Data warehouse & mining

Data mining - Process of extracting information to identify patterns trends and useful data. used by organisation to solve business prob.

Type of data - - Relational , Data warehouse.
- Data repository
- Transactional data.

Application - Sales analysis - Healthcare - finance
medical analysis - education -
credit card fraud.

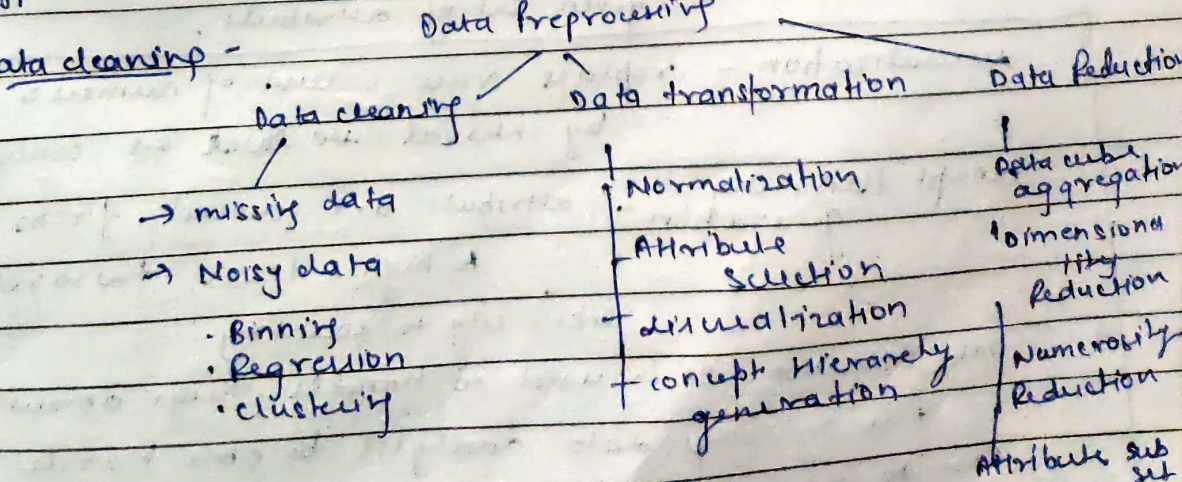
Advantage - - cost efficient , decision making
- help businesses in modification operation/strategy

disadvantage - - time consuming.
- need training to work on analytics
- Not precise.
- right mining tool is challenging to find.

Data warehouse - refers to a place where data can be stored for useful mining. It is like computer with lot of storage capacity. data from various sources eg. are copied to warehouse. where it can be ~~fetch~~ fetched.

+ Data preprocessing - is a data mining technique which is used to transform the raw data in useful and efficient format.

Data cleaning -



Date..... Page.....

* Data cleaning - data can have irrelevant and missing parts. to handle this cleaning is done.

a) Missing data — Ignore tuple
— Fill missing val. — mean, median

b) Noisy data — Noisy data is meaningless data cannot be interpreted by machine. generated due to faulty data collection, error.

- binning method - works on sorted data
- divided into segment and various method are applied.
- value can be replaced by mean - boundry value.

- Regression - data can be made smooth by fitting it to regression function.

- clustering - this approach group data in cluster. outlier may be undetected or will fall outside cluster.

* - Data transformation - step is taken to tranform data in appropriate forms suitable for mining.

- Normalization - scale data in specified range.

decimal scaling: $v' = \frac{v_i}{10^j}$, $v_i < 1$

min max: $v' = \frac{v - min(A)}{max(A) - min(A)}$

z-score: $v' = \frac{v - \bar{A}}{\sigma_A}$ — mean, — S.D

Attribute slection - new attribute are constructed from given set of attribute.

discretization - replace raw values of numeric attribute by interval or conceptual level

concept hierarchy generation - attribute are covered from lower level to higher level hierarchy.
like - city to country.

* Data Reduction - It is used to handle huge amount of data due to large data analysis become harder in such case to get rid of this we use data reduction.

- Data cube aggregation - It is applied for construction of data cube, like quately, yearly
- Attribute subset selection - highly relevant attribute should be used rest can be discarded.
- Numerosity redu - This enables to store the model of data instead of whole data ex. Regression model. Actual data is replaced by mathematical model.
- Dimensionality reduction - Reduce size of data by encoding mechanism. Ex- PCA- Principal component analysis. Such reduction can be lossy or lossless.

* Data integration - It is process of ~~converting~~ combining data from multiple heterogenous sources into single unified view.
  - Two approaches to data mining -
    1) Tight coupling - Data from various sources is combined by process of ETL - Extract, Transformation Loading.
    2) loose coupling - an interface is provided that takes query from user and transform in a way Database can understand.
       - data remains in actual source database.

Issue - data redundancy
- Schema integration
- Data value conflicts.

- Data compression - technique reduces the size of file using different encoding mechanism. (Huffman - encoding).
  - lossless compression
  - Lossy compression.

Date..... Page.....

* Data generalization - Process of summerizing data by replacing relatively low level values with higher level concepts.
It is form of descriptive mining.

1) Data cube approach — OLAP approach
- efficient
- computation and result are stored in data cube.
- It involve count(), sum(), avg(), max()
- used for knowledge discovery

2) Attribute oriented - online data analysis, query oriented.
- generalization based on attributes.
tuples are merged and their count are accumulated.
- It includes attribute removal
attribute generalization.

* • Analytical characterisation -
It is statistical approach for preprocessing data to filter out irrelevant attribute or rank relevant attribute. This preprocessing is called analytical characterization.

Reasons - decide which dimension must be included
- produce high lvl generalization
- reduce attribute that support patern easily.

Mining class comparison

class discrimination or comparison that distinguish a target class from its contrasting class. they should be comparable. Ex - person, address and item are not comparable.

comparison method and implementation
- Data collection -
- Dimension relevance analysis -
- Synchronous generalization -
- Presentation of derived comparison - can be visualise in chosen way.

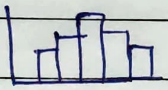
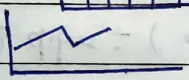
ex - we want to pgorm, compare graduate and undergraduate students using discriminant rule.

- Measuring central tendency

  Mean, median, mode.

- Measuring dispersion of data.
  - Range.
  - mean duration.
  - standard duration.

- Graph display of statistical class discription

  - Histogram 
  - line graph 
  - Bar graph 
  - Pie chart 

Association Rule mining - It finds interesting association and relationship among large sets of data item.

ex - market - Basket analysis.

Support count = Frequency occurence of a itemset.

Support (Milk, Bread, diaper) = 2.

Association rule - An implication of form $X \rightarrow Y$, X, Y are two Itemsets.

- Rule Evaluation metrics

ID	Data
1	Bread, milk
2	Bread, Diaper Beer, egg
3	Milk, diaper, Beer coke
4	Bread, milk, diaper beer
5	Bread, milk, Diaper coke.

Support - Number of transaction that include X, Y as part of rule as total percentage of all transaction. It means how frequently item together.

0.667

- confidence - (It is ratio of no. of traication that include all item in {B} as well as no of transaction that includes all item in A) to the no of transaction that includes all item in {A}.

- confidence $(X \to Y)$ = support $(X \cup Y) \div$ support (X).

- It measure how often each item in Y appear in transaction that contain X

- lift - $(X \to Y)$ = ~~conf~~ $(X \to Y) \div$ Supp (X)

{Milk, diaper} ⟶ {Beer}.

support = σ(Milk, diaper, beer) ÷ (T) = 2/5 = 0.4

confidence = σ(Milk, diaper, Beer) ÷ σ(Milk, diaper)

= 2/3

lift = Sup(Milk, diaper, Beer) ÷ supp (Milk, diaper) * Supp (Beer)

= 0.4/0.6 * 0.6 1.11

lift > 1 - appear together more than expected

lift ≥ 1

lift < 1 - appear together less than expected.

$$lift = \frac{Supp(X,Y)}{Supp(X) * supp(Y)}$$

Apriori algorithm

this algorithm is used to calculate association rules b/w object also called frequent pattern mapping. Name is apriori as it uses prior knowledge of frequent itemset property. We apply iterative approach.

ID	item
T1	91, 92, 95
T2	92, 94
T3	92, 93
T4	91, 92 94
T5	91, 93
T6	92, 93
T7	91, 93
T8	91, 92, 93, 95
T9	91, 92, 93

Limitation - slow algorithm
- Not efficient for large data
go 10^{14} frequent 1-itemset it generate 10^{17} candidate into two length which in turn will be tested
- take memory.

Step-I create table containing support count-

91	6
92	7
93	6
94	2
95	2

min-supp = 2

Step-II remove set having support count less than minimum support count.

Step-II - generate candidate set C_2

9192	4
9193	4
9194	1
9195	2
9293	4
9294	2
9295	2
9394	0
9395	1
9495	0

remove who have less than two min. support.

Step-3 generate candidate set 3

9192 93 = 2
91,92,95 = 2.

Step-4 generate candidate set C_4
↳ we have discovered all frequent item sets Now strong association rules come into picture for that we need to calculate confidence.