Total pages:3

IV YEAR B.TECH EXAMINATION, DECEMBER - 2021

Data Mining & Warehousing

150712

Time: 3 Hours **Maximum Marks: 70** **Minimum Marks: 28**

Note:	
	1. *Answer all five questions. All questions carry equal marks.* 2. *In each question part a, b, c are compulsory and part d has internal choice. Out of which part a & b carries 2 marks each, part c carry 3 marks and part d carry 7 marks.* 3. *All parts of each question are to be attempted at one place.* 4. *Assume suitable value for missing data, if any*

Question No.			Marks	Course Outcomes
1.	(a)	What is data mining? How does KDD differ from data mining?	2	5
	(b)	What is Temporal Database?	2	1
	(c)	How is a data warehouse different from a database? How are they similar?	3	1,2
	(d)	Write an essay on Application of Data Mining	7	5
		OR		
		Give the architecture of a data mining system. What are the essential components of a data mining system? Describe the purpose of each of these components.	7	4
2.	(a)	What do you understand by OLTP technology?	2	1
	(b)	Discuss the advantages of Data Cube?	2	1
	(c)	What do you understand by multi-dimensional data model?	3	2
	(d)	Discuss about the following schemas in the context of Data Warehousing (i) Star Schema (ii) Snowflake Schema (iii) Fact Constellation Schema	7	3
		OR		

<table>
<tr><td></td><td></td><td>What do you understand by data cube? Discuss different operations (with suitable examples) which can be performed on a data cube.</td><td>7</td><td>3</td></tr>
<tr><td>3.</td><td>(a)</td><td>Discuss the need and importance of data preprocessing.</td><td>2</td><td>4</td></tr>
<tr><td></td><td>(b)</td><td>Describe why concept hierarchies are useful in data mining.</td><td>2</td><td>4</td></tr>
<tr><td></td><td>(c)</td><td>What do you understand by data transformation? Discuss with suitable examples.</td><td>3</td><td>3</td></tr>
<tr><td></td><td>(d)</td><td>Discuss with suitable examples different steps of data preprocessing.</td><td>7</td><td>3,4</td></tr>
<tr><td></td><td></td><td>OR</td><td></td><td></td></tr>
<tr><td></td><td></td><td>In real world data, records with missing values for some attributes are common occurrence. Describe various methods and ways for handling this problem.</td><td>7</td><td>3</td></tr>
<tr><td>4.</td><td>(a)</td><td>What do you understand by strong association rules?</td><td>2</td><td>5</td></tr>
<tr><td></td><td>(b)</td><td>What are the different parameters to measure association (rules) and how they are measured?</td><td>2</td><td>5</td></tr>
<tr><td></td><td>(c)</td><td>What do you understand by the market basket analysis? Discuss its importance.</td><td>3</td><td>5</td></tr>
<tr><td></td><td>(d)</td><td>Write the Apiori Algorithm and discuss its working with suitable example.</td><td>7</td><td>5</td></tr>
<tr><td></td><td></td><td>OR</td><td></td><td></td></tr>
<tr><td></td><td></td><td>A database has four transactions. Let Minimum support = 50% and Minimum confidence = 80%
<table>
<tr><th>Tid</th><th>Items</th></tr>
<tr><td>100</td><td>D,K</td></tr>
<tr><td>200</td><td>A,C,D,E</td></tr>
<tr><td>300</td><td>A,B,C,E</td></tr>
<tr><td>400</td><td>A,B,D</td></tr>
</table>
(i) Find all frequent items using Apriori algorithm.
(ii) List all strong association rules</td><td>7</td><td>5</td></tr>
<tr><td>5.</td><td>(a)</td><td>What do you understand by the term Cluster ? Give suitable example.</td><td>2</td><td>6</td></tr>
</table>

	(b)	What do you understand by the term Classification? Give suitable example	2	6
	(c)	Discuss the issues related to prediction.	3	4
	(d)	List different Clustering algorithms and discuss any one with suitable example.	7	6
		OR		
	(e)	List different Classification algorithms and discuss any one with suitable example.	7	6

****

DEC - 2021
DATA MINING & WAREHOUSING
150712

Data mining is a technology to discover previously unknown, hidden and useful patterns or relationship from a large database. KDD is a process and data mining is an important step of KDD process

(b) Temporal data is simply data that has certain features that support time sensitive status for entries.

Examples : Land use of India in 2021

: Event sequences etc.

(c) A database is an organized collection of data store in a such a way that facilitates easier search, retrieval, manipulation and analysis of data.

A datawarehouse is also a type of database that integrates copies of transaction data from disparate source and provide them for analysis.

1(b) Some areas where datamining are being used, listed below:

- Market Basket Analysis
- Floor management
- Healthcare
- Education
- CRM
- Fraud detection
- Webpage Analysis
- Intrusion detection
- Financial banking
- Research Analysis
- Bio - informatics

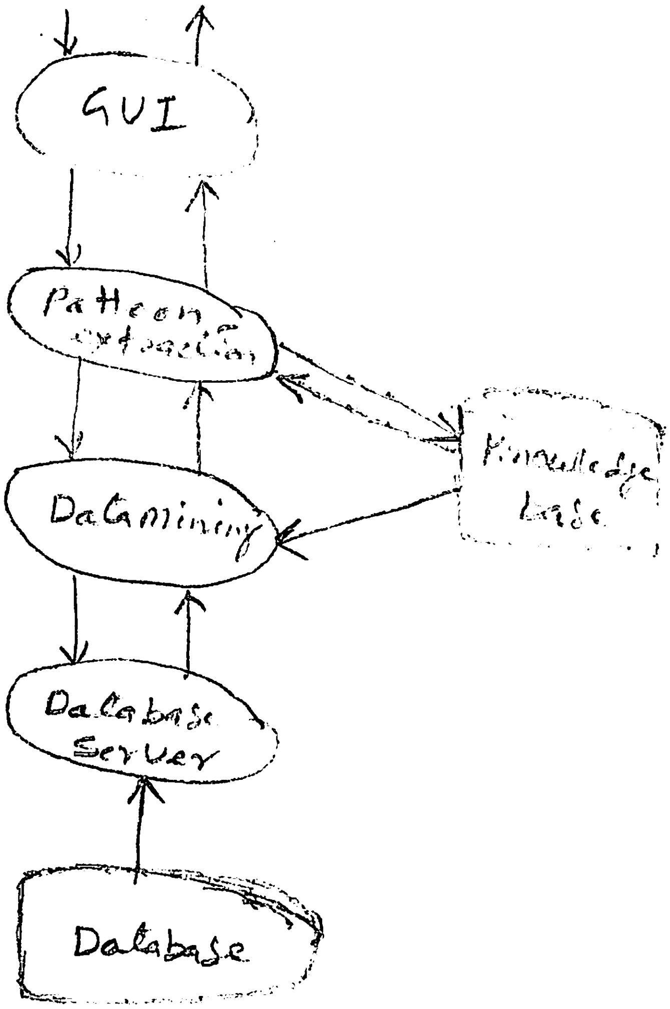
and many more.

Market Basket data analysis identify the purchasing behavior of the buyer and this information can be used for making strategy in future for improving sell. Also information retrieved from market Basket analysis helps the retailer to know the buyer's need and change the store layout accordingly. The huge amount of data is generated by the healthcare transactions. This data is too complex and voluminous to be processed ana analysed by traditional methods. Datamining provides the methodology and technology to extract useful knowledge from such data in decision making.

Educational data mining (EDM) is a new emerging field. Institutions can identify the learning patterns of the students by using EDM and can use it to develop new effective techniques.

OR

Architecture of Datamining System



Database Repository:

This may be a single source of database

or it may be a set of databases. Here data cleaning and data integration techniques may be applied on the data.

Database Server;

This is an important component of any mining system which is responsible for fetching the relevent data from database.

Knowledgebase:

This is the domain knowledge that is used to guide the search

Datamining Engine:

This is the heart of overall data mining system. It may consist of set of functional modules for one or more mining task.

Pattern Evaluation Module:

This component assists in searching interesting pattern. It uses interestingness threshold furnished by the user

Graphical User Interface (GUI)

This component communicate between user and the datamining system. This assists the user several ways.

(a) OLTP (online transactional processing) is also called Operational database management system. It allows to view and update data in real time.

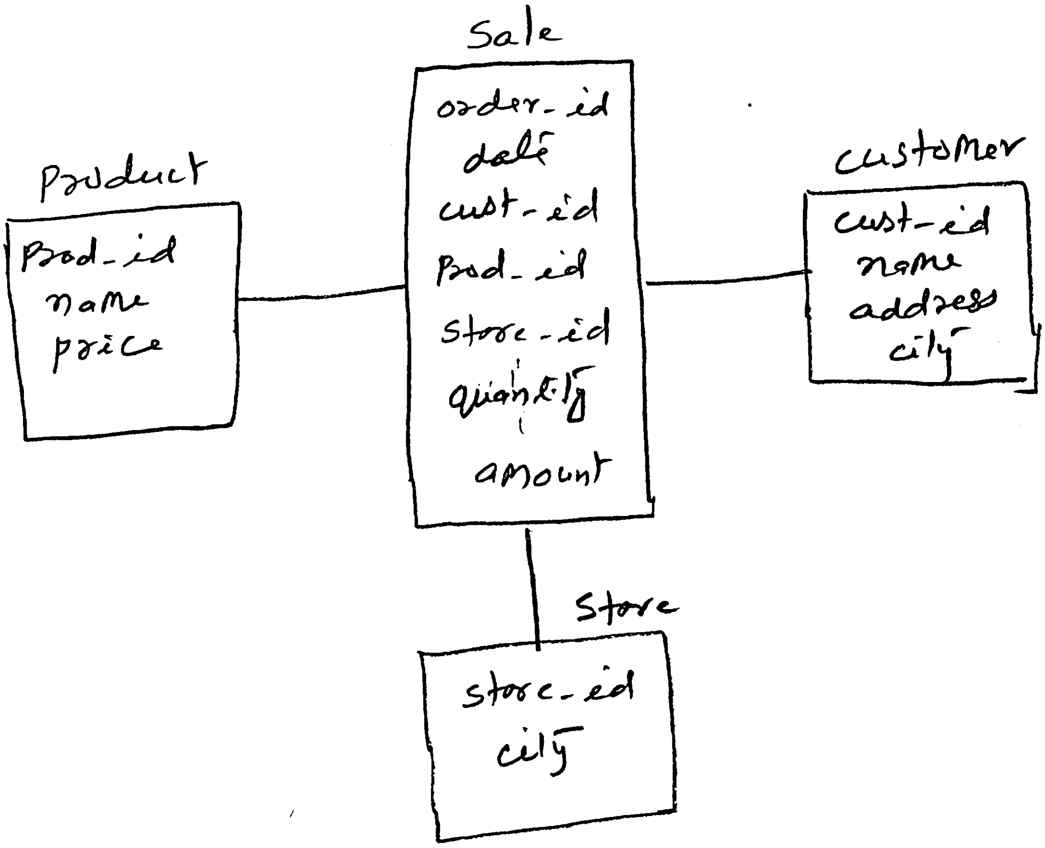
(b) Advantage of data cube

- Data cube ease in aggregating and summarizing the data
- Datacube provides better visualization of data
- Datacube stores the data (huge) in a very simplified way

(C) Data cube is a group of multidimensional matrices. Each dimention of a cube represents certain characteristics of the database for example daily, monthly or yearly sales.

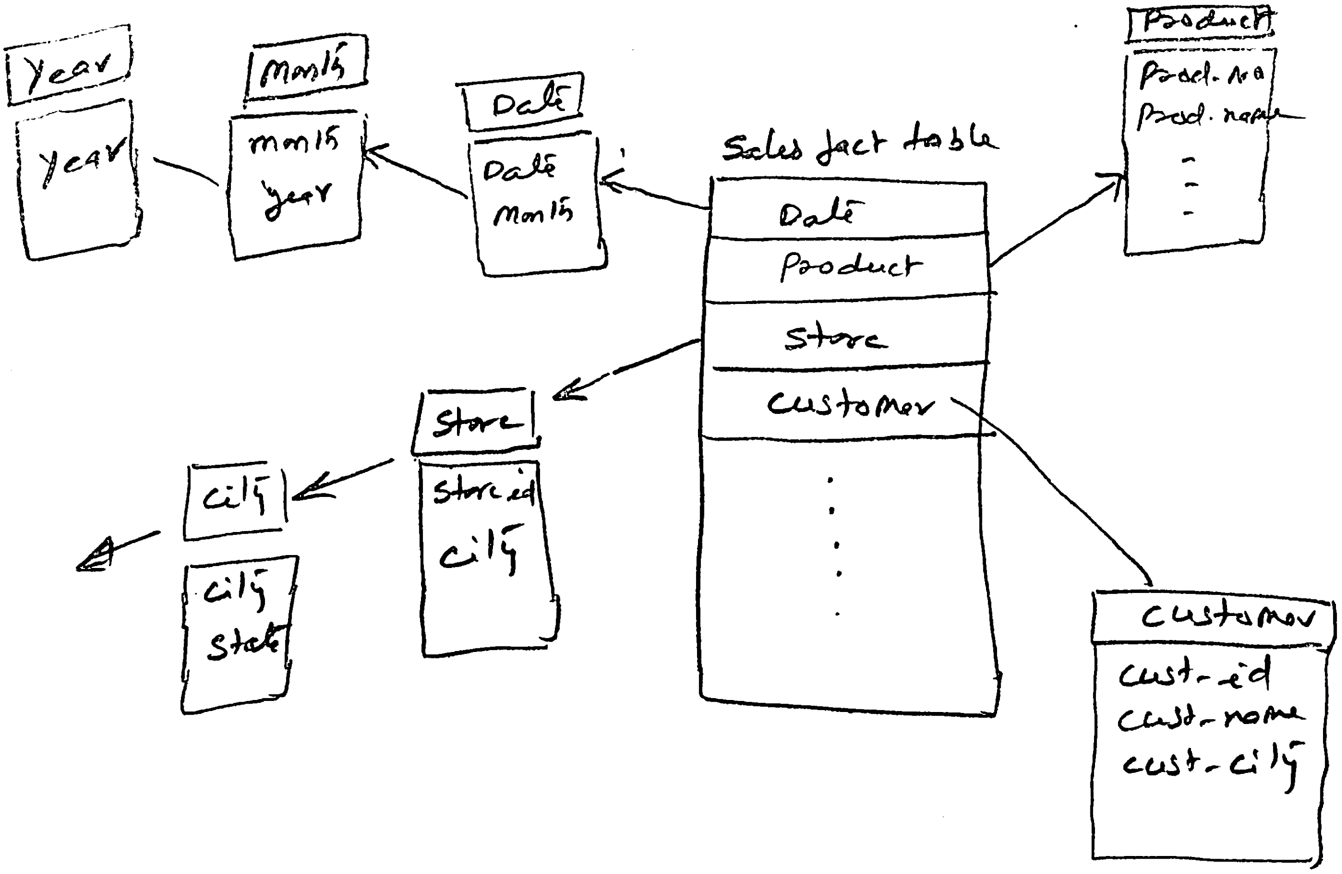
2(d) Star schema

- It is The simplest style of schema and This approach is most widly used to develop data warehouse and dimensional data marts. In This schema There is one fact table in the middle and a number of associated dimension table (look like star) as shown below:



Snowflake schema:

- This is an extension of star schema. There are additional dimensions added to star schema. This schema is known as snowflake due to its structure.



Fact constellation schema

- such schema can consist of multiple fact tables. This schema is also known as glaxy schema. It is Viewed as a collection of stars and hence name is glaxy. The shared dimension in this schema are known as comformed dimension. This type of schema is usually used for sophesticated applications. The multiple number of tables present in this schema makes it difficult and complex. Implementation of such schema is difficult.

OR

A data cube is a multidimensional data model used to store data. It is used to easily interpret data. It is especially useful when representing data together with dimensions as certain measures.

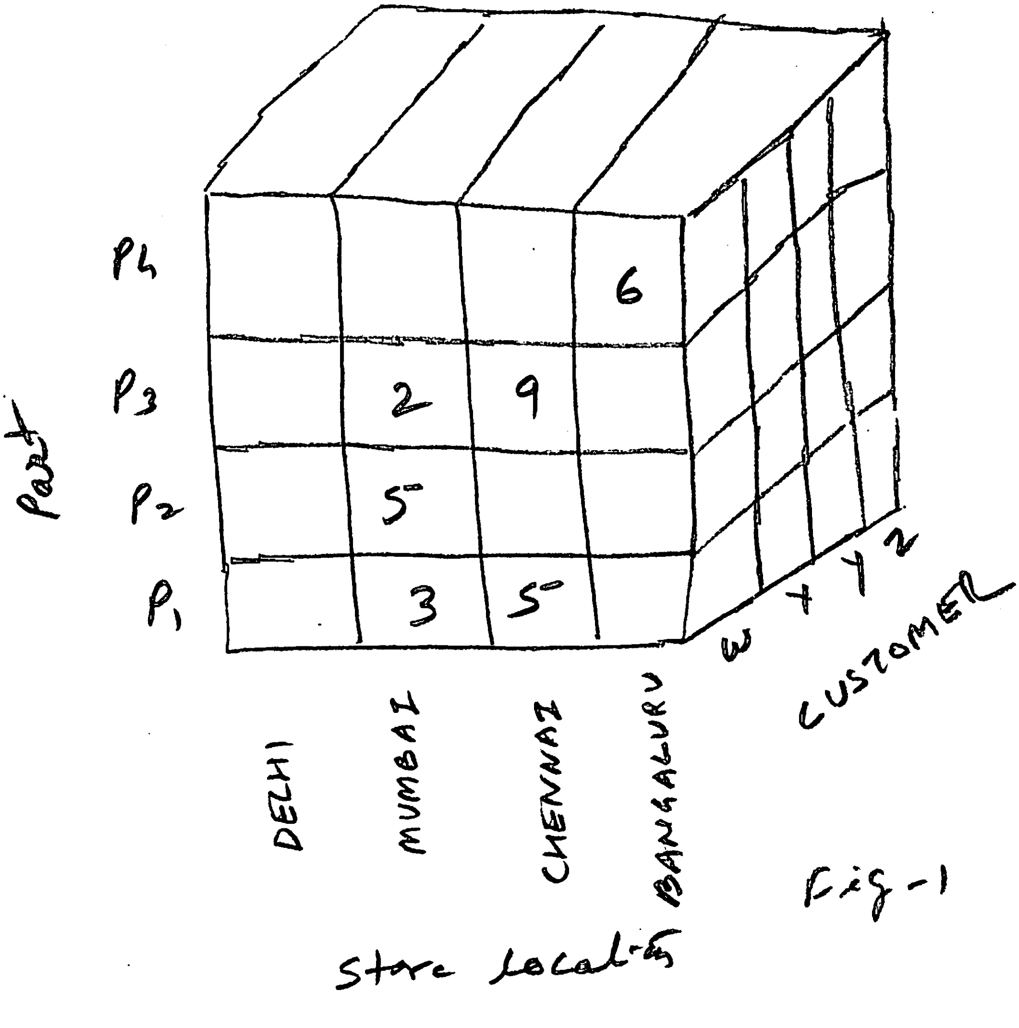


Fig-1

Following operations can be executed on the data cube inorder to view data from different perspective:

- Roll up:

This operation summarizes or aggregates the dimension either by perforing dimension reduction or concept hierarchy on a data cube.

If cities are divided in zone, as shown below

zone	cities
zone-I :	Delhi, Mumbai
zone-II :	Chennai, Bangaluru

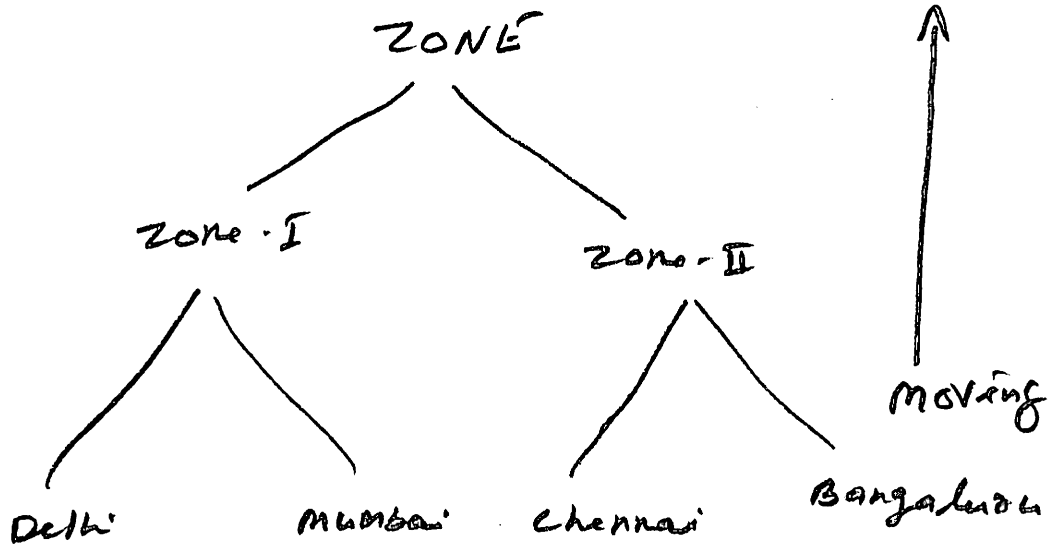


Fig-2

Now if apply on our example cube then new cube will be as shown below:

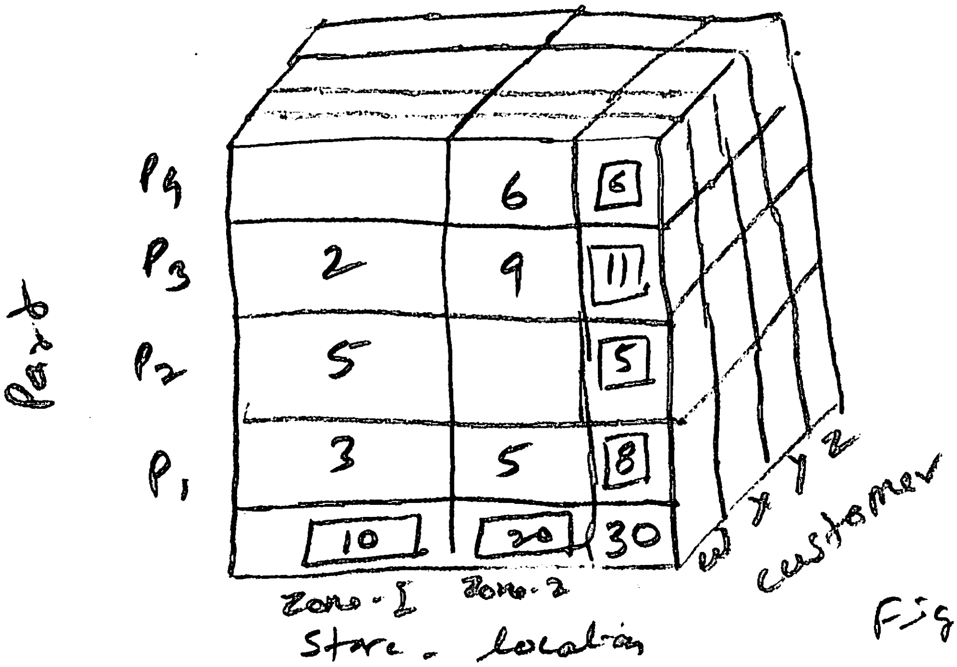


Fig-3

Drill Down:

when the drill-down operation is performed on any dimension, the data on the dimension is fragmented into granular form. This operation is just opposite of Roll-up. It can be done by:

- Moving down in the concept hierarchy
- Adding a new dimension

As this operation is just opposit of Roll-up. if we apply this operation on the cube of Fig. 3 on dimension zone then we will get the cube as shown in Fig 2.

Slice and Dice;

slice operation performs a selection on one dimension of the given cube and produces a new sub cube with reduce dimensionality.

The Dice operation performs a selection on two or more dimensions of the given cube and produces a new sub cube with reduce dimensionality

Pivot :

Pivot is not a calculative operation actually it rotates the data cube in order to view data cube from different dimension

3(a) Pre-processing is a technique to turn the raw-data (generated from diverse sources) into a form that is more suitable for work. By applying pre-processing on the raw data, quality of the data is improved to get the more accurate results.

(b) Concept hierarchies allows higher level knowledge patterns to be found. It allows Data mining at multiple levels of abstraction, which is common requirement for data mining applications.

(C) Data transformations are required because of syntactic reasons. Usually they lead to syntactic modifications with no alterations in semantics. For example there may be different format for date such as 1.1.2021 or 1/1/2021 or 1st JAN 2021 or 01/01/2021. As per the requirement of algorithm these format may be transformed in single one.

Data transformation can be done through several techniques:

- Normalization
- Attribute selection
- Discretization
- Concept hierarchy generation
- Data reduction

⋮

3(D) There are many important steps in the data preprocessing. Two import steps are

- Data cleaning:

Data gathering process may not be perfect and this may cause many irrelevant and missing parts. Further raw data can have missing values as well as there may be a lot of redundant data. The most common problems with raw data can be divided into 3 groups:

- Missing data
- Noisy data
- Inconsistent data

The method used to address these issue is called data cleaning

- Data Transformation:

After handling the issues mentioned above, data pre-processing moves on to the transformation stage. Here data is transformed into appropriate format. This can be done through several techniques including Normalization, attribute selection, Discritization etc. Normalization refers to scaling down of the dataset such that the normalized dataset falls in the range 0 and 1. There are different methods for data normalization

- Min-Max normalization
- Z-score normalization
- Normalization by decimal scaling

Example for Min-max normalization

Let the minimum and maximum values for the attribute income in a database are Rs. 12000 and Rs. 98000 respectively. Normalize income for Rs. 73600 can be calculated as below:

$$V' = [(V - min_A) / (max_A - min_A)] * [new\_max_A - new\_min_A)] + new\_min_A$$

$$= (73600 - 12000) / (98000 - 12000) * [(1-0)] + 0$$

$$= 0.716$$

OR

Missing data can be seen as inaccurate data since the information that is not there creates gaps that might be relevant to the final analysis. Missing data often appears when there is a problem in the collection phase, such as a glitch that causes a systems downtime, mistakes in data entry or issues with bio-metrics use among others.

One can ignore the part of the data set that has the missing values. This approach is only feasible when working with a big data set. In other cases, the best approach is fill the the missing values by inputting the values manually or using computational process that assign values through attribute mean or by calculating the most probable value.

4(a) The task is to find interesting associations or co-relations among a large set of data that is to identify sets of items that frequently occur together and then formulate rules that characterize their relationship. A typical example is Market basket analysis An association rule can be represented as

$$X \longrightarrow Y$$

where X and Y are called the antecedent and consequent respectively.

(b) Support and confidence are the parameters to measure the association.

$$\text{support} = \frac{(X \cup Y).\text{count}}{n}$$

$$\text{confidence} = \frac{(X \cup Y).\text{count}}{X.\text{count}}$$

where n represents total no of transaction X.count represents the support count of an itemset X.

(C.) Market Basket Analysis is an important application of association rule. This analysis identify the purchasing behavior of the customer. The information receive during this analysis can be used for making strategy in future for improving sell. and other decisions.

Apriori algorithm

```
Input
    D        // data base of Transactions
    min-sup  // The minimum support
Output
    L        // Frequent items in D

$L_1$ = find-frequent-1-itemset(D);
for(k=2, $L_{k-1} \neq \phi$, k++
{   $C_k$ = apriori-gen($L_{k-1}$);
    for each transaction t $\in$ D {
        $C_t$ = subset($C_k$, t)
        for each candidate c $\in C_t$
            c.count ++;

    }
    return L = $\cup_k L_k$
```

Algorithm first generate the frequent items of size 1 i.e. L_1. By using L_1 it generate C_2 & its used for generating L_2 and so on. It uses function apriori-gen(L_{k-1}) for getting C_k.

For its working please refer the solution of question 4 second part given below where frequent items were calculated for the minimum given support of 80%

OR

Given that min. support = 50%
min - confidence = 80%

(1) For finding L_1, first count value of each item is calculated

$|A| = 3$ ✓

$|B| = 2$ ✓

$|C| = 2$ ✓

$|D| = 3$ ✓

$|E| = 2$ ✓

$|K| = 1$ ✓

$\Rightarrow L_1 = \{A, B, C, D, E\}$

$C_2 = \{\underset{2}{AB}, \underset{2}{AC}, \underset{2}{AD}, \underset{2}{AE}, \underset{1}{BC}, \underset{1}{BD}, \underset{1}{BE}, \underset{1}{CD}, \underset{2}{CE}, \underset{1}{DE}\}$

$L_2 = \{AB, AC, AD, AE, CE\}$

$C_2 = \{\underset{1}{ABC}, \underset{1}{ABD}, \underset{1}{ABE}, \underset{1}{ACD}, \underset{2}{ACE}, \underset{1}{ADE}\}$

$L_2 = \{ACE\}$

$L = L_1 \cup L_2 = \{AB, AC, AD, AE, CE, ACE\}$

(ii) Now strong rules will be calculated for the frequent items obtained in (1)

$A \rightarrow B$

$$= \frac{|AB|}{|A|} = \frac{2}{3} = 0.66 = 66.66\%$$

$B \rightarrow A$

$$\frac{|AB|}{|B|} = \frac{2}{2} = 1.0 = 100\%$$

$A \rightarrow C$

$$= \frac{|AC|}{|A|} = \frac{2}{3} = 0.66 = 66\%$$

$C \rightarrow A$

$$= \frac{|AC|}{|C|} = \frac{2}{2} = 1.0 = 100\%$$

$A \rightarrow D$

$$= \frac{|AD|}{|A|} = \frac{2}{3} = 0.66 = 66.66\%$$

$D \rightarrow A$

$$= \frac{|AD|}{|D|} = \frac{2}{3} = 0.66 = 66.66$$

$A \rightarrow E$

$$= \frac{|AE|}{|A|} = \frac{2}{3} = 0.66 = 66.66$$

$C \rightarrow E$

$$\frac{|CE|}{|C|} = \frac{2}{2} = 1 = 100\%$$

$E \rightarrow C$

$$\frac{|CE|}{|E|} = \frac{2}{2} = 1 = 100\%$$

<u>ACE</u>

$A \rightarrow CE$

$$= \frac{|ACE|}{|A|} = \frac{2}{3} = 0.66 = 66.66\%$$

$AC \rightarrow E$

$$\frac{|ACE|}{|AC|} = \frac{2}{2} = 1 = 100\%$$

$E \rightarrow AC$

$$= \frac{|ACE|}{|E|} = \frac{2}{2} = 1 = 100\%$$

$AE \rightarrow C$

$$= \frac{|AEC|}{|AE|} = \frac{2}{2} = 1 = 100\%$$

$C \rightarrow AE \Rightarrow \frac{|AEC|}{|C|} = \frac{2}{2} = 100\%$

All associaton rules whoses confidence are 80% or more are the strong assocation rules.

(a) Cluster is a collection of data objects that are similar to one another in group

Example: Group of Male students and female students.

Clustering can help business to Manage their data better - image segmentation, grouping web page.

(b) Classification is the problem of identifying to which of a set of categories, a new object belongs to, on ~~a~~ the basis of a training set of data containing observations & whose categories of membership is known

Example: A classifier can be used to identify whether starting a particular project is safe or risky.. It is a two-step process 1 Learning step 2. classification step.

During 1 step classifier is trained & in the 2nd step model used to predict class labels & testing the constructed model on the test data.

C The accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new or previously unseen data. Following are the issues regarding preprocessing the data for prediction.

- Data cleaning
- Relevance analysis
- Data transformation
- Data reduction

D. Different clustering algorithm.

- Partitioning Methods
  - K-Means
  - K-Medoids
- Hierarchical Methods
  - Agglomerative
  - BIRCH
  - ROCK
- Density-Based methods
  - DBSCAN
  - OPTICS
- Grid Based Method

⋮

OR

Classification methods/algorithm

- Bayesian classification
  - Bayes theorem
  - Naive Bayesian
- Ruled based classification
- Support Vector Machine
- k-nearest neighbors algorithm
- Random forest

Students can describe any one

25