissimilarity.

$I_1 = I_2$

$x = y$ (1)

$x \neq y$ (0)

Distances

The Euclidean distance d between 2 ports x and y is given by

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

Where, x_i → no of dimensions

x_k & y_k are the k^{th} attribute of x and y

Euclidean distance generalised by the ~~Manhatta~~ Minkowski distance

$$d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k| \right)^{\frac{1}{r}}$$

Where

r → parameter

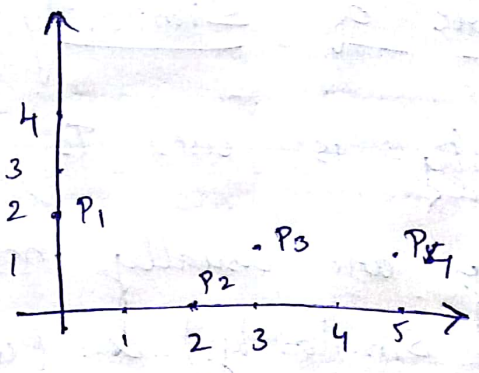
r = 1 city block (Manhattan L_1 norm) distance.

r = 2 Euclidean distance (L_2 norm)

r = ∞ Supremum (L_∞ or L_{max}) norm)

Part	x coordinate	y coordinate
P_1	0	2
P_2	2	0
P_3	3	1
P_4	5	1

x & y co-ordinate of



	P_1	P_2	P_3	P_4
P_1	0.0	2.8	3.2	5.1
P_2	2.7	0.0	1.4	3.2
P_3	3.2	1.4	00	2.0
P_4	5.1	3.0	2.0	0.0

Euclidean distance matrix

L_1	P_1	P_2	P_3	P_4
P_1	0.0	4.0	4.0	6.0
P_2	4.0	00	2.0	4.0
P_3	4.0	2.0	0.0	2.0
P_4	6.0	4.0	2.0	0.0

L_1 distance matrix

Euclidean - Method

$$(P_1 P_2) = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$$

$$= \sqrt{4+4}$$

$$= 2.8$$

Minkawhan Method

$$(P_1, P_2) = |x_1-x_2| + |y_1-y_2|$$

$$= |0-2| + |2-0|$$

$$= 4$$

$$=$$

8/2/18

10 marks

* Similarities between data objects

If $s(x,y)$ is the similarity between points than tipical properties of the similarity are the following.

1. $s(x,y) = 1$ only if $x = y$
2. $s(x,y) = s(y,x)$ for all x & y (Symmetry)

Examples of Proximity measures.

1. Similarity measures between object that contain only binary values. (attributes) are called Similarity co-efficients. We have values between 0 and 1.
2. let x and y be two objects that consist of n binary attributes.
3. Comparing of 2 objects i.e. 2 binary vectors leads to following four quantities (frequencies)

f_{00} = The no of attributes where $x = 0, y = 0$
f_{01} = The no of attributes where $x = 0, y = 1$.
f_{10} = The no of attributes where $x = 1, y = 0$.
f_{11} = The no of attributes where $x = 1, y = 1$.

Simple Matching Co-efficients (SMC)

Commonly used Similarity co-efficient defined a

$$SMC = \frac{\text{No matching co-efficient value}}{\text{Number of attributes}} =$$

Jaccard co-efficient

It is frequently used to handle objects

consisting of assymetric binary attribute

$$J = \frac{\text{No of matching presences}}{\text{No of attributes not involved in 00 matches}}$$

1. Consider the following two binary vectors. $x = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$

$y = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)$

find the SMC and Jaccard co-efficient.

Soln:

$f_{00} = 7$

$f_{01} = 2$

$f_{10} = 1$

$f_{11} = 0$

$$SMC = \frac{f_{00} + f_{01} + f_{10} + f_{11}}{f_{00} + f_{11} + f_{01} + f_{10}} = \frac{7}{7+2+1+0} = \frac{7}{10} = 0.7$$

$$J = \frac{f_{01} + f_{10} + f_{11}}{f_{11} + f_{01} + f_{10}} = \frac{2+1+0}{2+1+0} = 0$$

Cosine

Cosine Similarity

Documents are often represented as vectors, where each attribute represents the frequency with which a particular term or word occurs in a document.

→ Different forms of the same word.
→ Diffent document length.
→ Different word frequencies.

* A cosine similarity is one of the most common measure of document similarity

If x and y are two document vectors than $\cos(x,y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$

Where ; . indicates vector . product

i.e $x \cdot y = \sum_{n=1}^{k} x_k y_k$ and

$\|x\|$ is the length of vector x

$$\|x\| = \sqrt{\sum_{n=1}^{k} x_k^2} = \sqrt{x \cdot x}$$

$$\|y\| = \sqrt{\sum_{n=1}^{k} y_k^2} = \sqrt{y \cdot y}$$

1. Calculate the document similarity for the following 2 object, which might represent document vectors by using cosine similarity.

$x = (3, 2, 0, 5, 0, 0, 0, 2, 0)$

$y = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$

Soln

$$x \cdot y = (3)(1) + (2)(0) + (0)(0) + (5)(0) + (0)(0) + (0)(0) + (6)(1) + (2)(1) + (0)(0) + (0)(2)$$

$$= 3 + 2$$

$$= 5$$

$$\|x\| = \sqrt{3^2 + 2^2 + 5^2 + 2^2}$$

$$= \sqrt{9 + 4 + 25 + 4}$$

$$= \sqrt{42}$$

$$= 6.48$$

$$\|w\| = \sqrt{1 + 1 + 4}$$

$$= \sqrt{6}$$

$$= 2.44$$

$$= \frac{5}{(6.48 * 2.44)}$$

$$= 0.31$$

If the cosine similarity is 1 the angle b/w x and y is 0.

If cosise similarity is 0 than the angle b/w x and y is 90.