Data Mining (Trouble-Free)

8 sep 2022 (Sun)
(18/09/2022)

#1 Intro to Data Mining, Types of Data :-

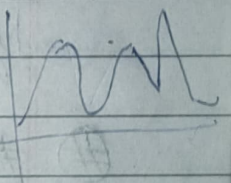
→ is defined as procedure of extracting info. from huge sets of data.
— also known as mining knowledge from data.

What types of data can be mined :-
└ 3 types — database (DB), datawarehouse, transactional data.

① Database Data (RDBMS) :- (Relational-DBMS)

→ set of tables — has rows & columns
tuples attributes

While mining databases, we can search for trends / data patterns.



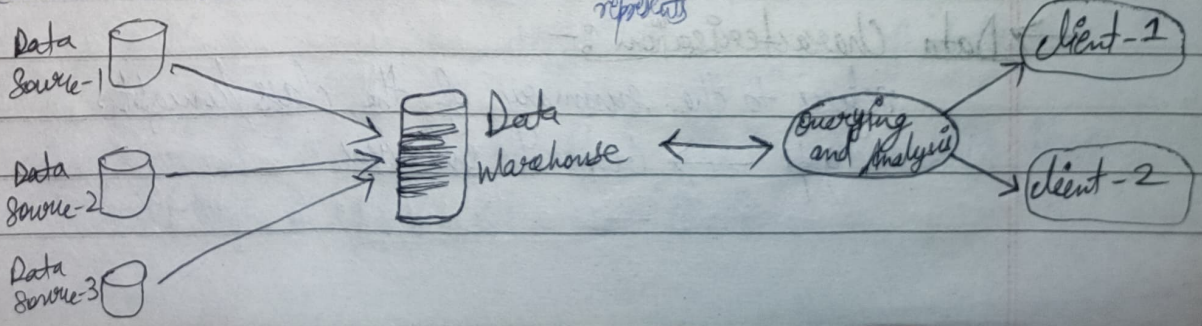
ex :- ① Analysing customer data to predict the credit risks of new customers based on prev. data.

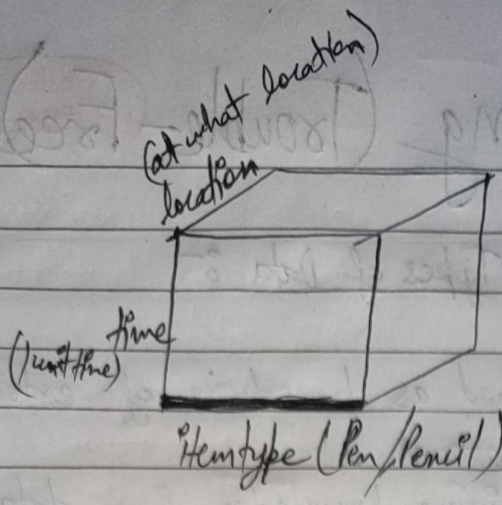
② Analysing sales data — (any deviations)
└ hike or loss (dip)

② DataWarehouse data :-

Collection of data integrated from diff^t sources with querying & decision making on data.

In data warehouse, data is stored in multidimensional structure (datacube) where each dimension represents each attribute.





(3.) Transactional database :-

Each record is called as transaction.

(Sales, flight booking, user clicks on web page.)

Transaction has transaction ID, list of other items making transaction from transaction DB, we can mine frequent patterns.

→ other types of data :-

Sequence data (stock market), data streams (continous data), spatial data (maps), engineering design data (IC circuits/designs), hypertext, multimedia, web data, etc.

#2 Data Mining Functionalities :-

(1.) Concept / class definitions (descriptions) :-

data is always associated with class / concepts.

descriptions can be done in 2 ways.

Data characterisation

Data Discrimination

→ Data Characterisation :-

refers to the summary of the class / concept.

o/p → General overview

→ Data Discrimination :-
└ Compares the common features of the two classes.

Then we notedown the changing values, and give o/p → barcharts/curves, etc.

② Mining frequent patterns, associations & correlations :-

Frequent Patterns :-
└→ Things which are found most commonly in data.
Frequent itemsets (data items/data objects)
Frequent Subsequence
Frequent Substructure

relationship analysis

→ Association Analysis :-
It's a way of identifying the relation b/w various item.
Ex:- Used to determine sales of items that are frequently purchased together.

→ Correlation Analysis :-
– mathematical technique.
– Shows how strongly pair of attributes are related together.

ex:-
Tall people, tend to have ~~not~~ more weight.
1st attribute — 2nd attribute

③ Classification and regression for predictive analysis :-
↓ prediction of data.

Classification :-
process of finding a model that distinguishes data items
(model: wrapper color for chocolates)
– decision tree is used for classification.

Aniket
Date
Page

Regression :-

Statistical methodology that is used for numeric prediction of missing data.

missing data → target data

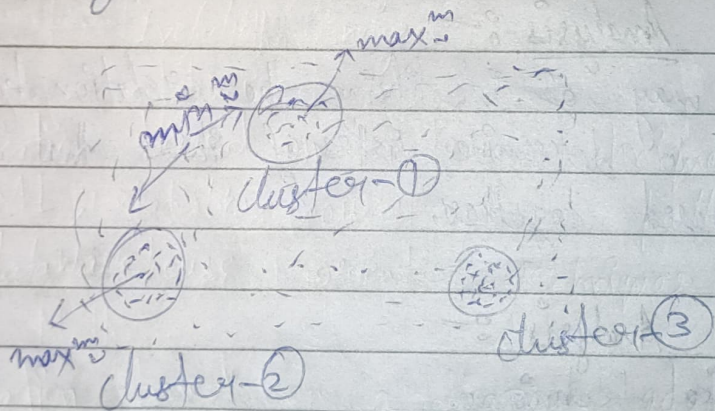
numeric prediction → done based on prev. data.

eg:- 2, 3, 4, 5, _, 7

predicted data = 6

(4.) **cluster Analysis :-**

The data items are clustered based on the principle of maximising the intraclass similarity and minimising the inter-class similarity.



⇒ Analysis of these clusters are cluster Analysis.

(5.) **Outlier Analysis :-** (anomaly mining)

eg:- {2, 4, 6, (7), 8, 10, 12}

↑ outlier (not even)

Among the data items in a db, there may be some items which do not follow the general behaviour of data. These data items → outliers (noise / exceptions).

#3. Interestingness of Patterns – What are Interesting Patterns?

* **Interestingness of patterns :–**

In a data mining system, everyday millions of data patterns are generated. Among all these patterns generated, how many are really interesting?
(means useful for user)

Actually, a small fraction of patterns generated would be of interest to any given user. This raises 3 questions :–

(1.) What makes patterns interesting? :–
→ a pattern is interesting if it's easy to understand by people (or user)
→ valid on new/old test data
→ potentially useful.

(2) Can data mining system generate all of the interesting patterns?
— refers to completeness of a data mining (dm) system.
In reality, it is not possible for a dm system to generate all interesting patterns.

(3.) Can dm systems generate only interestingness patterns?
— refers to optimisation of a dm system.
Remember, generating only interesting patterns ⇒ challenging.
If only interesting patterns are generated, it becomes easy and efficient for the user.
(time is saved).

#4. Classification of Data Mining Systems :–

Why classification needed?
dm → everywhere & anywhere (so that we can give what user actually wants/needs)
Data Mining (dm) systems are classified based on several criteria :–

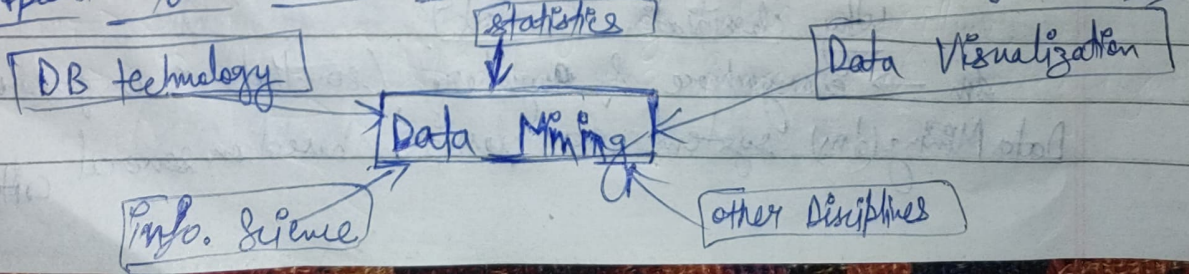
① Classification based on mined databases :-
based on type of database that is been mined.
- Relational
- Transactional
- Object-relational
- DataWarehouse

② Classification based on type of knowledge mined :-
- characterization
- Discrimination
- Association & Correlation Analysis
- classification.
- prediction
- outlier Analysis
- Evolution Analysis

③ Classification based on kinds of techniques used :-
- ML, statistics, neural networks, pattern recognition, datawarehouse oriented techniques etc.

④ Classification based on applications adapted :-
- Finance
- Telecommunications
- DNA
- Stock Markets
- Email
- etc.

Apart from above 4 classifications we have a general classification



Aniket
Date
Page

#5. Data Mining Task Primitives :-

A datamining task is represented in form of a dm query and is defined in terms of dm task primitives.

this allows the user to interactively communicate with the dm system.

5 dm task primitives :-

1. Set of task relevant data to be mined. (also called as Relevant attributes) (eg. supermarket → fruits)
2. Specifies the kind of knowledge to be mined. (i.e., functionalities)
3. The background knowledge to be used in discovery process. (concept hierarchies)
4. The interestingness measures and thresholds for Pattern evaluation
5. The expected representation for visualizing the discovery patterns to user.

whether in terms of rules / tables / Patterns / charts / graph

#6. Integration of Data Mining System with A Database or Data Warehouse System

Integration / association / combining

If there is no integration — no communication with db.

dm ⊕ db/dw

⇒ We have a total of 4 integration schemas :-

1. No Coupling :-

Dm system will not use any function. i.e., there is no communication with db.

for this it tackling directly comm. with other storage methods. (like file system)

Aniket
Date
Page

② Loose Coupling :- (≈ 15%)
- will use some of the functionalities (only upto some extent)
- better than no coupling.
- Suitable for small data sets.

③ Semi-Tight Coupling :-
- linked to the db. (> 50%)
- also, some of the dm primitives are also implemented in db.

④ Tight Coupling :- (≈ 100%)
- dm system is completely linked to db.
- most efficient among all.

The db system is fully integrated in such a way that it becomes part of the dm system.
- efficient & optimised implementation of dm.

#7. Major Issues in Data Mining :-

→ Mining Methodology & User-Interface/interaction issues :-

→ Mining diff. kinds of knowledge from databases.
→ interactive mining of knowledge at multiple levels of abstraction.
→ Handling Noisy / incomplete data.
→ Pattern evaluation.

→ Performance Issues :-

→ efficiency & scalability of dm algorithms.
→ Parallel, distributed and incremental mining algorithms

→ Diverse Data Types Issues :-

→ Handling of relational and complex types of data.
→ Mining information from heterogeneous db's and global info. systems.

#8. Data Preprocessing in Data Mining :- (-4 steps)

→ The process of transforming raw data into an understandable format.

— 4 major tasks :-

1. Data Cleaning
2. Data Integration
3. Data Reduction
4. Data Transformation

* Data Cleaning :- — cleaning of — missing data, noisy data.

Process of removal of incorrect, incomplete, inaccurate data, also replaces missing data.

1. Handling missing values :-

In place of missing values, we can replace with "NA",
with mean value (Normal data)
with median value (non-normal data)

- sometimes replaced with most probable values.
- missing values can be filled in (2) ways.

manual — small

automate — more efficient — large data sets

Aniket
Date
Page

(2) Handling noisy data :-

noisy data → inconsistent/error data.

methods to handle noisy data :-

(1) Binning :-
firstly, data is sorted. Then sorted data is stored in bins.

3 methods to handle data in bins :-
- smoothing by bin mean ($\frac{a+b+c+d}{4}$)
- smoothing by bin median (1, 2, 3, 4, 5)
- smoothing by bin boundary. (min & max)

(2) Regression :-
Numerical prediction of data.

(3) clustering :-
- similar data items are grouped at one place.
- dissimilar items — outside the cluster.

* Data Integration :-

Multiple heterogeneous sources of data are combined into single dataset.

There are 2 types of data integration :-

(1) Tight coupling :-
Data is combined together into a physical location.

(A) + (B) → (C)

(C) ← X (no return)

(2) Loose coupling :-
Only an interface is created and data is combined through the i/f (interface) and also accessed through i/f.
- Data remains in actual database only.

Aniket
Date
Page

* Data Reduction :-

Volume of data is reduced to make analysis easier.

methods for data reduction :-

① Dimensionality reduction :-
reduces no. of i/p variables in the dataset because large i/p variables ⟶ poor performance.

② Data Cube aggregation :-
Data is combined to construct a datacube
(Redundant, noisy data is removed)
(means duplicate records will be removed)

③ Attribute Subset Selection :- (i.e. Table columns)
Highly relevant attributes should be used, others should be discarded/removed.

④ Numerosity Reduction :-
Here, we store only model of data instead of entire data.
(sample)

* Data Transformation :-

Data is transformed into an appropriate form suitable for mining process.

4 methods for data transformation :-

① Normalisation :- Done in order to scale the data values in specified range (-1.0 to 1.0 or from 0 to 1)

② Attribute Selection :-
New attributes are created using older ones.

18-20 21-30

Aniket
Date
Page
10, 11, 12, 13, ~~14, 15, 16, 17, 18~~, 34, 35.

③ Discretization :-
Raw values are replaced by interval levels.

④ Concept hierarchy generation :-
attributes are converted from low level to high level.
Ex:- city → country
low level high level

#9. Frequent Patterns – Example, Market Basket Analysis :-

* Frequent Patterns :-
The patterns that appear frequently in a dataset.
(include frequent datatems, sequences, substructures)
eg. computer | computer, mouse, keyboard | graph, tree

eg:- Milk and Bread
└ both buyed together → frequent pattern

* Market-Basket-Analysis :-
Process of analysing customer buying habits, by finding the associations b/w the different items that a customer will place in their baskets.
— mainly useful for sellers.

→ Strategies used :-
1. Placing them together
2. Placing them at 2-diff ends.

— This analysis will help sellers to plan their shelf space for increased sales.

- frequent patterns are represented by association rules.

Ex:- Computer and Anti-virus.

General — specific to context

Computer ⇒ anti-virus software [support = 2%, confidence = 60%]
purchased purchased

out of all transactions 2% bought computer & anti-virus together.

→ Support :-
Identifies how frequently a rule is applied to given dataset.

$$S(P \rightarrow Q) = \frac{\sigma(P \cup Q)}{N}, \quad N = \text{total transactions}$$

→ Confidence :-
It defines frequent occurrence of 'items' of Q in transactions of P.

$$C(P \rightarrow Q) = P(Q/P)$$

P tends to Q

#10. Mining Methods – APRIORI algorithm with Example :-

* Mining Methods :-

→ Apriori algorithm
→ FP Growth Algorithm

→ Apriori Algorithm :- — Given by R. Agarwal & R-Srikant
L shows how objects are associated with each other

objective :-
— To generate an association.

Ex:- Min. support = 50%
Threshold confidence = 70%
(or min.)

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C1 (freq.)

Itemset	Support	min. support
{1}	2	2/total no. of transactions = 2/4 = 50%
{2}	3	3/4 = 75%
{3}	3	3/4 = 75%
~~{4}~~	1	1/4 = 25% < 50% (min. support)
{5}	3	3/4 = 75%

forms L1

Itemset (new) = {1,2,3,5} ⇒ form Pairs :- (1,2), (1,3), (1,5), (2,3), (2,5), (3,5)

C2 (freq.)

Itemset	Support	min. support
~~(1,2)~~	1	1/4 = 25%
(1,3)	2	2/4 = 50%
~~(1,5)~~	1	1/4 = 25%
(2,3)	2	2/4 = 50%
(2,5)	3	3/4 = 75%
(3,5)	2	2/4 = 50%

new Itemset = {(1,3), (2,3), (2,5), (3,5)}

Aniket
Date
Page

⇒ form triplets :- (1,2,3), (1,3,5), (2,3,5)

C3

Itemset	support	min. support
(1,2,3)	1	1/4 = 25%
(1,3,5)	1	1/4 = 25%
(2,3,5)	2	2/4 = 50%

↳ new Itemset = {(2,3,5)}

⇒ Now lets calculate support & confidence.

$$\text{Confidence} = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

Using (2,3,5) we can generate association rules.

Rules	Support	Confidence
(2^3) → 5	2	2/2 = 100%
(3^5) → 2	2	2/2 = 100%
(2^5) → 3	2	2/3 = 66%
2 → (3^5)	2	2/3 = 66%
5 → (2^3)	2	2/3 = 66%
3 → (2^5)	2	2/3 = 66%

Strong & perfect association rules among all.

- $\underbrace{(2 \wedge 3)}_{A} \rightarrow \underbrace{5}_{B}$ confidence $= \frac{S((2 \wedge 3) \cup 5)}{S(2 \wedge 3)} = \frac{2}{2} = 100\%$
- $\underbrace{2}_{A} \rightarrow \underbrace{(3 \wedge 5)}_{B}$ confidence $= \frac{S(2 \cup (3 \wedge 5))}{S(2)} = \frac{2}{3} = 66\%$

∴, $(2 \wedge 3) \rightarrow 5$ and $(3 \wedge 5) \rightarrow 2$ are association rules.

#11. Mining Methods – FP growth algorithm with example:-

* FP Growth Algorithm:-

FP → Frequent Pattern

→ its an efficient & scalable method for mining the complete set of FP using a tree structure for storing information about FP called FP tree.

ex:- min. support = 30%

Trans id	Items
1	E, A, D, B
2	D, A, E, C, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

⇒ all items are A, B, C, D, E

To write priorities:-

① more freq. → more priority
② same freq. → FCFS priority

- list out all the priorities

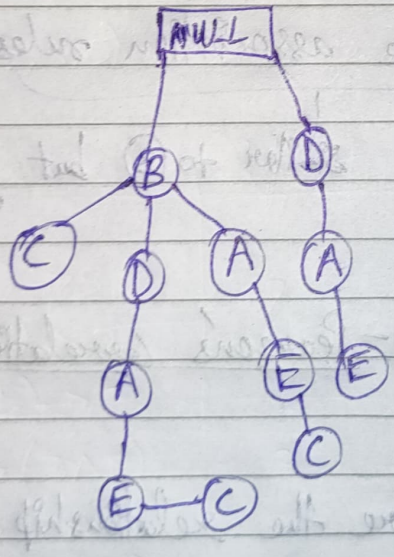
Itemset	Freq.	Priority
A	5	3
B	6	1
C	3	5
D	6	2
E	4	4

B D A E C

— order items according to their priority

Trans ID	Items	Ordered Items
1	E A D B	B D A E
2	D A E C B	B D A E C
3	C A B E	B A E C
4	B A D	B D A
5	D	D
6	D B	B D
7	A D E	D A E
8	B C	B C

— Construct the FP tree.



B – 1, 2, 3, 4, 5, 6
D – 1, 2, 3, 4
A – 1, 2, 3
E – 1, 2
C – 1

A – 1
E – 1
C – 1

D – 1, 2
A – 1
E – 1
C – 1

freq of D = 4 + 2
= 6

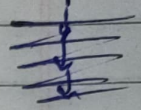
(you can check in prev page freq. of D is same or not. for correctness.)

Aniket
Date
Page

#12. Mining various kinds of Association Rules :-

* **Mining Various kinds of Association Rules :-** (4 types)

1. **Mining Multi-level Association rules :-**
   - using uniform support for all levels.
   - using reduced minimum support at lower levels.
   - using item or group based minimum support.



2. mining multidimensional Association Rules from relational-db or relational dW. — each attribute considered as a predicate, eg:- buy(x), sell(x) etc.
3. mining multi-dimensional association rules using static discretisation of quantitative attributes. — numerical intervals manually.
4. mining quantitative association rules. — similar to (3) but by dynamic, using binning

#13. Correlation Analysis - Pearson's correlation coefficient :-

* **Correlation Analysis :-**
  ↳ used to measure the relationship b/w two variables.

$$\gamma_{A,B} = \frac{\sum (A - A')(B - B')}{(n-1) \cdot \sigma_A \cdot \sigma_B}$$

gamma

- n = no. of tuples in dB
- Karle Pearson correlation coefficient
- A', B' = mean of A and B.
- σ_A, σ_B = standard deviation of A & B.

↑ ↑ → +ve correlation
↑ ↓ or ↓ ↑ → -ve correlation
& No correlation if no proportion is found

r can have $\longrightarrow$ 3 variables $(0, -1, +1)$
(or values)

$r = +1 \longrightarrow$ Perfect positive correlation
$r = 0 \longrightarrow$ No Correlation (no dependence)
$r = -1 \longrightarrow$ Perfect negative correlation

Ex:-