UNIT - I

1 Introduction to Data Warehousing

1.1 : Data Warehouse Systems and its Components

Q.1 What is data warehouse ?

Ans. : Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

Q.2 What is data warehouse ? How does it differ from a database ?

Ans. : • A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. A data warehouse stores historical data for purposes of decision support.

- A database an application-oriented collection of data that is organized, structured, coherent, with minimum and controlled redundancy, which may be accessed by several users in due time.
- Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions
- A data warehouse is a subject-oriented collection of data that is integrated, time-variant, non-volatile, which may be used to support the decision-making process.
- Data warehouses are databases that store and maintain analytical data separately from transaction-oriented databases for the purpose of decision support

Data organization in data warehouses is based on areas of interest, on the major subjects of the organization : customers, products, activities etc. Databases organize data based on enterprise applications resulted from its functions.

- The main objective of a data warehouse is to support the decision-making system, focusing on the subjects of the organization. The objective of a database is to support the operational system and information is organized on applications and processes.

Q.3 List and explain characteristics of data warehouse.

Ans. : Characteristics :

1. **Subject oriented :** Data are organized based on how the users refer to them. A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject
2. **Integrated :** All inconsistencies regarding naming convention and value representations are removed. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.
3. **Nonvolatile :** Data are stored in read-only format and do not change over time. Typical activities such as deletes, inserts, and changes that are performed in an operational application environment are completely nonexistent in a DW environment.
4. **Time variant :** Data are not current but normally time series.

Q.4 How are organizations using the information from data warehouses ?

Ans. : • Many organizations use this information to support business decision-making activities, including.

a) increasing customer focus, which includes the analysis of customer buying patterns.

b) repositioning products and managing product portfolios by comparing the performance of sales by quarter, by year, and by geographic regions in order to fine tune production strategies.

c) analyzing operations and looking for sources of profit.

d) managing the customer relationships, making environmental corrections, and managing the cost of corporate assets.

Q.5 What is data warehouse? Discuss most important issues in data warehouse implementation. [RGPV : June-15, Marks 7]

Ans. : most important issues in data warehouse implementation are as follows :

1. Data content
2. Metadata
3. Data distribution
4. Tools
5. Performance consideration.

- Steps for designing data warehouse:

1. Choosing the subject matter
2. Deciding what a fact table represents
3. Identifying data conforming the dimensions
4. Choosing the facts
5. Storing pre-calculation in the fact table
6. Rounding out the dimension tables
7. Choosing the duration of the database
8. The need to track slowly changing dimensions
9. Deciding the query priorities and query mode

Q.6 Explain data warehouse architecture.

Ans. : There are mainly three types of datawarehouse architectures :

1. **Single-tier architecture :** The objective of a single layer is to minimize the amount of data stored. This goal is to remove data redundancy. This architecture is not frequently used in practice.
2. **Two-tier architecture :** Two-layer architecture separates physically available sources and data warehouse. This architecture is not expandable and also not supporting a large number of end-users. It also has connectivity problems because of network limitations.
3. **Three-tier architecture :** Generally a data warehouses adopts a three-tier architecture. Following are the three tiers of the data warehouse architecture which is most widely used.

It consists of the top, middle and bottom tier.

1. **Bottom Tier :** The bottom tier of the architecture is the data warehouse database server. It is the relational database system. We use the back end tools and utilities to feed data into the bottom tier. The database of the data warehouse servers as the bottom tier. It is usually a relational database system. Data is cleansed, transformed, and loaded into this layer using back-end tools. These back end tools and utilities perform the extract, clean, load, and refresh functions.
2. **Middle Tier :** The middle tier in data warehouse is an OLAP server which is implemented using either ROLAP or MOLAP model. For a user, this application tier presents an abstracted view of the database. This layer also acts as a mediator between the end-user and the database. In the middle tier, we have the OLAP server that can be implemented in either of the following ways.
   - By Relational OLAP (ROLAP), which is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.
   - By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations.
3. **Top-Tier :** The top tier is a front-end client layer. Top tier is the tools and API that you connect and get data out from the data warehouse. This layer holds the query tools and reporting tools, analysis tools and data mining tools. It could be query tools, reporting tools, managed query tools, analysis tools and data mining tools.

Scanned with CamScanner

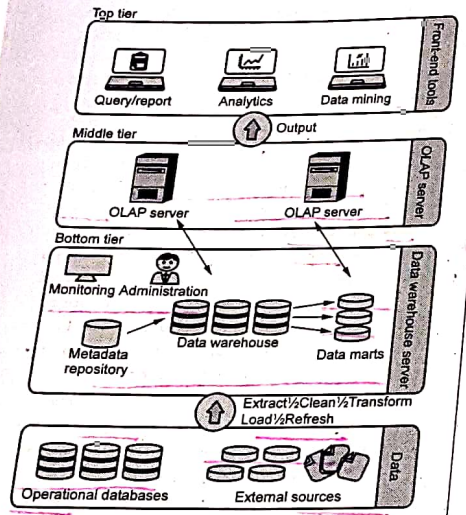


Fig. Q.6.1 Three tier architecture

Q.7 Describe problems of data warehousing.

Ans. : 1. Complexity of integration i.e. Hidden problems with source systems and data homogenization and underestimation of resources for data loading.

2. Required data not captured.
3. High maintenance
4. Long duration projects

Q.8 Why do we need data warehouses ?

Ans. : • Consolidation of information resources

• Improved query performance

• Separate research and decision support functions from the operational systems

• Foundation for data mining, data visualization, advanced reporting and OLAP tools

Q.9 Describe various methods used for cleaning data before it is loaded in the data warehouse.

[RGPV : June-15, Marks 7]

Ans. : • Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data

• The various methods for handling the problem of missing values in data tuples are as follows :

1. **Ignoring the tuple :** This is usually done when the class label is missing. This method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

2. **Manually filling in the missing value :** This approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined.

3. **Using a global constant to fill in the missing value :** Replace all missing attribute values by the same constant, such as a label like "Unknown," or $-\infty$. If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common that of "Unknown." Hence, although this method is simple, it is not recommended.

4. **Using a measure of central tendency for the attribute,** such as the mean (for symmetric numeric data), the median (for asymmetric numeric data), or the mode (for nominal data):

5. **Using the attribute mean for numeric values or attribute mode nominal values,** for all samples belonging to the same class as the given tuple.

6. **Using the most probable value to fill in the missing value :** This may be determined with regression, inference-based tools using Bayesian formalism, or decision tree induction.

Q.10 Why data warehouse is difficult to build ?

Ans. : Reasons are as follows:

1. Tendency of database to grow very large
2. Use of historical data
3. Heterogeneity of data source, which affects data conversion, quality and timeliness.

Q.11 Explain the following :
1. Operational Data Store 2. Data Mart 3. Meta Data

Ans.: 1. Operational data store (ODS) :

• An operational data store stores data for a specific application. It feeds the data warehouse a stream of desired raw data.

• It is the most common component of data warehouse environment.

• Data store is generally subject oriented, volatile, current commonly focused on customers, products, orders, policies, claims, etc.

• Its day-to-day function is to store the data for a single specific set of operational application.

• Its function is to feed the data warehouse data for the purpose of analysis.

2. Data Mart :

• A data mart is a lower-cost, scaled-down version of a data warehouse, usually designed to support a small group of users.

• Data Mart offer a targeted and less costly method of gaining the advantages associated with data warehousing and can be scaled up to a full DW environment over time.

3. Meta Data :

• The metadata is information that is kept about the warehouse. It is last component of DW environments.

• It is information that is kept about the warehouse rather than information kept within the warehouse.

• Legacy systems generally don't keep a record of characteristics of the data.

• The metadata is simply data about data.

• Metadata are abstractions - they are high level data that provide concise descriptions of lower-level data.

• The metadata are essential ingredients in the transformation of raw data into knowledge. They are the "keys" that allow us to handle the raw data.

Q.12 What is data cleaning ? How to handle noisy data in data mining ?

Ans. : Data cleaning : Raw data may have incomplete records, noise value, outliers and inconsistent data. Data cleaning is a first step in data preprocessing techniques which is used to find the missing value, smooth noise data, recognize outliers and correct inconsistent.

• These dirty data will effects on miming procedure and led to unreliable and poor output. Therefore it is important for some data cleaning routines.

How to handle noisy data in data mining ?

• Following methods are used for handling noisy data :

1. **Binning method :** First sort data and partition into (equi-depth) bins then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

2. **Clustering :** Detect and remove outliers.

3. **Combined computer and human inspection :** Detect suspicious values and check by human.

4. **Regression :** Smooth by fitting the data into regression functions.

1.2 : Design of Data Warehouse

Q.13 What is Top-down and Bottom-up approach ?

Ans. : • In Top-down approach, an organization has developed an enterprise data model, collected enterprise-wide business requirements, and decided to build an enterprise data warehouse with subset data marts.

• The Bottom-up approach, implying that the business priorities resulted in developing individual data marts, which are then integrated into enterprise data warehouse.

Q.14 List and explain technical issue while designing and implementing a data warehouse environment.

Ans. : • A number of technical issues are to be considered when designing and implementing a Data Warehouse environment.

1. The Hardware Platform that would house the Data Warehouse for parallel query scalability.
2. The DBMS that supports the warehouse database

Scanned with CamScanner

3. The communication infrastructure that connects the warehouse, data marts, operational systems, and end users
4. The hardware platform and software to support the metadata repository
5. The systems management framework that enables centralized management and administration to the entire environment.

Q.15 What are the user sophistication levels ?

Ans. : • The users can be classified on the basis of their skill level in accessing the warehouse :

1. Casual Users: These users are most comfortable retrieving information from the warehouse in pre-defined formats, and running preexisting queries and reports.
2. Power Users: These users need access tools that combine the simplicity of pre-defined queries and reports with a certain degree of flexibility.
3. Experts: These users tend to create their own queries and perform sophisticated analysis on the information they retrieve from the warehouse.

Q.16 Why not perform on-line analytical processing directly on such databases instead of spending additional time and resources to construct a separate data warehouse ?"

Ans. : • A major reason for such a separation is to help promote the high performance of both systems.

• An operational database is designed and tuned from known tasks and workloads, such as indexing and hashing using primary keys, searching for particular records, and optimizing "canned" queries.

• On the other hand, data warehouse queries are often complex. They involve the computation of large groups of data at summarized levels, and may require the use of special data organization, access, and implementation methods based on multidimensional views.

• Processing OLAP queries in operational databases would substantially degrade the performance of operational tasks.

• An operational database supports the concurrent processing of multiple transactions. Concurrency control and recovery mechanisms, such as locking and logging, are required to ensure the consistency and robustness of transactions.

Q.17 Discuss system development life cycle of a data warehouse. What factors should be considered while designing a data warehouse. [RGPV : June-14, marks 7]

Ans. : System development life cycle of a data warehouse :

• **Requirement gathering** can happen as one-to-one meetings with end-users or client. The meeting will include the business analyst and may include development team members such as architects.

• **User requirement analysis :** This process affects almost every decision throughout design and implementation of data warehouse project. User requirement analysis in data warehouse approach can fall within four categories: Data-driven, User-driven, Goal-driven and Mixed-driven approaches.

• **Prepare High Level and Low Level design documents :** Generally, High Level Design (HLD) is prepared by the architects. developers will prepare the Low Level Design (LLD) based on HLD.

• **Data Warehouse Development Process :** Once Low level design is implemented, the next step is the building data warehouse modules i.e. actual development. There are two steps in the development phase: ETL and Report.

• **Testing :** Testing is very import process of building data warehouse systems to make them work correctly and efficiently. There are three basic levels of testing performed on a data warehouse: Unit testing, Integration testing and System testing

• Following factors should be considered while **designing a data warehouse :**

1. Data Quality : In a data warehouse, data is coming from many disparate sources from all facets of an organization. When a data warehouse tries to combine inconsistent data from disparate sources, it encounters errors. Poor data quality results in faulty reporting and analytics necessary for optimal decision making.
2. Understanding Analytics: The powerful analytics tools and reports available through integrated data will provide credit union leaders with the ability to make precise decisions that impact the future success of their organizations.

3. Quality Assurance : The end user of a data warehouse is using Big Data reporting and analytics to make the best decisions possible.
4. Performance : A data warehouse must also be carefully designed to meet overall performance requirements.
5. Designing the Data Warehouse : the data warehouse is inadequate for the end user, there is a need for fixes and improvements immediately after initial delivery.
6. User Acceptance : People are not keen to changing their daily routine especially if the new process is not intuitive.
7. Cost : system should be cost effective.

Q.18 What are the different types of metadata is used in data warehouse ?

Ans.: • Metadata in a data warehouse fall into three major categories :

1. Operational Metadata
2. Extraction and Transformation Metadata
3. End-User Metadata

Operational Metadata :

• Data for the data warehouse comes from several operational systems of the enterprise.

• These source systems contain different data structures.

• The data elements selected for the data warehouse have various field lengths and data types.

• In selecting data from the source systems for the data warehouse, you split records, combine parts of records from different source files, and deal with multiple coding schemes and field lengths.

• When you deliver information to the end-users, you must be able to tie that back to the original source data sets.

• Operational metadata contain all of this information about the operational data sources.

Extraction and Transformation Metadata :

• Extraction and transformation metadata contain data about the extraction of data from the source systems, namely, the extraction frequencies, extraction methods, and business rules for the data extraction.

• This category of metadata contains information about all the data transformations that take place in the data staging area.

End-User Metadata :

• The end-user metadata is the navigational map of the data warehouse. It enables the end-users to find information from the data warehouse.

• The end-user metadata allows the end-users to use their own business terminology and look for information in those ways in which they normally think of the business.

1.3 : Data Marts

Q.19 What is a data mart ? Write the difference between Independents, dependent and distributed data marts. [RGPV : June-15, Marks 7]

Ans. : • A data mart is a lower-cost, scaled-down version of a data warehouse, usually designed to support a small group of users.

• Data Mart offer a targeted and less costly method of gaining the advantages associated with data warehousing and can be scaled up to a full DW environment over time.

• Difference between Independents, dependent and distributed data marts are as follows :

1. Independent data marts

• Independent data marts are isolated entities, entirely separate from the enterprise data warehouse.

• Their data derives from independent sources and they should be viewed as data pirates in the context of the enterprise data warehouse because their independent inputs, which are entirely separate from the enterprise data warehouse, have a high likelihood of producing data that does not match that of the warehouse.

• These independent data marts are sometimes referred to as and Teradata strongly discourages their use.

• This method of storing data is essentially the same as the mix of paper databases and mixed

Scanned with CamScanner

hierarchical and relational online databases spread among multiple departments that supports many businesses today. It is the sort of situation that businesses generally want to escape, not automate.

2. Dependent data marts

- Dependent data marts are derived from the enterprise data warehouse. Depending on how a dependent data mart is configured, it might or might not be useful.
- The recommended process uses only data that is derived from the enterprise data warehouse data store and also permits its users to have full access to the enterprise data store when the need to investigate more enterprise-wide issues arises.
- Dependent data marts can be built in one of two ways: either where a user can access both the data mart and the complete data warehouse, depending on need, or where access is limited exclusively to the data mart

3. Distributed data marts

- In dealing with distributed data warehouses we may apply the relational theory of distribution design, i.e. we first fragment the schema, then allocate the fragments to the nodes of a network.
- This allocation may replicate fragments, i.e. store the same fragment at more than one node. The goal of the distribution is to optimise global performance.

Q.20 List the benefits of data marts.

Ans. : Benefits of data marts :

- Allows the data to be accessed in lesser time
- Cost-efficient alternative to the bulky data warehouse
- Easy to use as designed according to the needs of specific user group
- Fastens the business processes.

Q.21 What are the reasons for creating a data mart ?

Ans. : Reasons are as follows :

a) Easy access to frequently needed data

b) Creates collective view by a group of users

c) Improves end-user response time

d) Ease of creation

e) Lower cost than implementing a full data warehouse

f) Potential users are more clearly defined than in a full data warehouse

g) Contains only business essential data and is less cluttered.

Q.22 Difference between data mart and data warehouse.

Ans. :

Data mart	Data warehouse
Often holds only one subject area	Holds multiple subject areas
May hold more summarized data	Holds very detailed information
Concentrates on integrating information from a given subject area or set of source systems	Works to integrate all data sources
Is built focused on a dimensional model using a star schema.	Does not necessarily use a dimensional model but feeds dimensional models

1.4 : Conceptual Modeling of Data Warehouses

Q.23 What is conceptual data modelling ?

Ans. : The conceptual model is also known as the data model that can be used to describe the conceptual schema when a database system is implemented. It hides the internal details of physical storage and targets on describing entities, data type, relationships and constraints.

Q.24 What is data warehouse modeling ?

Ans. : A data model is a graphical view of data created for analysis and design purposes. Data modeling includes designing data warehouse databases in detail, it follows principles and patterns established in Architecture for Data Warehousing and Business Intelligence.

Q.25 Explain the snowflake schema with the help of examples. [RGPV : June-14, Marks 7]

Ans. : • Snowflake schema : A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake.

- It is a variation of star schema, in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them.
- It normalizes dimensions to eliminate redundancy. The dimension data has been grouped into multiple tables instead of one large table to saves space, but this lead to more complex queries and reduced query performance.
- It is useful when there are low cardinality attributes in the dimensions
- The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space
- **Characteristics of Snowflake Schema :**

1. The main benefit of the snowflake schema it uses smaller disk space.
2. Easier to implement a dimension is added to the Schema
3. Due to multiple tables query performance is reduced
4. The primary challenge that you will face while using the snowflake Schema is that you need to perform more maintenance efforts because of the more lookup tables.

Q.26 Explain OLAP operations in multidimensional data model. Differentiate ROLAP, MOLAP and HOLAP. [RGPV : June-15, marks 7]

Ans. : • Star schema : A fact table in the middle connected to a set of dimension tables.

- Star schema consists of a fact table with a single table for each dimension.

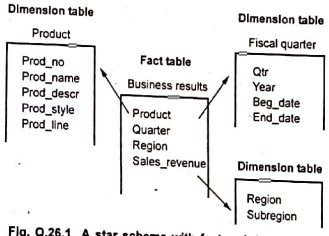


Fig. Q.26.1 A star schema with fact and dimensional tables

- **Snowflake schema :** A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake.
- It is a variation of star schema, in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them.
- The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space.

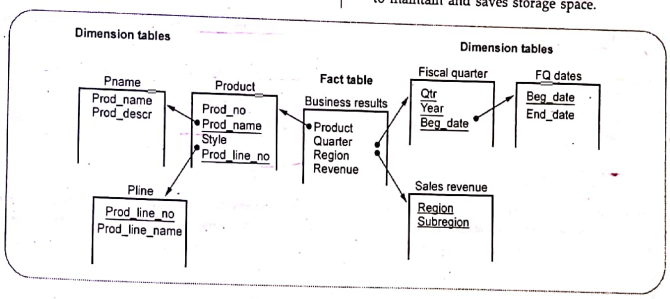


Fig. Q.25.1 Snowflake schema

Scanned with CamScanner

Q.27 Explain and compare the following by suitable example. **[RGPV : June-15, Marks 7]**
i)Fact constellations ii) Star schema

Ans. : i) Fact constellation

- The fact constellation architecture contains multiple fact tables that share many dimension tables.
- It is possible to construct fact constellation schema by splitting the original star schema into more star schemes each of them describes facts on another level of dimension hierarchies.
- The dimensions in this schema are large. They must be split into independent dimensions based on the levels of hierarchy.
- It is used mainly for the aggregate fact tables and for better understands.
- It is possible to create fact constellation schema by splitting original star schema into more star schema. It has many fact tables and some common dimension table

ii) Star schema

- Each dimension in a star schema is represented with only one-dimension table. This dimension table contains the set of attributes
- Every fact points to one tuple in each of the dimensions and has additional attributes
- Does not capture hierarchies directly.
- Usually the fact tables in a star schema are in third normal form, whereas dimensional tables are de-normalized

Q.28 Compare star schema and snowflake schema.

Ans. :

Star schema	Snowflake schema
Star schema is the simple and common modelling paradigm where the data warehouse comprises of a fact table with a single table for each dimension	Snowflake schema is the kind of the star schema which includes the hierarchical form of dimensional tables
Star schema does not use normalization	This schema uses normalization to eliminate redundancy of data
It contains fact and dimension tables	It contains sub-dimension tables including fact and dimension tables.
Simple to understand and easily designed	Hard to understand and design
High level of data redundancy	Low level of data redundancy
Cube processing is faster	Cube processing might be slow because of the complex join
It uses more space	It uses less space

Q.29 Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. Draw a stars schema diagram for the data warehouse.

Ans. : Star schema diagram :

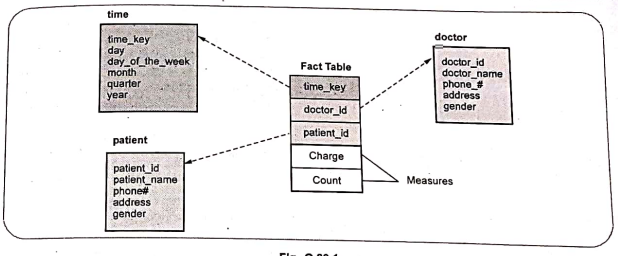


Fig. Q.29.1

Q.30 Suppose that a data warehouse for Big-University consists of the following four dimensions : student, course, semester, and instructor, and two measures count and avg. grade. When at the lowest conceptual level the avg grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination. Draw snowflakes schema diagram.

Ans. : Snowflakes schema diagram :

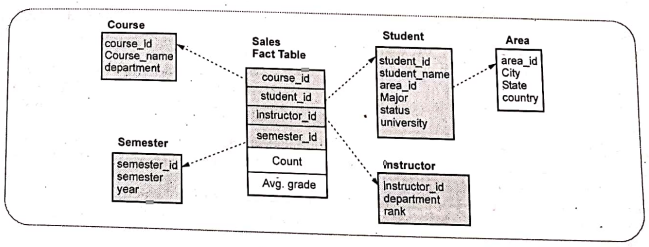


Fig. Q.30.1

1.5 Multidimensional Data Model and Aggregates

Q.31 Explain the concept that metadata is like a new center. Describe how the concept applies to the data warehouse environment. **[RGPV : June-14, marks 7]**

Ans. :

- The metadata is information that is kept about the warehouse. It is last component of DW environments.
- It is information that is kept about the warehouse rather than information kept within the warehouse.
- Legacy systems generally don't keep a record of characteristics of the data.
- The metadata is simply data about data.

Scanned with CamScanner

- Metadata are abstractions – they are high level data that provide concise descriptions of lower-level data.
- The metadata are essential ingredients in the transformation of raw data into knowledge. They are the "keys" that allow us to handle the raw data.
- Metadata in a data warehouse fall into three major categories:
  1. Operational Metadata
  2. Extraction and Transformation Metadata
  3. End-User Metadata

Operational Metadata

- Data for the data warehouse comes from several operational systems of the enterprise.
- These source systems contain different data structures.
- The data elements selected for the data warehouse have various field lengths and data types.
- In selecting data from the source systems for the data warehouse, you split records, combine parts of records from different source files, and deal with multiple coding schemes and field lengths.
- When you deliver information to the end-users, you must be able to tie that back to the original source data sets.
- Operational metadata contain all of this information about the operational data sources.

Extraction and Transformation Metadata

- Extraction and transformation metadata contain data about the extraction of data from the source systems, namely, the extraction frequencies, extraction methods, and business rules for the data extraction.
- This category of metadata contains information about all the data transformations that take place in the data staging area.

End-User Metadata

1. The end-user metadata is the navigational map of the data warehouse. It enables the end-users to find information from the data warehouse.
2. The end-user metadata allows the end-users to use their own business terminology and look for information in those ways in which they normally think of the business.

- All the metadata needed to get data into a data-staging area and prepare it for loading into dimensional model. In back room of the data warehouse there are different processes that runs, the data has to go through extraction, transformation process, and then the data needs to be loaded into the dimensional data model. So during this process the metadata is attached to all the individual processes and also includes the metadata

Important Points to Remember

- Data warehousing is the process of constructing and using a data warehouse.
- In Top-down Approach, an organization has developed an enterprise data model, collected enterprise-wide business requirements, and decided to build an enterprise data warehouse with subset data marts.
- The Bottom-up Approach, implying that the business priorities resulted in developing individual data marts, which are then integrated into enterprise data warehouse.
- A data mart is a lower-cost, scaled-down version of a data warehouse, usually designed to support a small group of users.
- Dependent data marts are derived from the enterprise data warehouse. Depending on how a dependent data mart is configured, it might or might not be useful.
- Independent data marts are isolated entities, entirely separate from the enterprise data warehouse.
- Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake.
- The fact constellation architecture contains multiple fact tables that share many dimension tables.
- The metadata is information that is kept about the warehouse. It is last component of DW environments.
- Metadata in a data warehouse fall into three major categories: Operational , Extraction and Transformation and End-User Metadata.

END...

UNIT - II

2 On-Line Analytical Processing (OLAP)

2.1 : OLAP

Q.1 What is Online Analytical Processing ?

Ans. : OLAP is a category of software that allows users to analyze information from multiple database systems at the same time. It is a technology that enables analysts to extract and view business data from different points of view.

Q.2 List three basic operations of OLAP.

Ans. : Three basic operations are Roll-up (Consolidation), Drill-down, Slicing and dicing.

Q.3 Define OLAP cube.

Ans. : The OLAP cube is a data structure optimized for very quick data analysis. The OLAP Cube consists of numeric facts called measures which are categorized by dimensions. OLAP Cube is also called the hypercube.

Q.4 Why run OLAP queries over data warehouse ?

Ans. : • Warehouse collects and combines data from multiple sources. Warehouse may organize the data in certain formats to support OLAP queries

- OLAP queries are complex and touch large amounts of data. They may lock the database for long periods of time. They negatively affects all other OLTP transactions.

Q.5 What is OLAP ? Explain different types of OLAP.

Ans. : • OLAP designates a category of applications and technologies that allow the collection, storage, manipulation and reproduction of multidimensional data, with the goal of analysis.

- OLAP allows users to analyze database information from multiple database systems at one time.
- OLAP data is stored in multidimensional databases. OLAP processing is often used for data mining.
- OLAP products are typically designed for multiple?user environments, with the cost of the software based on the number of users.
- An OLAP Cube is a data structure that allows fast analysis of data. The arrangement of data into cubes overcomes a limitation of relational databases.
- It consists of numeric facts called measures which are categorized by dimensions.
- The OLAP cube consists of numeric facts called measures which are categorized by dimensions.
- Types of OLAP :
  1. Multidimensional OLAP (MOLAP)
  2. Relational OLAP (ROLAP)
  3. Hybrid OLAP (HOLAP)

Multidimensional OLAP (MOLAP) :

- MOLAP is the 'classic' form of OLAP and is sometimes referred to as just OLAP.
- MOLAP stores this data in an optimized multi-dimensional array storage, rather than in a relational database. Therefore it requires the pre-computation and storage of information in the cube - the operation known as processing.
- MOLAP tools generally utilize a pre-calculated data set referred to as a data cube.
- The data cube contains all the possible answers to a given range of questions.
- MOLAP tools have a very fast response time and the ability to quickly write back data into the data set.

TECHNICAL™

Scanned with CamScanner

Relational OLAP (ROLAP)

- ROLAP is a kind of online analytical processing (OLAP) that analyzes data using multidimensional data models
- ROLAP can handle large volumes of data.
- Although its use of a relational database means that it requires more processing time, it can be accessed by any SQL tool, and it does not have to be a tool specifically for OLAPs.
- Compared to MOLAP, ROLAP tools are much better at controlling non-aggregated facts such as textual descriptions.

Hybrid OLAP (HOLAP) :

- There is no clear agreement across the industry as to what constitutes Hybrid OLAP, except that a database will divide data between relational and specialized storage.
- For example, HOLAP database will use relational tables to hold the larger quantities of detailed data, and use specialized storage for at least some aspects of the smaller quantities of more-aggregate or less-detailed data.
- HOLAP addresses the shortcomings of MOLAP and ROLAP by combining the capabilities of both approaches.
- HOLAP tools can utilize both pre-calculated cubes and relational data sources

Q.6 Explain typical OLAP operations in brief.
[RGPV : June-14, Marks 7]

Ans. : i) Drill Up

- It is also called roll-up operation. It performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction
- The roll-up operation shown aggregates the data by ascending the location hierarchy from the level of city to the level of country.
- When roll-up is performed by dimension reduction, one or more dimensions are removed from the given cube.
- For example, consider a sales data cube containing only the two dimensions location and time. Roll-up may be performed by removing, say, the time dimension, resulting in an aggregation of the total sales by location, rather than by location and by time.

ii) Slice and Dice

- The slice operation performs a selection on one dimension of the given cube, resulting in a sub-cube.
- The dice operation defines a sub-cube by performing a selection on two or more dimensions
- The slice operation produces a sliced OLAP cube by allowing the analyst to pick specific value for one of the dimensions.
- **Roll up (drill-up) :** Summarize data. By climbing up hierarchy or by dimension reduction
- **Drill down (roll down) :** Reverse of roll-up. From higher level summary to lower level summary or detailed data, or introducing new dimensions
- **Slice and dice :** Project and select
- **Pivot (rotate) :** Reorient the cube, visualization, 3D to series of 2D planes.
- **Drill across :** Involving (across) more than one fact table
- Drill Through : Through the bottom level of the cube to its back-end relational tables

Q.7 Compare OLTP and data warehousing.

Ans. :

OLTP	DATA WAREHOUSING
It supports operational data	It supports consolidation data
Purpose is to control and run fundamental business tasks	Purpose is to help with planning, problem solving, and decision support
Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Relatively standardized and simple queries and returning relatively few records	Often complex queries involving aggregations
Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Highly normalized with many tables	Typically de-normalized with fewer tables

2.2 : Multidimensional View and Data Cube

Q.8 What is a data cube ?

Ans. : A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

Q.9 Define dimension and dimension table.

Ans. : • Dimensions are the perspectives or entities with respect to which an organization wants to keep records.

- Each dimension may have a table associated with it, called a dimension table.

Q.10 What is multidimensional data model ? Discuss 2-D, 3-D and 4-D representation of data cube.
[RGPV : June-15, Marks 7]

Ans. : • Multidimensional databases helps to provide data-related answers to complex business queries quickly and accurately. Data warehouses and Online Analytical Processing (OLAP) tools are based on a multidimensional data model. OLAP in data warehousing enables users to view data from different angles and dimensions.

- The data cube summarizes the measure with respect to a set of n dimensions and provides summarizations for all subsets of them. Fig. Q.10.1 shows data cube.

Product \ Year	1999	2000	2001	2002	ALL
Chairs	25	37	89	21	172
Tables	10	30	0	45	85
Desks	56	84	9	35	184
Shelves	19	20	0	71	110
Boards	5	16	11	15	47
ALL	115	187	109	187	598

Fig. Q.10.1 Data cube

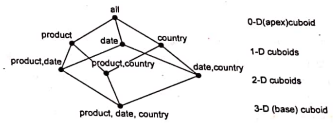
- The most detailed part of the cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.

Product \ Year	1999	2000	2001	2002	ALL
Chairs	25	37	89	21	172
Tables	10	30	0	45	85
Desks	56	84	9	35	184
Shelves	19	20	0	71	110
Boards	5	16	11	15	47
ALL	115	187	109	187	598

Base cuboid

Apex cuboid

Fig. Q.10.2 Base cuboid



Q.11 What is cuboid ? Explain any three OLAP operation on data cube with example.

Ans. : • **Cuboid :** The data cube is used by the users of the decision support system to see their data. In cube, each dimension represents some attribute in the database and cells represent some measure of interest. Users can apply queries on cubes to retrieve data.

- A data cube is a lattice of cuboids.
- In the data warehousing research literature, a data cube is often referred to as a cuboid. Given a set of dimensions, we can generate a cuboid for each of the possible subsets of the given dimensions.
- The result would form a lattice of cuboids, each showing the data at a different level of summarization, or group by. The lattice of cuboids is then referred to as a data cube.
- Also refer Q.6.

Q.12 Define data cube and explain 3 operations on it.

Ans. : • Data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts

- OLAP is based on a multidimensional data model which views data in the form of a data cube. A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions

TECHNICAL PUBLICATIONS™ I. A. Dhotre

Scanned with CamScanner

- For example, all electronics may create a sales data warehouse in order to keep records of the store's sales with respect to the dimensions time, item, branch and location.
- These dimensions allow the store to keep track of things like monthly sales of items and the branches and locations at which the items were sold.
- Each dimension may have a table associated with it, called a dimension table, which further describes the dimension.
- Also refer Q.6.

Q.13 What is fact table and dimension table ?

Ans. : • A fact table consists of facts of a particular business process e.g., sales revenue by month by product. Facts are also known as measurements or metrics. A fact table record captures a measurement or a metric.

- A fact table is used in the dimensional model in data warehouse design. A fact table is found at the center of a star schema or snowflake schema surrounded by dimension tables.
- Dimension tables are generally used to define dimensions; they include the values, dimension keys as well as attributes.
- Typically, dimension tables are small in size, ranging from a few to numerous thousand rows.

Q.14 Explain benefits of OLAP.

Ans. : • One main benefit of OLAP is consistency of information and calculations.

- "What if" scenarios are some of the most popular uses of OLAP software and are made eminently more possible by multidimensional processing.
- It allows a manager to pull down data from an OLAP database in broad or specific terms.
- OLAP creates a single platform for all the information and business needs, planning, budgeting, forecasting, reporting and analysis.

Q.15 Compare fact table and dimension table.

Ans. :

Parameters	Fact Table	Dimension Table
Basic	Fact table contains the measurement along the attributes of a dimension table	Dimension table contains the attributes along which fact table calculates the metric.
Attributes and Records	It contains less attributes and more records.	It contains more attributes and less records
Table Size	Fact table grows vertically	Dimension table grows horizontally
Key	Fact table contains a primary key which is a concatenation of primary keys of all dimension table	Each dimension table contains its primary key
Creation	Fact table can be created only when dimension tables are completed.	Dimension tables need to be created first.
Schema	A schema contains less number of fact tables	A schema contains more number of dimension tables.
Attributes	Fact table can have data in numeric as well as textual format.	Dimension table always contains attributes in textual format.

2.3 Data Cube Implementations

Q.16 How many cuboids are there in an n-dimensional data cube ?

Ans. : If there were no hierarchies associated with each dimension, then the total number of cuboids for an n-dimensional data cube is 2^n.

Q.17 Explain data cube materialization.

Ans. : 1. No materialization : Do not pre-compute any of the "non-base" cuboids. This leads to computing expensive multidimensional aggregates on the fly, which can be extremely slow.

2. Full materialization : Pre-compute all of the cuboids. The resulting lattice of computed cuboids is referred to as the full cube. This

choice typically requires huge amounts of memory space in order to store all of the pre-computed cuboids.

3. Partial materialization : Selectively compute a proper subset of the whole set of possible cuboids.

Q.18 Why most data warehouse system support index structures ? Discuss methods to index OLAP data. [RGPV : June-14, Marks 7]

Ans. : • Indexes are database objects associated with database tables and created to speed up access to data within the tables.

- Indexing techniques have already been in existence for decades for the OLTP relational database system but they cannot handle large volume of data and complex and iterative queries that are common in OLAP applications.
- Indexes are data structures which hold field values from the indexed column(s) and pointers to the related record(s). This data structure is then sorted and binary searches are performed to quickly find the record.
- Data Warehouse (DW) is a large repository of information accessed through an OLAP application. This application provides users with tools to iteratively query the DW in order to make better and faster decisions.
- The information stored in a DW is clean, static, integrated, and time varying, and is obtained through many different sources.
- Requests for information from a DW are usually complex and iterative queries. Complex queries could take several hours or days to process because the queries have to process through a large amount of data.
- So index structure is used in data warehouse.

To index OLAP data by bitmap indexing and join indexing

1. Bitmap Indexing :

- The bitmap indexing method is popular in OLAP products because it allows quick searching in data cubes.
- In the bitmap index for a given attribute, there is a distinct bit vector (Bv) for each value v in the domain of the attribute. If the domain of a given attribute consists of n values, then n bits are needed for each entry in the bitmap index.
- If the attribute has the value v for a given row in the data table, then the bit representing that value is set to 1 in the corresponding row of the bitmap index. All other bits for that row are set to 0.

2. Join Indexing :

- The join indexing method gained popularity from its use in relational database query processing.
- Join indexing registers the joinable rows of two relations from a relational database.
- The join index records can identify joinable tuples without performing costly join operations. Join indexing is especially useful for maintaining the relationship between a foreign key and its matching primary keys, from the joinable relation.

Q.19 Explain OLAP query processing.

Ans. : OLAP query processing is as follows :

1. Determine which operations should be performed on the available cuboids.
   - This involves transforming any selection, projection, roll-up and drill-down operations specified in the query into corresponding SQL and/or OLAP operations.
   - For example, slicing and dicing a data cube may correspond to selection and/or projection operations on a materialized cuboid.
2. Determine to which materialized cuboid(s) the relevant operations should be applied.
   - This involves identifying all of the materialized cuboids that may potentially be used to answer the query, pruning the above set using knowledge of "dominance" relationships among the cuboids, estimating the costs of using the remaining materialized cuboids, and selecting the cuboid with the least cost.

• June 2016 • Dec. 2016 • Dec. 2017
TECHNICAL™
Scanned with CamScanner

2.4 Guidelines for OLAP Implementation

Q.20 Briefly explain guidelines for OLAP implementation.

Ans. : OLAP implementation guidelines are as follows :

1. **Vision :** The OLAP team must, in consultation with the users, develop a clear vision for the OLAP system. This vision including the business objectives should be clearly defined, understood, and shared by the stakeholders.
2. **Senior management support :** The OLAP project should be fully supported by the senior managers.
3. **Selecting an OLAP tool :** The OLAP team should familiarize themselves with the ROLAP and MOLAP tools available in the market. Since tools are quite different, careful planning may be required in selecting a tool that is appropriate for the enterprise.
4. **Corporate strategy :** The OLAP strategy should fit in with the enterprise strategy and business objectives. A good fit will result in the OLAP tools being used more widely.
5. **Focus on the users :** The OLAP project should be focused on the users. Users should, in consultation with the technical professional, decide what tasks will be done first and what will be done later.
6. **Joint management :** The OLAP project must be managed by both the IT and business professionals.
7. **Review and adapt :** Regular reviews of the project may be required to ensure that the project is meeting the current needs of the enterprise.

2.5 : Difference between OLAP and OLTP

Q.21 Difference between OLTP and OLAP.

☞ [RGPV : June-15, Marks 7]

Ans. :

OLTP	OLAP
It support long transactions, usually complex queries	It support short transaction both query and updates
It is operational processing system	It is informational processing system
Purpose of data is to control and run fundamental business tasks	Purpose of data is to help with planning, problem solving, and decision support
Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Space requirements can be relatively small if historical data is archived	Space requirement is larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
Tables in OLTP database are normalized.	Tables in OLAP database are not normalized.
OLTP and its transactions are the original source of data	Different OLTPs database becomes the source of data for OLAP
DB size: 100 MB to GB	DB size : 100 GB to TB
It contains current data	It contains historical data.
It is application oriented	It focus on information out.

2.6 : OLAP Servers : ROLAP, MOLAP, HOLAP Queries

Q.22 Differentiate between ROLAP, MOLAP and HOLAP.

Ans. :

Parameters	ROLAP	MOLAP	HOLAP
Implementation	An implementation based on relational DBMS	An implementation based on multidimensional DBMS	An implementation using both relational and multidimensional techniques
Advantages	• It can handle large amounts of data • Can leverage functionalities inherent in the relational database	• Excellent performance • Can perform complex calculations	• HOLAP tools can utilize both pre-calculated cubes and relational data sources
Disadvantages	• Performance can be slow • Limited by SQL functionalities	• Limited in the amount of data it can handle • Requires additional investment	HOLAP supports disadvantages of MOLAP

Q.23 Explain difference between ROLAP and MOLAP.

Ans. :

Characteristics	ROLAP	MOLAP
SCHEMA	• User star Schema • Additional dimensions can be added dynamically.	• User Data cubes • Addition dimensions require recreation of data cube.
Database Size	Medium to large	Small to medium
Architecture	Client/Server	Client/Server
Access	Support ad-hoc requests	Limited to pre-defined dimensions
Speed	• Good with small data sets. • Average for medium to large data set.	• Faster for small to medium data sets. • Average for large data sets.
Flexibility and Scalability	High	Low

Q.24 List advantages and disadvantages of MOLAP and RLOAP.

Ans. : Advantages of MOLAP:

1. **Excellent performance :** MOLAP cubes are built for fast data retrieval, and are optimal for slicing and dicing operations.
2. **Can perform complex calculations :** All calculations have been pre-generated when the cube is created.

Disadvantages of MOLAP :

1. Limited in the amount of data it can handle
2. Requires additional investment: Cube technology are often proprietary and do not already exist in the organization.

Scanned with CamScanner

Advantages of ROLAP :

1. **It can handle large amounts of data :** The data size limitation of ROLAP technology is the limitation on data size of the underlying relational database. In other words, ROLAP itself places no limitation on data amount.
2. **Can** leverage functionalities inherent in the relational database: Often, relational database already comes with a host of functionalities.

Disadvantages of ROLAP :

1. Performance can be slow: Because each ROLAP report is essentially a SQL query in the relational database, the query time can be long if the underlying data size is large.

- Limited by SQL functionalities : Because ROLAP technology mainly relies on generating SQL statements to query the relational database, and SQL statements do not fit all needs.

Important Points to Remember

- OLAP is a category of software that allows users to analyze information from multiple database systems at the same time
- The OLAP cube is a data structure optimized for very quick data analysis
- Three basic operations of OLAPs are Roll-up (Consolidation), Drill-down, Slicing and Dicing
- Dimensions are the perspectives or entities with respect to which an organization wants to keep records.
- Each dimension may have a table associated with it, called a dimension table.
- The bitmap indexing method is popular in OLAP products because it allows quick searching in data cubes.
- The join indexing method gained popularity from its use in relational database query processing.

END...

UNIT - III

3 Introduction to Data Mining

3.1 Knowledge Discovery

Q.1 What is knowledge discovery in databases ? Why it is important and what are its applications ? [RGPV : June-14, Marks-7]

Ans. : With the enormous amount of data stored in files database and other repositories, it is increasingly important. If not necessary to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Datamining also popularly known as Knowledge Discovery in Database (KDD), to the nontrivial extraction of implicit. Previously unknown and potentially useful information from data in database. While data mining and Knowledge Discovery in Database (KDD) are frequently treated as synonyms, datamining is actually part of the knowledge discovery process The Fig. Q.1.1 shows iterative knowledge discovery process.

The knowledge discovery in database process comprises of a few steps learning from raw data collection to some from of new knowledge. The iterative process consist of the following steps.

- **Data cleaning :** Also known as data cleaning, it is a phase in which noise data and irrelevant data are removed from the collection.
- **Data integration :** At this stage, multiple data source, often heterogeneous, may be combined in a common source.
- **Data selection :** At this stage, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation :** Also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining :** It is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation :** In this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation :** Is the final phase in which the discovered knowledge to visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

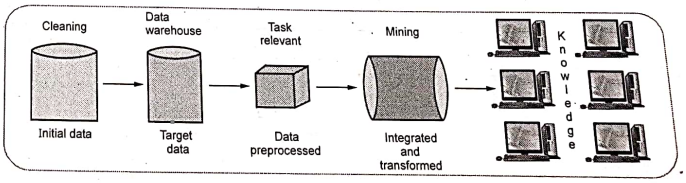


Fig. Q.1.1 KDD various model

Scanned with CamScanner

Q.2 What is the importance of visualization of discovered patterns? Explain the role of presentation in pattern visualization. Discuss various visualization techniques in KDD.

Ans. : Visualize : "To form a mental vision, image or picture of (something not visible or present to the sight or of an abstraction); to make visible to the mind or imagination."

Data visualization techniques :

1. Pixel oriented visualization techniques :

- A simple way to visualize the value of a dimension is to use a pixel where the color of the pixel reflects the dimension's value.
- For a data set of m dimensions pixel oriented techniques create m windows on the screen, one for each dimension.
- The m dimension values of a record are mapped to m pixels at the corresponding position in the windows.
- The color of the pixel reflects other corresponding values.
- Inside a window, the data values are arranged in some global order shared by all windows.
- E.g. All Electronics maintains a customer information table, which consists of 4 dimensions : Income, credit_limit, transaction_volume and age. We analyze the correlation between income and other attributes by visualization.
- We sort all customers in income in ascending order and use this order to layout the customer data in the 4 visualization windows as shown in Fig. Q.2.1.
- The pixel colors are chosen so that the smaller the value, the lighter the shading.
- Using pixel based visualization we can easily observe that credit_limit increase as income increases customer whose income is in the middle range are more likely to purchase more from All Electronics, these is no clear correlation between income and age.

2. Geometric projection visualization techniques :

- A draw back of pixel-oriented visualization techniques is that they cannot help us much in understanding the distribution of data in a multidimensional space.
- Geometric projection techniques help users find interesting projections of multidimensional data sets.
- A scatter plot displays 2-D data point using Cartesian co-ordinates. A third dimensional can be added using different colors of shapes to represent different data points.
- E.g. Where x and y are two spatial attributes and the third dimension is represented by different shapes.
- Through this visualization, we can see that points of types "+" and "X" tend to be collocated.

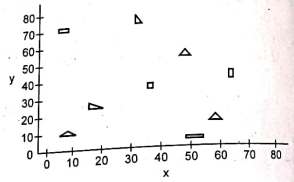


Fig. Q.2.2 Visualization of 2D data set using scatter plo

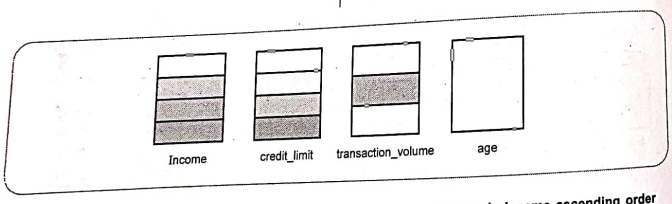


Fig. Q.2.1 Pixel oriented visualization of 4 attributes by sorting all customers in income ascending order

3. Icon based visualization techniques :

- It uses small icons to represent multidimensional data values.
- Two popular icon based techniques :

a) **Chern off faces :** They display multidimensional data of up to 18 variables as a cartoon human face.

Fig. Q.2.3 Chern off faces each face represents an 'n' dimensional data points (n < 18)

b) **Stick figures :** It maps multidimensional data to five - piece stick figure, where each figure has 4 limbs and a body.

- 2 dimensions are mapped to the display axes and the remaining dimensions are mapped to the angle and or length of the limbs.

4. **Hierarchical visualization techniques :** The subspaces are visualized in a hierarchical manner.

Q.3 Explain integrating data mining system with a database or data warehouse system. List some integration scheme.

Ans. : • The data mining system needs to be integrated with database or the data warehouse system. If the data mining system is not integrated with any database or data warehouse system then there will be no system to communicate with it.

- This scheme is known as non-coupling scheme. In this scheme the main focus is put on data mining design and for developing efficient and effective algorithms for mining the available data sets.
- The list of integration schemes :

1. **No coupling -** In this scheme the data mining system does not utilize any of the database or data warehouse functions. It then fetches the data from a particular source and process that data using some data mining algorithms. The data mining result is stored in other file.
2. **Loose coupling -** In this scheme the data mining system may use some of the functions of database and data warehouse system. It then fetches the data from data respiratory managed by these systems and perform data mining on that data. It then stores the mining result either in a file or in a designated place in a database or data warehouse.
3. **Semi-tight coupling -** In this scheme the data mining system is along with the kinking the efficient implementation of data mining primitives can be provided in database or data warehouse systems.
4. **Tight coupling -** In this coupling scheme data mining system is smoothly integrated into database or data warehouse system. The data mining subsystem is treated as one functional component of an information system.

3.2 Data Mining Functionalities

Q.4 What is data mining ? Draw and explain architecture of a typical data mining system.

Ans. : • Data mining refers to extracting or mining knowledge from large amounts of data. It is a process of discovering interesting patterns or knowledge from a large amount of data stored either in databases, data warehouses, or other information repositories.

- It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.
- Fig. Q.4.1 shows typical architecture of data mining system. (See Fig. Q.4.1 on next page)
- Components of data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.
- **Database, data warehouse, WWW, or other information repository :** This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.
- **Database or data warehouse server :** The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

• June 2016 • Dec. 2016 • Dec. 2017

Scanned with CamScanner

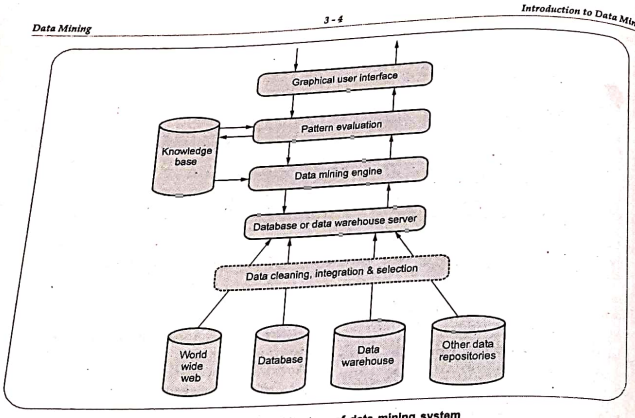


Fig. Q.4.1 typical architecture of data mining system

- *The knowledge base is helpful in the whole data mining process. It might be useful for guiding the search or evaluating the interestingness of the result patterns. The knowledge base might even contain user beliefs and data from user experiences that can be useful in the process of data mining.*
- *The **data mining engine** is the core component of any data mining system. It consists of a number of modules for performing data mining tasks including association, classification, characterization, clustering, prediction, time-series analysis etc.*
- *The **pattern evaluation module** is mainly responsible for the measure of interestingness of the pattern by using a threshold value. It interacts with the data mining engine to focus the search towards interesting patterns.*
- *The **graphical user interface** module communicates between the user and the data mining system. This module helps the user use the system easily and efficiently without knowing the real complexity behind the process.*
- When the user specifies a query or a task, this module interacts with the data mining system and displays the result in an easily understandable manner.

Q.5 What is data mining ? Discuss its various strategies. Also list out its applications.

[RGPV : June-14, Marks-7]

Ans. : Data mining strategies :

Classification	• Learning is supervised. • The dependent variable is categorical. • Well-defined classes. • Current rather than future behavior
Estimation	• Learning is supervised. • The dependent variable is numeric. • Well-defined classes. • Current rather than future behavior.
Prediction	• The emphasis is on predicting future rather than current outcomes. • The output attribute may be categorical or numeric.
• The nature of the data determines whether a model is suitable for classification, estimation, or prediction.	

Applications :

1. Data mining applications in Marketing
2. Data mining applications in HealthCare
3. Data mining applications in Retail Industry : Retail industry collects large amount of data on sales and customer shopping history. Retail data mining helps in analyzing client behavior, customer buying patterns and trends and lead to better customer service, good customer satisfaction and minimize the cost of business.
4. Data mining applications in Banking : The banking industry has hugely benefited from the advancements in digital technology. Data mining is becoming strategically important area for many business organizations including banking sector.

Q.6 What is data mining ? In your answer, address the following :

(a) Is it another hype ?

(b) Is it a simple transformation of technology developed from databases, statistics, and machine learning ?

Ans. : • Data mining refers to the process or method that extracts interesting knowledge or patterns from large amounts of data.

- It is a computational procedure of finding patterns in the bulk of data and extracting the required useful data which is further transformed in understandable information.
- In data mining, the sets of useful data are extracted involving methods like database systems, machine learning, statistics and artificial intelligence.

(a) Is it another hype ?

- Data mining is not hype. Instead, the need for data mining has arisen due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Thus, data mining can be viewed as the result of the natural evolution of information technology.

(b) Is it a simple transformation of technology developed from databases, statistics, and machine learning ?

- No. Data mining is more than a simple transformation of technology developed from databases, statistics, and machine learning.
- Instead, data mining involves an integration, rather than a simple transformation, of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing.

Q.7 What are the different functions of data mining ?

Ans. : • Different functions of data mining are characterization, association and correlation analysis, classification, prediction, clustering analysis and evolution analysis.

1. **Characterization** is a summarization of the general characteristics or features of a target class of data. For example, the characteristics of students can be produced, generating a profile of all the University in first year computing engineering students.
2. **Association** is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data
3. **Classification** differs from prediction in that the former constructs a set of models that describe and distinguish data classes, whereas the latter builds a model to predict some missing or unavailable, and often numerical, data values.
4. **Clustering** can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.
5. **Data evolution analysis** describes and models regularities for objects whose behavior changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct features of such an analysis include

Scanned with CamScanner

time-series data analysis, sequence or periodicity pattern matching.

Q.8 Describe the steps involved in data mining when viewed as a process of knowledge discovery.

Ans. : • Fig. Q.8.1 shows data mining as a step in the process of knowledge discovery.

• The steps involved in data mining when viewed as a process of knowledge discovery are as follows :

1. **Data cleaning :** A process that removes or transforms noise and inconsistent data
2. **Data integration :** Where multiple data sources 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.
5. **Data mining :** An essential process where intelligent and efficient methods are applied in order to extract patterns
6. **Pattern evaluation :** A process that identifies the truly interesting patterns representing knowledge based on some interestingness measures.
7. **Knowledge presentation :** Where visualization and knowledge representation techniques are used to present the mined knowledge to the user

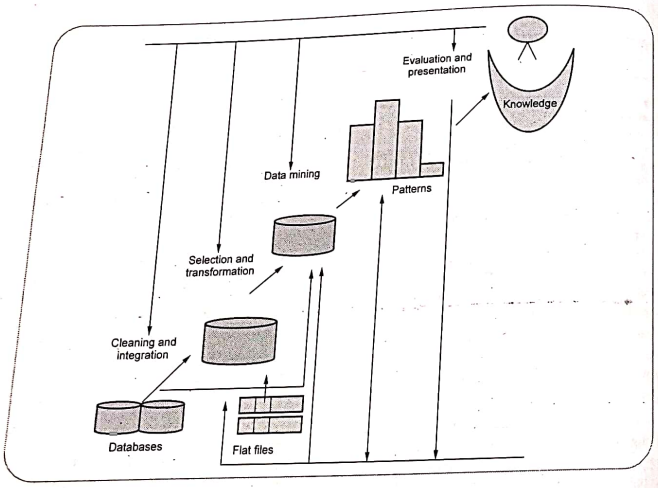


Fig. Q.8.1 Data mining as a step in the process of knowledge discovery

Q.9 Present an example where data mining is crucial to the success of a business.

Ans. : • A department store, for example, can use data mining to assist with its target marketing mail campaign.

• Using data mining functions such as association, the store can use the mined strong association rules to determine which products bought by one group of customers are likely to lead to the buying of certain other products.

• With this information, the store can then mail marketing materials only to those kinds of customers who exhibit a high likelihood of purchasing additional products.

• Data query processing is used for data or information retrieval and does not have the means for finding association rules.

• Similarly, simple statistical analysis cannot handle large amounts of data such as those of customer records in a department store.

3.3 Data Mining System Categorization and its Issues

Q.10 Explain predictive and descriptive model of data mining.

Ans. : Predictive model :

• **Predictive :** A set of techniques that analyze current and historical data to determine what is most likely to (not) happen

• It provides organizations with actionable insights based on data. It provides an estimation regarding the likelihood of a future outcome.

• To do this, a variety of techniques are used, such as machine learning, data mining, modeling and game theory.

• Predictive modeling can, for example, help to identify any risks or opportunities in the future.

• Predictive analytics can be used in all departments, from predicting customer behaviour in sales and marketing, to forecasting demand for operations or determining risk profiles for finance.

• A very well-known application of predictive analytics is credit scoring used by financial services to determine the likelihood of customers making future credit payments on time. Determining such a risk profile requires a vast amount of data, including public and social data

• Historical and transactional data are used to identify patterns, and statistical models and algorithms are used to capture relationships in various datasets.

• Predictive analytics has taken off in the big data era, and there are many tools available for organisations to predict future outcomes.

Descriptive Analysis :

• Descriptive Analytics, the conventional form of business intelligence and data analysis, seeks to provide a depiction or "summary view" of facts and figures in an understandable format, to either inform or prepare data for further analysis.

• It uses two primary techniques, namely data aggregation and data mining to report past events.

• It presents past data in an easily digestible format for the benefit of a wide business audience.

• Descriptive : A set of techniques for reviewing and examining the data set(s) to understand the data and analyze business performance.

• Descriptive analytics helps organisations understand what happened in the past. It helps to understand the relationship between customers and products and the objective is to gain an understanding of what approach to take in the future: learn from past behaviour to influence future outcomes.

• Example of descriptive analytics are company reports that simply provide a historic review of an organization's operations, sales, financials, customers, and stakeholders.

• Descriptive analytics helps to describe and present data in a format which can be easily understood by a wide variety of business readers.

Scanned with CamScanner

Q.11 Describe three challenges to data mining regarding data mining methodology.

Ans. : • Challenges to data mining regarding data mining methodology include the following:

1. mining different kinds of knowledge in databases
2. interactive mining of knowledge at multiple levels of abstraction,
3. incorporation of background knowledge

1. **Mining different kinds of knowledge in databases :** Different users are interested in different kinds of knowledge and will require a wide range of data analysis and knowledge discovery tasks such as data characterization, discrimination, association, classification, clustering, trend and deviation analysis, and similarity analysis.
   - Each of these tasks will use the same database in different ways and will require different data mining techniques.
2. **Interactive mining of knowledge at multiple levels of abstraction :** Interactive mining, with the use of OLAP operations on a data cube, allows users to focus the search for patterns, providing and refining data mining requests based on returned results.
   - The user can then interactively view the data and discover patterns at multiple granularities and from different angles.
3. **Incorporation of background knowledge :** Background knowledge, or information regarding the domain under study such as integrity constraints and deduction rules, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction.
   - This helps to focus and speed up a data mining process or judge the interestingness of discovered patterns

Q.12 Explain data mining task primitives.

Ans. : • Data mining query is used for specifying data mining task, which is input to the data mining system. Data mining query is defined in terms of data mining task primitives.

- User can communicate with the data mining system using these primitive. Data mining primitives specify the following:

1. The set of task-relevant data to be mined : This is related to the portion of database or set of data in which the user is interested.
2. The kind of knowledge to be mined : This is related to the function of data mining to be performed like characterization, discrimination, association or correlation analysis, classification, prediction, clustering etc.
3. Background knowledge to be used in the discovery process: This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found.
4. The interestingness measures and thresholds for pattern evaluation: They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns.
5. The expected representation for visualizing the discovered patterns: This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes.

Datamining task primitives	Parameters
Task-relevant data	• Database or data warehouse 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/correlation • Classification/prediction • Clustering
Background knowledge	• Concept hierarchies • User beliefs about relationships in the data

Pattern interestingness measures	• Simplicity • Certainty (e.g., confidence) • Utility (e.g., support) • Novelty
Visualization of discovered patterns.	• Rules, tables, reports, charts, graphs, decision trees, and cubes • Drill-down and roll-up

Q.13 Discuss whether or not each of the following activities is a data mining task.
i)Computing the total sales of a company.
ii)Predicting the future stock price of a company using historical records.
iii)Predicting the outcomes of tossing a pair of dice

Ans. :

i) **Computing the total sales of a company :** No. This is simple accounting.

ii) **Predicting the future stock price of a company using historical records :** Yes. We would attempt to create a model that can predict the continuous value of the stock price. This is an example of the area of data mining known as predictive modelling. We could use regression for this modelling, although researchers in many fields have developed a wide variety of techniques for predicting time series.

iii) **Predicting the outcomes of tossing a pair of dice :** No. Since the die is fair, this is a probability calculation. If the die were not fair, and we needed to estimate the probabilities of each outcome from the data, then this is more like the problems considered by data mining. However, in this specific case, solutions to this problem were developed by mathematicians a long time ago, and thus, we would not consider it to be data mining.

Q.14 Discuss whether or not each of the following activities is a data mining task.
(a)Dividing the customers of a company according to their gender.
(b)Dividing the customers of a company according to their profitability.
(c)Monitoring the heart rate of a patient for abnormalities.
(d)Monitoring seismic waves for earthquake activities.

Ans. :

(a) **Dividing the customers of a company according to their gender :** No. This is a simple database query.

(b) **Dividing the customers of a company according to their profitability :** No. This is an accounting calculation, followed by the application of a threshold. However, predicting the profitability of a new customer would be data mining.

(c) **Monitoring the heart rate of a patient for abnormalities :** Yes. We would build a model of the normal behavior of heart rate and raise an alarm when an unusual heart behavior occurred. This would involve the area of data mining known as anomaly detection. This could also be considered as a classification problem if we had examples of both normal and abnormal heart behavior.

(d) **Monitoring seismic waves for earthquake activities :** Yes. In this case, we would build a model of different types of seismic wave behavior associated with earthquake activities and raise an alarm when one of these different types of seismic activity was observed. This is an example of the area of data mining known as classification.

Q.15 Define the term data mining. Discuss the major issues in data mining.

Ans. : • Data mining refers to extracting or mining knowledge from large amounts of data. It is a process of discovering interesting patterns or knowledge from a large amount of data stored either in databases, data warehouses, or other information repositories.

- It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

Major Issues in Data Mining

- Mining methodology and user interaction
  - Mining different kinds of knowledge in database.

TECHNICAL PUBLICATIONS™

Scanned with CamScanner

- Interactive mining of knowledge at multiple levels of abstraction.
- Incorporation of background knowledge.
- Data mining query language and Ad-hoc data mining.
- Expression and visualization of data mining results.
- Handling noise and incomplete data.
- Pattern evaluation : The interesting problem.

- Performance and scalabilty.
  - Efficiency and scalabilty of data mining algorithms.
  - Parallel, distributed and incremental mining problem.
- Issue relating to the diversity of data type
  - Handling relational and complex types of data mining information form heterogeneous database and global information system(www).
- Issue related to application and social impacts
  - Application of discovered knowledge Domain specific data mining tools.
  - Intelligent query processing
  - Process control and decision making.
  - Integration of the discovered knowledge with existing knowledge : A knowledge fusion problem protection of data security, integrity and privacy.

3.4 Data Processing

Q.16 What do you mean by data preprocessing ? Why it is needed ? ☞[RGPV : June-15, Marks-7]

Ans. : • Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.

- Aim to reduce the data size, find the relation between data and normalized them.
- Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.
- Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications.
- Steps during preprocessing :
  1. **Data cleaning :** Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
  2. **Data integration :** Data with different representations are put together and conflicts within the data are resolved.
  3. **Data transformation :** Data is normalized, aggregated and generalized.
  4. **Data reduction :** This step aims to present a reduced representation of the data in a data warehouse.
  5. **Data discretization :** Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

Q.17 Briefly explain :
i)Data Cleaning ii) Data Integration ☞[RGPV : June-14, Marks-7]

Ans. : i) Data cleaning : Row data may have incomplete records, noise value, outliers and inconsistent data. Data cleaning is a first step in data preprocessing techniques which is used to find the missing value, smooth noise data, recognize outliers and correct inconsistent.

- These dirty data will effects on miming procedure and led to unreliable and poor output. Therefore it is important for some data cleaning routines.

How to handle noisy data in data mining ?

- Following methods are used for handling noisy data :
  1. **Binning method :** First sort data and partition into (equi-depth) bins then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
  2. **Clustering :** Detect and remove outliers
  3. **Combined computer and human inspection :** Detect suspicious values and check by human
  4. **Regression :** Smooth by fitting the data into regression functions

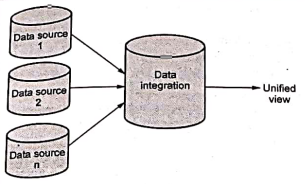


Fig. Q.17.1

i) Data integration :

Data Integration is the process of integrating data from multiple sources and probably has a single view over all these sources.

These sources may include multiple databases, data cubes, or flat files.

These sources may include multiple databases, data cubes, or flat files. One of the most well-known implementation of data integration is building an enterprise's data warehouse.

The benefit of a data warehouse enables a business to perform analyses based on the data in the data warehouse.

There are mainly 2 major approaches for data integration :

1. Tight Coupling : In tight coupling data is combined from different sources into a single physical location through the process of ETL - Extraction, Transformation and Loading.
2. Loose Coupling : In loose coupling data only remains in the actual source databases. In this approach, an interface is provided that takes query from user and transforms it in a way the source database can understand and then sends the query directly to the source databases to obtain the result.

8 Describe the issue to be considered during a integration. ☞[RGPV : June-15, Marks-7]

. : • The issue to be considered during data gration are Schema integration and object ching

- Examples of metadata for each attribute include the name, meaning, data type, and range of values permitted for the attribute, and null rules for handling blank, zero, or null values.
- Such metadata can be used to help avoid errors in schema integration. The metadata may also be used to help transform the data.
- Redundancy is another important issue. An attribute may be redundant if it can be "derived" from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.
- Some redundancies can be detected by correlation analysis. Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data.

3.5 Data Reduction & Data Mining Statistics

Q.19 What do you mean by data reduction ? What are the strategies of the data reduction ? ☞[RGPV : June-14, Marks-7]

OR Explain the various strategies of the data reduction ? ☞[RGPV : June-15, Marks-7]

Ans. : • Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.

- That is, mining on the reduced data set should be more efficient yet produce the same analytical results. Strategies for data reduction include the following :
  1. Data cube aggregation, where aggregation operations are applied to the data in the construction of a data cube.
  2. Attribute subset selection, where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.
  3. Dimensionality reduction, where encoding mechanisms are used to reduce the data set size.
  4. Numerosity reduction, where the data are replaced or estimated by alternative, smaller data representations such as parametric models

• June 2016 • Dec. 2016 • Dec. 2017

Scanned with CamScanner

or nonparametric methods such as clustering sampling, and the use of histograms

Q.20 Explain following terms : Numerosity reduction, data integration, data transformation

Ans. : • **Numerosity reduction :** The data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, and the use of histograms.

• **Data integration** combines data from multiple sources to form a coherent data store. Metadata, correlation analysis, data conflict detection, and the resolution of semantic heterogeneity contribute toward smooth data integration.

Data transformation :

• In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following :

1. Smoothing, which works to remove noise from the data. Such techniques include binning, regression, and clustering.
2. Aggregation, where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to computer monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.
3. Generalization of the data, where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higher-level concepts, like city or country. Similarly, values for numerical attributes, like age, may be mapped to higher-level concepts, like youth, middle-aged, and senior.
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.

Q.21 Explain sampling methods for data reduction.

Ans. : • Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random sample (or subset) of the data.

• Different methods of sampling are Simple Random Sampling (SRS), Stratified Sampling, Cluster Sampling, Systematic Sampling and Multistage Sampling

1. Simple random sample without replacement of size s : This is created by drawing s of the N tuples fromD (s < N), where the probability of drawing any tuple in D is 1=N, that is, all tuples are equally likely to be sampled.
2. Simple random sample with replacement (SRSWR) of size s : This is similar to SRSWOR except that each time a tuple is drawn from D it is recorded and then replaced. That is, after a tuple is drawn, it is placed back in D so that it may be drawn again.
3. Cluster sample : If the tuples in D are grouped into M mutually disjoint "clusters," then an SRS of s clusters can be obtained, where s < M. For example, tuples in a database are usually retrieved a page at a time, so that each page can be considered a cluster. For example, in a spatial database, we may choose to define clusters geographically based on how closely different areas are located.
4. Stratified sample : If D is divided into mutually disjoint parts called strata, a stratified sample of D is generated by obtaining an SRS at each stratum. This helps ensure a representative sample, especially when the data are skewed. For example, a stratified sample may be obtained from customer data, where a stratum is created for each customer age group. In this way, the age group having the smallest number of customers will be sure to be represented.

Q.22 What is outlier analysis ? Why outlier mining is important ? Briefly describe the different approaches : statistical-based outlier detection, distance-based outlier detection and deviation-based outlier detection.

Ans. : Outlier analysis is used in various types of dataset, such as graphical dataset, numerical dataset, text dataset and can also be used on the pictures etc. The identification of outlier can lead to the discovery

of useful and meaningful knowledge. Outlier detection is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outlier or anomalies. Finding outliers from as collection of pattern is a popular problem in the field of data mining. A key challenge with outlier analysis and detection is that it is not a well formulated problem like clustering so outlier detection as a branch of data mining requires more attention. Outlier analysis and detection has various applications in numerous fields such as fraud detection, credit card, discovering computer intrusion and criminal behaviors, medical and public health outlier detection, industrial damage detection. General idea of application is to find out data which deviates from normal behavior of data set.

Statistical approach :

Statistical approaches were the oldest algorithms used for outlier identification. The statistical approach assumes that data follows some standard or predefined distribution or probability model, and aims to identify outliers with respect to the model using a discordance test (those outliers which do not follows such distributions). A discordance test is used to detect whether a given object is an outliers or not. The general idea behind statistical methods for outlier detection is to learn a generative model fitting the given data set, and then identify those objects in low-probability regions of the model as outliers. Statistical methods perform poorly on high-dimensional data. Statistical methods for outlier detection can be divided into two major categories.

(a) Parametric methods : Model using parametric technique grow only with model complexity not data size. Assume that the normal data objects are generated by a parametric distribution. Regression, scatter-point method are popular parametric method. This Fig. Q.22.1 shows scatter-point method to detecting outliers.

(b) Nonparametric methods : The model of normal data is learned from the input data, rather than assuming one a priori. It do not make any assumption about statistical distribution of data. Histogram, Kernel density function. Kernel feature space are methods of non-parametric techniques we

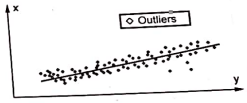


Fig. Q.22.1

can use statistical approach. Density-based approach and deviation-based approach for numerical data. Frequency based approaches have been defined to detect outliers in categorical data.

Proximity based approach : Proximity-based approaches assume that the proximity of an outlier object to its nearest neighbors significantly deviates from the proximity of the object to most of the other objects in the data set. There are three types of proximity-based outlier detection methods.

(a) Density based approach : The density-based approach of classic outlier finds the density allocation of the information and identify outliers as those present in low-density region. Partitioning methods are designed to find spherical shaped clusters but it is difficult to find clusters of arbitrary shape such as 'S' and oval density-based methods overcome this limitation.

This method compares the density around a point with its local neighbor's densities. The relative density of a point compared to its neighbors is computed as an outlier score. Density based outlier detection method uses density distribution of data points within data set. Allocate a Local Outlier Factor (LOF) to every point based on the neighboring density of its environs. LOF value of an object is based on the average of the ratios of the local reachability density of the area around the object and the local reachability densities of its neighbors. The size of the neighborhood of the object is determined by the area containing a user-supplied minimum number of points (MinPts). DBSCAN, OPTICS, DENCLUE methods are used for density-based approach. In this Fig. Q.22.2 C1, C3 stands for clusters and 0 for outliers.

(b) Distance based approach : Distance based outlier analysis is one of the most used and widely accepted technique that used in data mining and

Scanned with CamScanner

Fig. Q.22.2

machine learning. It completely depends on the concept of local neighborhood of data points. The concept is also termed as Nearest Neighbor analysis and it can also applied for different purpose such as classification, clustering and most importantly outlier analysis. In distance based approach similarity between two objects is measured with the help of distance between the two objects in data space. If this distance exceeds a particular threshold, then the data object will be called as the outlier. To measure distance any appropriate distance measure can be used such as Manhattan distance. Euclidean distance, Mahalanobis distance, or some other measure of dissimilarity. In this approach most-used algorithm is K-means algorithm which uses Eulidean distance to find outliers. K-Medoids algorithm and KNN algorithms are also types of distance based approaches.

(c) Grid based approach : Grid-based method quantizes the space into a finite number of cells which form a grid structure on which all the operations for clustering are performed. Main advantage of using the method is fastest processing time that depends only on size of grids instead of data size. For problems with highly irregular data distribution. The resolution of grid must be too fine to obtain good clustering quality CELL is a grid-based method for distance-based outlier detection. CELL method organizes objects into groups using a grid-all objects in a cell form a group. When the data set is very large so that most of the data are stored on disk, the CELL method may incur many random accesses to disk, Which is costly. An alternative method was proposed, which uses a very small amount of main memory (around 1 % of the data set) to mine all outliers by scanning the data set three times.

(d) The deviation-based approach :

- This approach identifies outliers by examining the main characteristics of objects in a group.
- Objects that "deviate" from this description are considered outliers.
- Hence, in this approach the term deviation is typically used to refer to outliers.
- There are two techniques for deviation-based outlier detection.
- The first sequentially compares objects in a set, while the second employs an OLAP data cube approach.

3.6 Guidelines for Successful Data Mining

Q.23 Briefly suggest some important guidelines for successful data mining. [RGPV : June-14, Marks-7]

Ans. : • Successful data mining involves careful determining the aims and selecting appropriate data.

- The following steps should be normally be followed :

1. Using DM guidelines in your data mining projects makes the difference between successful data mining project and disappointment
2. Make sure project stakeholders know that data mining is not a silver bullet that magically solves business problems. Data mining is a business process aided by computer support. As with any business problem, stakeholders need to find a solvable problem and work on the solution.
3. Limit the scope of your initial project: Start with limited objectives and schedules. When you achieve success, move on to more complex projects.
4. Identify a steering committee : A data mining project is a group effort. Data mining requires business users who understand the issues in the data, as well as people who understand analysis. Those who own the data will need to provide access as well

5. Avoid the data dump : Always set up the business problem, define the project goals and get support.

- Critical Success Factors (CSF) are as follows :

1) Top management support is essential;
2) Substantial group process is required;
3) Data collection is a significant problem;
4) Managers need to be continuously encouraged to consider the relevance of CSF for them;
5) The use of the factors must be pushed down through the organization;
6) Survey development will most likely be needed];
7) Develop realistic project schedules;
8) Match skills to needs at the proper time;
9) Establish and control the performance of contractors.

Important Points to Remember

- Data mining refers to extracting or mining knowledge from large amounts of data.
- Data integration is the process of integrating data from multiple sources and probably has a single view over all these sources.
- Different functions of data mining are characterization, association and correlation analysis, classification, prediction, clustering analysis and evolution analysis.
- Predictive modeling can, for example, help to identify any risks or opportunities in the future.
- Descriptive analytics helps organisations understand what happened in the past.
- Data mining query is used for specifying data mining task, which is input to the data mining system.
- Data mining query is defined in terms of data mining task primitives.
- Information allows us to expand our knowledge beyond the range of our senses.
- Knowledge is a deterministic process. When someone "memorizes" information, then they have amassed knowledge.
- Discrete attributes has a finite or countable infinite set of values.
- Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.
- *Data cleaning* routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data.
- In data transformation, the data are transformed or consolidated into forms appropriate for mining.
- Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals.

END...

TECHNICAL™

Scanned with CamScanner

UNIT - IV

4 Association Rule Mining

4.1 : The Task and a Naïve Algorithm

Q.1 How does association rules differ from traditional production rules ? Explain.

[RGPV : June-15, Marks 7]

Ans. : • Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a transactional database, relational database or other information repository.

• An example of an association rule would be "If a customer buys a dozen eggs, he is 80 % likely to also purchase milk."

• Association rule mining can be viewed as a two-step process :

1. Find all frequent item sets : By definition, each of these item sets will occur at least as frequently as a predetermined minimum support count, min sup.
2. Generate strong association rules from the frequent item sets : By definition, these rules must satisfy minimum support and minimum confidence.

• An association rule is commonly understood to be an expression of the form : $X \Rightarrow Y$ where X and Y are sets of items such that $X \cap Y = \phi$

• The association rule $X \Rightarrow Y$ means that transactions containing items from set X tend to contain items from set Y.

• Association rules show attribute value conditions that occur frequently together in a given data set. A typical example of association rule mining is Market Basket Analysis.

• Data is collected using bar-code scanners in supermarkets. Such market basket databases consist of a large number of transaction records.

• Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together.

• They could use this data for adjusting store layouts, for cross-selling, for promotions, for catalog design, and to identify customer segments based on buying patterns.

• Association rules provide information of this type in the form of if-then statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature.

• In addition to the antecedent (if) and the consequent (then), an association rule has two numbers that express the degree of uncertainty about the rule.

• In association analysis, the antecedent and consequent are sets of items (called itemsets) that are disjoint (do not have any items in common).

• The first number is called the support for the rule. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule.

• The other number is known as the confidence of the rule. Confidence is the ratio of the number of transactions that include all items in the consequent, as well as the antecedent (the support) to the number of transactions that include all items in the antecedent.

Q.2 Explain the following terms :

i) Support count **ii) Support**

iii) Frequent itemset **iv) Closed itemset**

Ans. : i) Support count : The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known, simply, as the frequency, support count or count of the itemset.

ii) Support : The support supp(X) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset.

$$\text{Supp}(X) = \frac{\text{Number of transactions which contain the itemset X}}{\text{Total number of transactions}}$$

Q.3 Explain the following terms :

i) Constraint based rule mining **ii) Closed and maximal frequent itemsets.**

Ans. : i) Constraint based rule mining :

• Constraint-based rule mining allows users to focus the search for rules by providing meta rules and additional mining constraints.

• Such mining is facilitated with the use of a declarative data mining query language and user interface, and poses great challenges for mining query optimization.

• Rule constraints can be classified into five categories: antimonotonic, monotonic, succinct, convertible, and inconvertible.

• Constraints belonging to the first four of these categories can be used during frequent itemset mining to guide the process, leading to more efficient and effective mining.

ii) Closed and maximal frequent item : Refer Q.12.

4.2 : Apriori Algorithms

Q.4 Explain apriori algorithms with example. **[RGPV : June-15, Marks 7]**

Ans. : • The Apriori algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.

• Innovative way to find association rules on large scale, allowing implication outcomes that consist of more than one item. It based on minimum support threshold.

• Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n binary attributes called items. Let $D = \{t_1, t_2, ..., t_n\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I.

• A rule is defined as an implication of the form $X \rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \phi$ The sets of items X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively.

• **Support** = The support supp (X) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset.

$$\text{Supp}(X) = \frac{\text{Number of transactions which contain the itemset X}}{\text{Total number of transactions}}$$

• Confidence : The confidence of a rule is defined.

$$\text{conf}(X \rightarrow Y) = \text{supp}(X \cup Y) / \text{supp}(X)$$

Pseudocode :

L_1 = {frequent items};

for ($k = 2$; $L_{k-1} \neq \phi$; k++) **do begin**

TECHNICAL PUBLICATIONS

Scanned with CamScanner

C_k = Candidates generated from L_{k-1} (that is : cartesian product $L_{k-1} \times L_{k-1}$ and eliminating any $k-1$ size itemset that is not frequent);

for each transaction t in database **do**

increment the count of all candidates in

C_k that are contained in t

L_k = Candidates in C_k with min_sup

end

return $\cup_k L_k$;

Q.5 What is the 'Apriori property' ? How is it used by the APRIORI algorithm ? What are the drawbacks of a Apriori algorithm. ☞[RGPV : June-14, Marks 7]

Ans. : • **Apriori property :** All nonempty subsets of a frequent itemset must also be frequent.

- Apriori property is used by using join and prune actions.

1. The join step : To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself. This set of candidates is denoted C_k. Let l_1 and l_2 be itemsets in L_{k-1}. Apriori assumes that items within a transaction or itemset are sorted in lexicographic.
2. The prune step : C_k is a superset of L_k, its members may or may not be frequent, but all of the frequent k-itemsets are included in C_k. A scan of the database to determine the count of each candidate in C_k would result in the determination of L_k.

Drawbacks of the apriori algorithm :

1. Needs several iterations of the data.
2. Uses a uniform minimum support threshold.
3. Difficulties to find rarely occuring events.
4. Alternative methods (other than appriori) can address this by using a non-uniform minimum support threshold.
5. Some competing alternative approaches focus on partition and sampling.

Q.6 Generate frequent itemsets and generate association rules based on it using aprori algorithm. Minimum support is 50 % and minimum confidence is 70 %.

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

Ans. : Apriori algorithm :

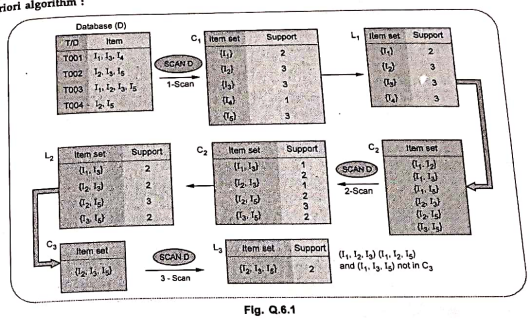


Fig. Q.6.1

Q.7 Using Apriori algorithm, generate frequent Item sets (min_sup > = 33.3 %) for the following transaction database.

Trans_id	Itemlist
T_1	{A, B, D, K}
T_2	{A, B, C, D, E}
T_3	{A, B, C, E}
T_4	{B, D}
T_5	{A, C}
T_6	{B, D}

Ans. : (See Fig. Q.7.1 on next page)

Q.8 State the Apriori property. Generate candidate itemsets, frequent itemsets and association rules using Apriori algorithm on the following data set with minimum support count is 2.

	TID	List of items_IDs
1.	T100	11, 12, 15
2.	T200	12, 14
3.	T300	12, 13
4.	T400	11, 12, 14
5.	T500	11, 13
6.	T600	12, 13
7.	T700	11, 13
8.	T800	11, 12, 13, 15
9.	T900	11, 12, 13

PUBLICATIONS
An Up-Thrust for Knowledge
Scanned with CamScanner

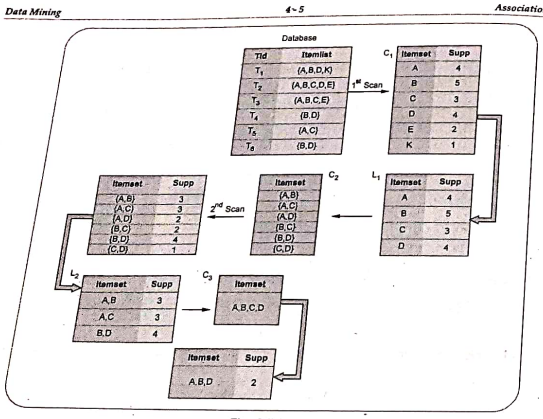
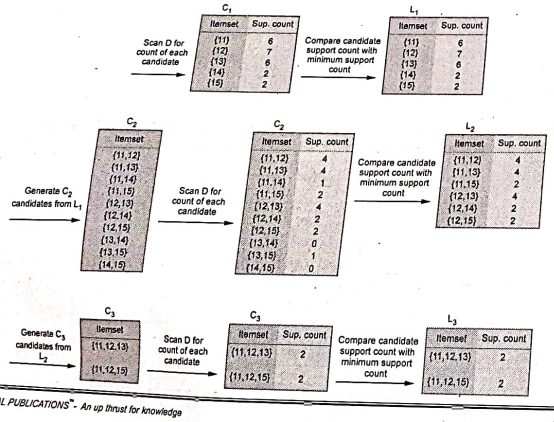


Fig. Q.7.1

Ans. :



Q.9 Explain with suitable example, the frequent item set generation in Apriori algorithm.

Ans. : • The Apriori algorithmis an influential algorithm for mining frequent itemsets for boolean association rules.

- Innovative way to find association rules on large scale, allowing implication outcomes that consist of more than one item. It based on minimum support threshold.
- Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n binary attributes called items. Let $D = \{t_1, t_2, ..., t_n\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I.
- A rule is defined as an implication of the form $X \rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \phi$. The sets of items X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively
- Apriori algorithm for finding frequent itemsets :

1. In the first iteration of the algorithm, each item is a member of the set of candidate 1-itemsets, C_1. The algorithm simply scans all of the transactions in order to count the number of occurrences of each item.
2. Suppose that the minimum support count required is 2, that is, min sup = 2. The set of frequent 1-itemsets, L_1, can then be determined. It consists of the candidate 1-itemsets satisfying minimum support.
3. To discover the set of frequent 2-itemsets, L_2, the algorithm uses the join $L_1 \times L_1$ to generate a candidate set of 2-itemsets, C_2. The C_2 consists of 2-itemsets.
4. The transactions in D are scanned and the support count of each candidate itemset in C_2 is accumulated.
5. The set of frequent 2-itemsets, L_2, is then determined, consisting of those candidate 2-itemsets in C_2 having minimum support.
6. The generation of the set of candidate 3-itemsets, $C_3 = L_2 \times L_2$
7. The transactions in D are scanned in order to determine L_3, consisting of those candidate 3-itemsets in C_3 having minimum support.
8. The algorithm uses $L_3 \times L_3$ to generate a candidate set of 4-itemsets, C_4.

Q.10 What are the limitations of the Apriori approach for mining? Briefly describe the techniques to improve the efficiency for handling this problem.

Ans. : Limitations of the Apriori approach for mining

1. Needs several iterations of the data.
2. Uses a uniform minimum support threshold.
3. Difficulties to find rarely occuring events.
4. Alternative methods (other than Apriori) can address this by using a non-uniform minimum support thresold.
5. Some competing alternative approaches focus on partition and sampling.

- Techniques for improving efficiency of Apriori algorithm are as follows :

1. **Hash-based technique :** A hash-based technique can be used to reduce the size of the candidate k-itemsets, C_k, for $k > 1$.

- For example, when scanning each transaction in the database to generate the frequent 1-itemsets, L_1, from the candidate 1-itemsets in C_1, we can generate all of the 2-itemsets for each transaction, hash them into the different buckets of a hash table structure, and increase the corresponding bucket counts.

2. **Transaction reduction :** A transaction that does not contain any frequent k-itemsets cannot contain any frequent (k+1)-item sets. Therefore, such a transaction can be marked or removed from further consideration because subsequent scans of the database for j-itemsets, where $j > k$, will not require it.
3. **Partitioning :** A partitioning technique can be used that requires just two database scans to mine the frequent itemsets.

Sampling : The basic idea of the sampling approach is to pick a random sample S of the given data D, and then search for frequent itemsets in S instead of D.

• Dec. 2013 • Dec. 2014 • June 2015 • Dec. 2015
• June 2016 • Dec. 2016 • Dec. 2017

Scanned with CamScanner

4.3 : Improving the Efficiency of the Apriori Algorithm

Q.11 Explain hash based item and partition based item.

Ans. :

1) ***Hash-based item :*** *The main drawback of frequent-pattern itemset is that they are very large in number to compute and store computer to present leads to the introduction of closed frequent itemset and maximal frequent itemset in has based technique can be applied so that it reduce the size of the candidate. K-itemset C_K for K > 1.*

For example : When we scanning each transaction in the dataset to generate the frequent 1-itemset L_1 from the candidate 1-itemset C_1 we can generate all of 2-itemset for each transaction has item into corresponding bucket count in the hash table.

*H (A, B) = ((Order of A) * 10 + Order of B)) mod 6*

B-Address	1	2	3	4	5	6
B-count	3	2	3	2	4	2
Basket	{A C}		{A C}	{A B}	{A C}	{A D}
Count	{A E}	{A D}	A B	{A B}	{A C}	{A B}
	{E E}	{A E}	{A B}		{A B}	
			{A B}		{A E}	

2) ***Transaction based :*** *Purpose of improve apriori to reduced transaction does not contain frequent K-item can not be next frequent of (K + 1). So improve Apriori algorithm.*

3) ***Partition based :*** *Partition based Apriori can reduce the number of database scan to, it divide the database into small partition such that each partition can be handled in the main memory. Let the partition P of the database $R_1, R_2 .... R_P$ can be following scan in database.*

Scan I : It to finds the local large itemset in each partition R_i. $\{A \mid A \cdot \text{count} \in S\} \cdot |R_i|$, the local large itemset can be store minimum time local memory L_i, can be found by using a level-wise algorithm such as Apriori.

Scan II : It uses the property that a large itemset in the whole database must be locally large in at least one partition of the database.

Then the union of the local large itemset found in each partition are used as the candidate and counted through the whole database to find all the large itemset.

For example :

T_{ID}	Itemset
T_1	Bread, Butter, Beer
T_2	Butter, Beer, Milk
T_3	Butter
T_4	Bread, Butter

Scan D

L^1 = {{Bread}, {Butter}, {Beer}, {Bread, Butter}, {Bread, Beer}, {Butter, Beer}, {Bread, Butter, Beer}, {Milk}, {Butter, Milk}, {Beer, Milk}, {Butter, Beer, Milk}}

L^2 = {{Butter} {Bread} {Bread, Butter}}

C = {{Bread}, {Butter}, {Beer} {Bread, Butter}, {Bread, Beer}, {Butter, Beer}, {Bread, Butter, Beer}, {Milk}, {Butter, Milk}, {Beer, Milk}, {Butter, Beer, Milk}} $\xrightarrow[\text{count support}]{\text{scan D}}$

L = {{Bread} {Butter}, {Beer}, {Bread, Butter}, {Butter, Beer}}.

Q.12 What is frequent item-set ? Explain maximal itemset and close itemset with example.

[RGPV : June-15, Marks 7]

Ans. : • A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset.

• The set {computer, antivirus software} is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset.

This is also known, simply, as the frequency, support count, or count of the itemset.

• Frequent itemsets that cannot be extended with any item without making them infrequent are called maximal frequent itemsets.

• Exact support counts of the subsets cannot be directly derived from support of the maximal frequent itemset.

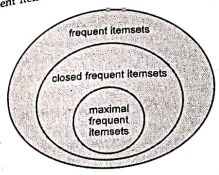


Fig. Q.12.1

Closed itemsets :

• An alternative approach is to try to retain some of the support information in the compacted representation.

• A closed itemset is an itemset whose all immediate supersets have different support count.

• A closed frequent itemset is a closed itemset that satisfies the minimum support threshold.

• Maximal frequent itemsets are closed by definition.

• An itemset X is closed in a data set S if there exists no proper super-itemset5 Y such that Y has the same support count as X in S. An itemset X is a closed frequent itemset in set S if X is both closed and frequent in S.

4.4 : Mining Frequent Patterns without Candidate Generation (FP-Growth)

Q.13 Write an algorithm for discovering itemsets without candidate algorithm.

[RGPV : June-15, Marks 7]

Ans. : Algorithm : FP growth. Mine frequent itemsets using an FP-tree by pattern fragment growth.

Input : 1. Transaction database (D);

2. The minimum support count threshold *(min sup)*

Output : The complete set of frequent patterns.

Method :

1. The FP-tree is constructed in the following steps :

a) Scan the transaction database D once. Collect F, the set of frequent items and their support counts. Sort F in support count descending order as L, the list of frequent items.

b) Create the root of an FP-tree and label it as "null." For each transaction *Trans* in D do the following.

• Select and sort the frequent items in *Trans* according to the order of L.

• Let the sorted frequent item list in *Trans* be [p | P], where p is the first element and P is the remaining list.

• Call insert tree ([p | P], T), which is performed as follows :

• If T has a child N such that N.*item-name*= p.*item-name*, then increment N's count by 1; else create a new node N, and let its count be 1, its parent link be linked to T, and its node-link to the nodes with the same *item-name* via the node-link structure.

• If P is nonempty, call insert tree(P, N) recursively.

2. The FP-tree is mined by calling FP growth(FP *tree, null*).

Q.14 Give a brief note on mining frequent patterns without candidate generation.

[RGPV : June-14, Marks 7]

Ans. : • Let P be a frequent itemset, S be the set of transactions contain P and x be an item. If x is a frequent item in S, {x} ∪ P must be a frequent itemset.

• No candidate generation. A compact data structure, FP-tree, to store information for frequent pattern mining.

• Recursive mining algorithm for mining complete set of frequent patterns.

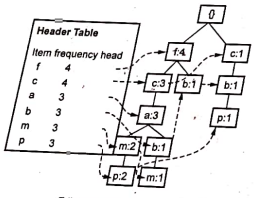
• Dec. 2013 • Dec. 2014 • June 2015 • Dec. 20
• June 2016 • Dec. 2016 • Dec. 2017

Scanned with CamScanner

Itemset	Support count
A	3
B	3
C	4
F	4
M	3
P	3

Order the items : F = 4, C = 4, A = 3, B = 3, M = 3, P = 3,

TID	Items	Rearranging
1	F, C, A, M, P	F, C, A, M, P
2	F, C, A, B, M	F, C, A, B, M
3	F, B	F, B
4	C, B, P	C, B, P
5	F, C, A, M, P	F, C, A, M, P



F-list = f-c-a-b-m-p

Fig. Q.14.1

Q.15 Write a note on FP growth algorithm to geerate frequent itemset.

Ans. : FP Growth Algorithm

Let $I = \{I_1, I_2, ...I_m\}$ be a set of item and transaction database $T = \{T_1, T_2, T_3, ... T_n\}$ then $T_i = (I \in [1...n])$ is a transaction which contains a set of items I. Support is occurrence frequency A in database, A is a frequent pattern if the support A is no less than a prede-fined minimum support call threshold (–), the problem of finding the complete set of frequent pattern is call frequent pattern mining problem.

F-P growth tree can be design based on the following observation.

1) It to be perform one scan of database to be identify the set of frequent item.
2) Store in item after frequent, this process can be avoid repeatedly scanning database.
3) Multiple transaction share an identical frequent itemset, after they can be merged into one with the number of occurrences registered as count.
4) The same node is encountered in another transaction, increment the support count of the common node or call prefix. After this process can be generated tree traversal, because point to its occurences in the tree or chain of node-link.

Construction of FP-Tree : FP-Growth Method

1) Create or consist of one root labeled as always "Null" if set of item prefix sub trees as the children of the root node, then frequent-item is in Header.
2) Each node in the prefix sub tree consist of three fields

   a) item-name b) Count c) node-link
3) Then entry in the frequent-item header table consist of two fields

   a) item-name b) Head of node-link
4) Always each item transaction are processed in L order.

FP - Growth Method : Example

T_{ID}	Item
T_{10}	A, B, E
T_{20}	B, D
T_{30}	B, C
T_{40}	A, B, D
T_{50}	A, C
T_{60}	B, C
T_{70}	A, C
T_{80}	A, B, C, E
T_{90}	A, B, C

i) Count minimum support according you but here will be minimum support 2(are required). After minimum support count database.

Item	Support
A	6
B	7
C	6
D	2
E	2

According support 2/9 = 22 %

ii) I-scan of database is same as apriori often than support count.

iii) Frequent item is sorted order of descending support.

Item	Support
B	7
A	6
C	6
D	2
E	2

iv) Then resulting L = {B = 7, A = 6, C = 6, D = 2, E = 2}

v) Then generate tree root node is "null".

vi) Each node prefix subtree consist fields item-name, supp.count, node link.

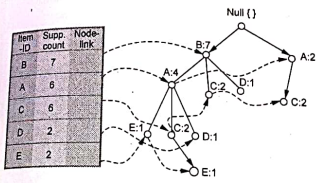


Fig. Q.15.1 FP growth tree

FP-Tree by Creating Conditional Pattern Tree

i) Start from each frequent length 1-pattern

ii) Conditional pattern based of prifix path of FP-tree is based on suffix pattern.

iii) FP-tree suffix pattern generated conditional pattern.

Item	Conditional pattern	Conditional FP-tree	Frequent pattern
E	{(B, A : 1), (BAC : 1)}	(B : 2, A : 2)	BE : 2, AE : 2, BAE : 2
D	{(BA : 1), (B : 1)}	(B : 2)	BD : 2
C	{(BA : 1), (B : 2), (A : 2)}	(B : 4, A : 2, A : 2)	BC : 4, AC : 2, BAC : 2
B	{(B : 4)}	(B : 4)	BA : 4

E as suffix base : E-conditional pattern 2 corresponding prefix path.

Fig. Q.15.2

Benefit of FP-Growth Tree

i) FP-tree avoid costly or repeated database scans.

ii) FP-tree based mining adopts a pattern-fragment growth method to avoid the costly generation of a large number of candidate sets.

iii) FP-growth eliminate repeated database scan.

iv) FP-growth is an order of faster than apriori.

v) Easy to handle large database as compared apriori.

Q.16 List advantages and disadvantages of FP growth.

Ans. : Advantages of FP-Growth

1. Only 2 passes over data-set
2. "Compresses" data-set
3. No candidate generation
4. Much faster than Apriori.

Disadvantages of FP-Growth

1. FP-Tree may not fit in memory
2. FP-Tree is expensive to build.

4.5 : Performance Evaluation of Algorithms

Q.17 Explain measures for finding rule interestingness (support, confidence) with example.

• June 2016 • Dec. 2016 • Dec. 2017

Scanned with CamScanner

Ans. : • Frequent items will play a role in the frequent-pattern mining it is necessary to perform one scan of transaction database DB to identify the set of frequent item. A compact data structure can be designed based on the following observation,.

i) Pattern a set of items/subsequences, substructures that occurs frequently in a data set.

ii) The set of frequent item of each transaction can be stored in some compact structure.

iii) Multiple transaction share a set of frequent items, it may be possible to merge the shared sets with the number of occurrences registered as count.

iv) If two transaction share a common prefix, according to some sorted order of frequent items.

v) Support and confidence.

a) Support count : The support count of an item A is denoted by A. Count in a dataset T is the number of transaction in T that contain X. Assume T has η transactions.

$$Support = \frac{(A \cup B) \cdot count}{\eta}$$

b) Confidence : The rule hold in T with confidence % of transaction that contain A also contain B.

$$Conf = P_l (B/A)$$

$$Confidence = \frac{(A \cup B) \cdot count}{A \cdot count}$$

Important Points to Remember

- A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset.
- Constraint-based rule mining allows users to focus the search for rules by providing meta rules and additional mining constraints.
- Market basket analysis is an example of frequent itemset mining.
- The purpose of market basket analysis is to determine what products customers purchase together.
- The Apriori algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.
- Techniques for improving efficiency of Apriori algorithm are Hash-based technique and transaction reduction.
- Multilevel association rules involve concepts at different levels of abstraction. Multidimensional association rules involve more than one dimension or predicate.

END...

UNIT - V

5 Classification

5.1 : Decision Tree

Q.1 What is classification ?

Ans. : Classification is a technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be "sunny", "rainy" or "cloudy".

Q.2 What is classification ? Explain.

Ans. : • **Classification** is a technique used to predict group membership for data instances. For example : You may wish to use classification to predict the day. The weather on a particular day will be "sunny", "rainy" or "cloudy".

• **Clustering** is a process of partitioning a set of data in a set of meaningful subclasses. Every data in the subclass shares a common trait. It helps a user understand the natural grouping or structure in a data set. There are two main areas of application of classification methods in information retrieval :

1. Document clustering (Relationship between document)
2. Keyword clustering (Relationship between terms)

• A computer can process words. The principle on which automated classification system is based is statistics, whereas documents can be classified on the basis of the words they contain.

• **Automatic indexing** deals with the automation of the process whereby we assign index terms to documents, that is the terms which are used later to retrieve a subject of documents from a file. **Automatic classification** is concerned with procedures and systems which can make comparison between terms used to index documents.

• **Keyword classification** is the automatic construction of classification schemes through an examination of keywords used in indexing documents. Primary objective is to derive grouping of keywords.

• Keyword classification can be used to characterize the document in a collection as an aid to retrieval. It is required as a device for providing characterizations for all the documents in a collection.

• While describing the frequency of a word the first step is that the document is translated into machine-readable code. A computer program develops a machine listing of all words in the document, arranged by frequency of occurrence and alphabetically within frequency.

Q.3 Explain decision tree method for data classification give the suitable example.

☞ [RGPV : June-15, Marks 7]

Ans. :

• A decision tree is a tree where each node represents a feature(attribute), each link(branch) represents a decision(rule) and each leaf represents an outcome(categorical or continues value).

• A decision tree or a classification tree is a tree in which each internal node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature.

• Each leaf of the tree is labeled with a class or a probability distribution over the classes.

• A decision tree is a simple representation for classifying examples

• In this method a set of training examples is broken down into smaller and smaller subsets while at the

Scanned with CamScanner

same time an associated decision tree get incrementally developed.

- At the end of the learning process, a decision tree covering the training set is returned.
- The key idea is to use a decision tree to partition the data space into cluster regions and empty regions.
- Decision tree consists of
  1. Notes : Test for the value of a certain attribute.
  2. Edges : Correspond to the outcome of a test and connect to the next node or leaf.
  3. Leaves : Terminal nodes that predict the outcome.
- A tree is defined as a set of logical conditions on attributes; a leaf represents the subset of instances corresponding to the conjunction of conditions along its branch or path back to the root.
- A simple approach to ranking is to estimate the probability of an instance's membership in a class and assign that probability as the instance's rank. A decision tree can easily be used to estimate these probabilities
- There are several steps involved in the building of a decision tree.

1. **Splitting :** The process of partitioning the data set into subsets. Splits are formed on a particular variable and in a particular location. For each split, two determinations are made : the predictor variable used for the split, called the splitting variable, and the set of values for the predictor variable, called the split point.
2. **Pruning :** The shortening of branches of the tree. Pruning is the process of reducing the size of the tree by turning some branch nodes into leaf nodes, and removing the leaf nodes under the original branch.
3. **Tree Selection :** The process of finding the smallest tree that fits the data. Usually this is the tree that yields the lowest cross-validated error.

- The Fig. Q.3.1 shows an example of a decision tree to determine what kind of contact lens a person may wear.
- The choices (classes) are none, soft and hard. The attributes that we can obtain from the person are their tear production rate (reduced or normal), whether they have astigmatism (yes/no), their age category (presbyopic or not, young or not), their spectacle prescription (myopia or hypermetropia).

Q.4 Write short note on : Decision trees rules.

☞ **[RGPV : June-14, Marks 4]**

Ans. : The decision tree learning algorithm recursively learns the tree as follows :

1. Assign all training instances to the root of the tree. Set current node to root node.

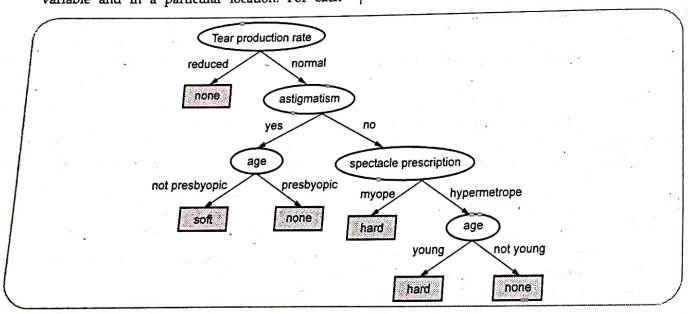


Fig. Q.3.1

2. For each attribute
   a. Partition all data instances at the node by the value of the attribute.
   b. Compute the information gain ratio from the partitioning.
3. Identify feature that results in the greatest information gain ratio. Set this feature to be the splitting criterion at the current node.
   a. If the best information gain ratio is 0, tag the current node as a leaf and return.
4. Partition all instances according to attribute value of the best feature.
5. Denote each partition as a child node of the current node.
6. For each child node :
   a. If the child node is "pure" (has instances from only one class) tag it as a leaf and return.
   b. If not set the child node as the current node and go to to step 2.

Q.5 Define the following :
a. Input variable
b. leaf node
c. Internal nodes
d. depth.

Ans. :

a. **Input variable :** Each member of the set { x_1, x_2, ... x_n} is called an input variable.

b. **Leaf node :** A node without further branches is called a leaf node. The leaf nodes return class labels and, in some implementations, they return the probability scores.

c. **Internal nodes** are the decision or test points. Each internal node refers to an input variable or an attribute. The top internal node is called the root.

d. **Depth :** The depth of a node is the minimum number of steps required to reach the node from the root.

Q.6 List the advantages and disadvantages of decision tree.

Ans. : Advantages :

1. Rules are simple and easy to understand.
2. Decision trees can handle both nominal and numerical attributes.
3. Decision trees are capable of handling datasets that may have errors.
4. Decision trees are capable of handling datasets that may have missing values.
5. Decision trees are considered to be a nonparametric method.
6. Decision trees are self-explanatory.

Disadvantages :

1. Most of the algorithms require that the target attribute will have only discrete values.
2. Some problem are difficult to solve like XOR.
3. Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
4. Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.

Q.7 Write short note on : Tree induction algorithm

☞ **[RGPV : June-14, marks-4]**

Ans. : • Decision tree induction is the learning of decision tree from class labeled training tuples.

- Following is basic algorithm for inducing a decision tree from training tuples:

Input :

- Data partition (D) which is a set of training tuples and their associated class labels;
- Attribute list, the set of candidate attributes;
- Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.

Output : A decision tree

Algorithm:

(1) create a node N;

(2) if tuples in D are all of the same class, C then

(3) return N as a leaf node labeled with the class C;

(4) if attribute list is empty then

Scanned with CamScanner

(5) return N as a leaf node labeled with the majority class in D;

(6) apply Attribute selection method(D, attribute list) to find the "best" splitting criterion;

(7) label node N with splitting criterion;

(8) if splitting attribute is discrete-valued and multiway splits allowed then

(9) attribute_list attribute_list - splitting attribute;

(10) for each outcome j of splitting criterion

(11) let Dj be the set of data tuples in D satisfying outcome j;

(12) if Dj is empty then

(13) attach a leaf labeled with the majority class in D to node N;

(14) else attach the node returned by Generate decision tree (Dj, attribute list) to node N; end for

(15) return N;

- The algorithm is called with three parameters: D, attribute list, and Attribute selection method.

5.2 : Split Algorithms Based on the Gini Index

Q.8 Write short note on : Gini Index. **[RGPV : June-15, Marks 7]**

Ans. : • One of the decision tree algorithms is CART (Classification and Regression Tree).

- Classification Tree : When decision or target variable is categorical, the decision tree is classification decision tree.
- Regression Tree : When the decision or target variable is continuous variable, the decision tree is called regression decision tree.
- CART algorithm can be used for building both Classification and Regression Decision Trees. The impurity measure used in building decision tree in CART is Gini Index. The decision tree built by CART algorithm is always a binary decision tree.
- Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.
- Gini index, entropy and twoing rule are some of the frequently used impurity measures.
- Gini Index for a given node t :

$$GINI(t) = \sum_j p(j|t)(1-p(j|t)) = \sum_j p(j|t)^2$$

Maximum of $1-1/n_c$ (number of classes) when records are equally distributed among all classes = maximal impurity.

- Minimum of 0 when all records belong to one class = complete purity.
- Entropy at a given node by :'

$$\text{Entropy}(t) = \sum_j p(j|t) \log p(j|t)$$

- Maximum ($\log n_c$) when records are equally distributed among all classes(maximal impurity).
- Minimum (0.0) when all records belongs to one class (maximal purity).
- Entropy is the only function that satisfies all of the following three properties

1. When node is pure, measure should be zero
2. When impurity is maximal (i.e. all classes equally likely), measure should be maximal
3. Measure should obey multistage property

- When a node p is split into k partitions (children) the quality of the split is computed as a weighted sum :

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i) = \sum_j p(j|t)^2$$

where n_i = number of records at child i, and n = number of records at node p.

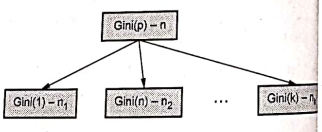


Fig. Q.8.1

- A problem with all impurity measures is that they depend only on the number of (training) patterns of different classes on either side of the hyperplane. Thus, if we change the class regions without changing the effective areas of class regions on either side of a hyperplane, the impurity measure of the hyperplane will not change.
- Thus the impurity measures do not really capture the geometric structure of class distributions. Also, all the algorithms need to optimize on some average of impurity of the child nodes and often it is not clear what kind of average is proper.

Q.9 Write short note on : Split algorithms based on the Gini Index. **[RGPV : June - 14, Marks - 3]**

Ans. :

- Used in CART, SLIQ, SPRINT.
- All attributes are assumed continuous-valued. Assume there exist several possible split values for each attribute.
- It can be modified for categorical attributes.
- The Gini index considers a binary split for each attribute. Let's first consider the case where A is a discrete-valued attribute having v distinct values $\{a_1, a_2, \ldots, a_v\}$, occurring in D.
- For example, if income has three possible values, namely {low, medium, high}, then the possible subsets are {low, medium, high}, {low, medium}, {low, high}, {medium, high}, {low}, {medium}, {high}, and { }.
- We exclude the power set, {low, medium, high} and the empty set from consideration since, conceptually, they do not represent a split. Therefore, there are $2^v - 2$ possible ways to form two partitions of the data (D) based on a binary split on A.
- When a node p is split into k partitions (children), the quality of split is computed as,

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

where, n_i = Number of records at child i,

n = Number of records at node p.

5.3 : Overfitting and Pruning

Q.10 What is decision tree pruning? Explain error based pruning and reduced error pruning method for tree pruning.

Ans. : • Pruning is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances.

- Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting
- Pruning means to change the model by deleting the child nodes of a branch node. The pruned node is regarded as a leaf node. Leaf nodes cannot be pruned.
- A decision tree consists of a root node, several branch nodes, and several leaf nodes.

1. The root node represents the top of the tree. It does not have a parent node, however, it has different child nodes.
2. Branch nodes are in the middle of the tree. A branch node has a parent node and several child nodes.
3. Leaf nodes represent the bottom of the tree. A leaf node has a parent node. It does not have child nodes.

Error based pruning :

- Error based pruning can be used to prune a decision tree and it does not require the use of validation data.
- Error-based pruning considers the E errors among the N training examples at a leaf of the tree to give an estimate of the error probability for that node.
- The assumption is that these are E events in N independent trials which is, of course, not perfectly true.
- Using the binomial theorem, condence limits can be calculated for the probability of error for a given condence level
- EBP also performs subtree raising. In the case of subtree raising, an internal node can be replaced by the subtree of one of its children rather than a leaf.

Scanned with CamScanner

- EBP will estimate that more errors are made than actually observed on the validation set for certainty factors lower than 100. Hence, if we ignore subtree raising, then EBP applied to a validation set will tend to prune more than reduced error pruning (REP) for all CF < 100 and will certainly prune more of the tree for smaller CFs.

Reduced Error Pruning

- It is simplest and most understandable method in decision tree pruning.
- This method considers each of the decision nodes in the tree to be candidates for pruning, consist of removing the subtree rooted at that node, making it a leaf node.
- The available data is divided into three parts : the training examples, the validation examples used for pruning the tree, and a set of test examples used to provide an unbiased estimate of accuracy over future unseen examples.
- If the error rate of the new tree would be equal to or smaller than that of the original tree and that subtree contains no subtree with the same property, then subtree is replaced by leaf node, means pruning is done.
- Otherwise don't prune it. The advantage of this method is its linear computational complexity.
- When the test set is much smaller than the training set, this method may lead to over pruning.

Q.11 What is information gain ? How information gain helps pruning ?

Ans. :

- Entropy measures the impurity of a collection. Information Gain is defined in terms of Entropy.
- Information gain tells us how important a given attribute of the feature vectors is.
- Information gain of attribute A is the reduction in entropy caused by partitioning the set of examples S.

$$Gain\ (S, A) \equiv Entropy\,(S) - \sum_{v \in Values} \frac{|S_v|}{|S|} Entropy\,(S_v)$$

where Values (A) is the set of all possible values for attribute A and S_v is the subset of S for which attribute A has value v

Pruning by Information Gain :

- The simplest technique is to prune out portions of the tree that result in the least information gain.
- This procedure does not require any additional data, and only bases the pruning on the information that is already computed when the tree is being built from training data.

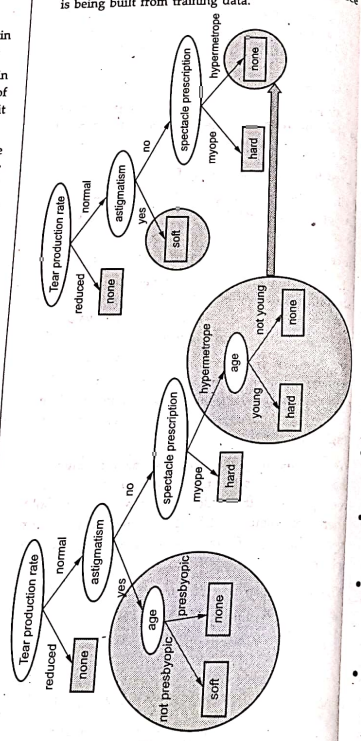


Fig. Q.11.1

- The process of information gain based pruning requires us to identify "twigs", nodes whose children are all leaves.
- "Pruning" a twig removes all of the leaves which are the children of the twig, and makes the twig a leaf. The figure below illustrates this.

The algorithm for pruning is as follows :

1. Catalog all twigs in the tree
2. Count the total number of leaves in the tree.
3. While the number of leaves in the tree exceeds the desired number :
   a. Find the twig with the least Information Gain
   b. Remove all child nodes of the twig.
   c. Relabel twig as a leaf.
   d. Update the leaf count

Q.12 Which attribute is "best" ?

Ans. : • Information gain measures how well a given attribute separates the training examples according to their target classification.

- ID3 uses this information gain measure to select among the candidate attributes at each step while growing the tree.
- In order to define information gain precisely, we use a measure commonly used in information theory, called entropy.
- Entropy characterizes the (im)purity of an arbitrary collection of examples.
- Putting together a decision tree is all a matter of choosing which attribute to test at each node in the tree.
- We shall define a measure called information gain which will be used to decide which attribute to test at each node.
- Information gain is itself calculated using a measure called entropy, which we first define for the case of a binary decision problem and then define for the general case.
- Given a binary categorization, C, and a set of examples, S, for which the proportion of examples categorized as positive by C is p+ and the proportion of examples categorized as negative by C is p-, then the entropy of S is :

$$Entropy\ (S) = -p+\log_2(p+) - p-\log_2(p-)$$

Where

S is a sample of training examples

p+ is the proportion of positive examples

p– is the proportion of negative examples

- Imagine having a set of boxes with some balls in. If all the balls were in a single box, then this would be nicely ordered, and it would be extremely easy to find a particular ball.
- If, however, the balls were distributed amongst the boxes, this would not be so nicely ordered, and it might take quite a while to find a particular ball.
- If we were going to define a measure based on this notion of purity, we would want to be able to calculate a value for each box based on the number of balls in it, then take the sum of these as the overall measure.
- We would want to reward two situations : nearly empty boxes, and boxes with nearly all the balls in. This is the basis for the general entropy measure, which is defined as follows.
- Given an arbitrary categorisation, C into categories $c_1, \ldots, c_n$, and a set of examples, S, for which the proportion of examples in c_i is p_i, then the entropy of S is :

$$Entropy\ (S) = \sum_{i=1}^{n} -p_i \log_2(p_i)$$

Q.13 How to evaluate a decision tree ?

Ans. :

- A decision tree can help you make tough choices between different paths and outcomes, but only if you evaluate the model correctly.
- Decision trees are graphic models of possible decisions and all related possible outcomes in a tree form, with the outcomes shown as "branches" off each choice. You can use a decision tree to help you make all kinds of business decisions, including new product development, new marketing strategies and workforce changes.

TECHNICAL™

Scanned with CamScanner

- Decision tree is evaluated as follows :
  1. First, evaluate whether the splits of the tree make sense. Conduct sanity checks by validating the decision rules with domain experts, and determine if the decision rules are sound.
- Next, look at the depth and nodes of the tree. Having too many layers and obtaining nodes with few members might be signs of overfitting.
- In overfitting, the model fits the training set well, but it performs poorly on the new samples in the testing set.
- Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship.
- Overfitting is when a classifier fits the training data too tightly. Such a classifier works well on the training data but not on independent test data. It is a general problem that plagues all machine learning methods.
- Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations.
- We can determine whether a predictive model is underfitting or overfitting the training data by looking at the prediction error on the training data and the evaluation data. Fig. Q.13.1 shows overfiting.

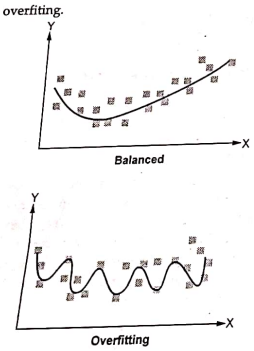


Fig. Q.13.1 Overfitting

- Reasons for overfitting :
  1. Noisy data
  2. Training set is too small
  3. Large number of features
- To prevent over-fitting we have several options :
  1. Restrict the number of adjustable parameters the network has - e.g. by reducing the number of hidden units, or by forcing connections to share the same weight values.
  2. Stop the training early , before it has had time to learn the training data too well.
  3. Add some form of regularization term to the error/cost function to encourage smoother network mappings.
  4. Add noise to the training patterns to smear out the data points.

Q.14 Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?

Ans. : • The decision tree built may overfit the training data. There could be too many branches, some of which may reflect anomalies in the training data due to noise or outliers.

- Tree pruning addresses this issue of overfitting the data by removing the least reliable branches.
- This generally results in a more compact and reliable decision tree that is faster and more accurate in its classification of data.
- The drawback of using a separate set of tuples to evaluate pruning is that it may not be representative of the training tuples used to create the original decision tree.
- If the separate set of tuples are skewed, then using them to evaluate the pruned tree would not be a good indicator of the pruned tree's classification accuracy.
- Furthermore, using a separate set of tuples to evaluate pruning means there are less tuples to use for creation and testing of the tree.
- While this is considered a drawback in machine learning, it may not be so in data mining due to the availability of larger data sets.

Q.15 Explain overfitting. What are the reason for overfitting ?

Ans. : • Training error can be reduced by making the hypothesis more sensitive to training data, but this may lead to overfitting and poor generalization.

- Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship.
- Overfitting is when a classifier fits the training data too tightly. Such a classifier works well on the training data but not on independent test data. It is a general problem that plagues all machine learning methods.
- Because of overfitting, low error on training data and high error on test data.
- Overfitting occurs when a model begins to memorize training data rather than learning to generalize from trend.
- The more difficult a criterion is to predict, the more noise exists in past information that need to be ignored. The problem is determining which part to ignore.
- Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations.

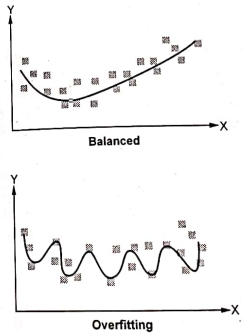


Fig. Q.15.1 Overfitting

- We can determine whether a predictive model is underfitting or overfitting the training data by looking at the prediction error on the training data and the evaluation data.
- Fig. Q.15.1 shows overfiting.
- Reasons for overfitting
  1. Noisy data 2. Training set is too small
  3. Large number of features
- Method to avoid overfittings :
  1. Limit the number of hidden nodes.
  2. Stop training early to avoid a perfect explanation of the training set, and
  3. Apply weight decay to limit the size of the weights, and thus of the function class implemented by the network.

Q.16 Explain underfitting. What are the reason for underfitting ? How do we know if we are underfitting or overfitting ?

Ans. : • **Underfitting :** If we put too few variables in the model, leaving out variables that could help explain the response, we are **underfitting**.

- **Consequences :**
  1. Fitted model is not good for prediction of new data - prediction is biased.
  2. Regression coefficients are biased.
  3. Estimate of error variance is too large.
- **Underfitting examples :**
  1. The learning time may be prohibitively large, and the learning stage was prematurely terminated.
  2. The learner did not use a sufficient number of iterations.
  3. The learner tries to fit a straight line to a training set whose examples exhibit a quadratic nature,
- To prevent under-fitting we need to make sure that :
  1. The network has enough hidden units to represent the required mappings.
  2. The network is trained for long enough that the error/cost function is sufficiently minimized.

TECHNICAL™

Scanned with CamScanner

- How do we know if we are underfitting or overfitting ?
  1. If by increasing capacity we decrease generalization error, then we are underfitting, otherwise we are overfitting.
  2. If the error in representing the training set is relatively large and the generalization error is large, then underfitting;
  3. If the error in representing the training set is relatively small and the generalization error is large, then overfitting;
  4. There are many features but relatively small training set.

Q.17 Write short note on : Overfitting and prunig.

[RGPV : June-14, Marks 4]

Ans. : Refer Q.10 and Q.15.

5.4 : Naïve Bayes Method

Q.18 Write short note on : Naïve Bayes method.

[RGPV : June-15, Marks 7]

OR What is Bay's theorem ? Explain. Also discuss Bay's classifier using some example.

Ans. : • Bayes' theorem is a method to revise the probability of an event given additional information. Baye's theorem calculates a conditional probability called a posterior or revised probability.

- Bayes' theorem is a result in probability theory that relates conditional probabilities. If A and B denote two events, $P(A|B)$ and $P(B|A)$ are in general different.
- Bayes theorem gives a relation between $P(A|B)$ and $P(B|A)$. An important application of Baye's theorem is that it gives a rule how to update or revise the strengths of evidence - based beliefs in light of new evidence a posteriori.
- A **prior probability** is an initial probability value originality obtained before any additional information is obtained.
- A **posterior probability** is a probability value that has been revised by using additional information that is later obtained.
- Suppose that $B_1, B_2, B_3 \ldots B_n$ partition the outcomes of an experiment and that A is another event. For any number, k, with $1 \le k \le n$, we have the formula :

$$P(B_k/A) = \frac{P(A/B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A/B_i) \cdot P(B_i)}$$

Naive Bayes Classifier :

- Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong independence assumptions between the features.
- It is highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.
- A Naive Bayes Classifier is a program which predicts a class value given a set of attributes.
- For each known class value,
  1. Calculate probabilities for each attribute conditional on the class value.
  2. Use the product rule to obtain a joint conditional probability for the attributes.
  3. Use Bayes rule to derive conditional probabilities for the class variable.
- Once this has been done for all class values, output the class with the highest probability.
- Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.
- The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class.

Conditional Probability :

- Let A and B be two events such that $P(A) > 0$. We denote $P(B|A)$ the probability of B given that A has occurred. Since A is known to have occurred, it

becomes the new sample space space replacing the original S. From this, the definition is,

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

OR $P(A \cap B) = P(A)\,P(B|A)$

- The notation $P(B|A)$ is read **"the probability of event B given event A"**. It is the probability of an event B is given the occurrence of the event A.
- We say that, the probability that both A and B occur is equal to the probability that A occurs times the probability that B occurs given that A has occurred. We call **$P(B|A)$ is the conditional probability of B given A,** i.e. the probability that B will occur given that A has occurred.
- Similarly, the conditional probability of an event A, given B by,

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

- The probability $P(A|B)$ simply reflects the fact that the probability of an event A may depend on a second event B. If A and B are mutually exclusive $A \cap B = \phi$ and $P(A|B) = 0$.
- Another way to look at the conditional probability formula is :

$$P(\text{Second/First}) = \frac{P(\text{First choice and second choice})}{P(\text{First choice})}$$

Q.19 At a certain university, 4 % of men are over 6 feet tall and 1% of women are over 6 feet tall. The total student population is divided in the ratio 3:2 in favour of women. If a student is selected at random from among all those over six feet tall, what is the probability that the student is a woman ?

Ans. : Let us assume following :

M = {Student is Male},

F = {Student is Female},

T = {Student is over 6 feet tall}.

Given data :

$P(M) = 2/5$,

$P(F) = 3/5$,

$P(T|M) = 4/100$

$P(T|F) = 1/100$.

We require to find $P(F|T)$?

Using Bayes' Theorem we have :

$$P(F/T) = \frac{P(T/F)\,P(F)}{P(T/F)\,P(F) + P(T/M)\,P(M)}$$

$$= \frac{\frac{1}{100} \times \frac{3}{5}}{\frac{1}{100} \times \frac{3}{5} + \frac{4}{100} \times \frac{2}{5}} = \frac{\frac{3}{500}}{\frac{3}{500} + \frac{8}{500}}$$

$$P(F/T) = \frac{3}{11}.$$

Q.20 Write short note on Bayesian belief networks.

Ans. : • Bayesian Belief Networks (BBN) are also known as belief networks, Bayesian networks, and probabilistic networks.

- A BBN is a special type of diagram (called a directed graph) together with an associated set of probability tables.
- The graph consists of nodes and arcs. The nodes represent variables, which can be discrete or continuous. The arcs represent causal relationships between variables.
- A belief network is defined by two components : Directed acyclic graph and a set of conditional probability tables
- BBNs enable us to model and reason about uncertainty. BBNs accommodate both subjective probabilities and probabilities based on objective data.
- The most important use of BBNs is in revising probabilities in the light of actual observations of events.
- Each node in the directed acyclic graph represents a random variable. The variables may be discrete or continuous-valued. They may correspond to actual attributes given in the data or to "hidden variables" believed to form a relationship.
- Each arc represents a probabilistic dependence. If an arc is drawn from a node Y to a node Z, then Y is a parent or immediate predecessor of Z, and Z is a descendant of Y. Each variable is conditionally

TECHNICAL PUBLICATIONS

Scanned with CamScanner

independent of its non-descendants in the graph, given its parents.

- Fig. Q.20.1 shows simple belief network.

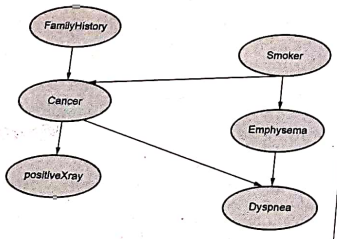


Fig. Q.20.1 Simple belief network

- Belief network has one Conditional Probability Table (CPT) for each variable.
- Probability theory is the body of knowledge that enables us to reason formally about uncertain events.
- The probability P of an uncertain event A, written P(A), is defined by the frequency of that event based on previous observations. This is called frequency based probability.
- In general, a person belief in a statement a will depend on some body of knowledge K. We write this as P(a|K). The expression P(a|K) thus represents a belief measure.
- Sometimes, for simplicity, when K remains constant we just write P(a), but you must be aware that this is a simplification.
- The notion of degree of belief P(A|K) is an uncertain event A is conditional on a body of knowledge K. In general, we write P(A|B) to rep represent a belief in A under the assumption that B is known.

5.5 : Cluster Analysis

Q.21 What is cluster ?

Ans. : Cluster is a group of objects that belong to the same class. In other words the similar object are grouped in one cluster and dissimilar are grouped in other cluster.

Q.22 Explain clustering.

Ans. : Clustering is a process of partitioning a set of data in a set of meaningful subclasses. Every data in the subclass shares a common trait. It helps a user understand the natural grouping or structure in a data set.

Q.23 What are the desirable properties of a clustering algorithm ?

Ans. :

1. Scalability (in terms of both time and space)
2. Ability to deal with different data types
3. Minimal requirements for domain knowledge to determine input parameters
4. Interpretability and usability

Q.24 What is cluster analysis ?

Ans. : Cluster analysis is process of grouping a set of data-objects into clusters.

Q.25 What is Learning ? Discuss supervised and Un supervised learning.

Ans. : • Learning is any change in a system that allows it to perform better the second time on repetition of the same task or on another task drawn from the same population. One part of learning is acquiring knowledge and new information; and the other part is problem-solving.

- Supervised learning in which the network is trained by providing it with input and matching output patterns. These input-output pairs are usually provided by an external teacher.
- Unsupervised learning algorithms aim to learn rapidly and can be used in real-time. Unsupervised learning is frequently employed for data clustering, feature extraction etc.

Q.26 Explain desired features of cluster analysis method.

Ans. : Features are as follows :

1. **Scalability :** Data-mining problems can be large and therefore a cluster-analysis method should be able to deal with large problems gracefully. Ideally, performance should be linear with data-size.
2. **Only one Scan of the Dataset :** For large problems, data must be stored on disk, so cost of I/O disk can become significant in solving the problem.
3. **Ability to Stop and Resume :** For large dataset, cluster-analysis may require huge processor-time to complete the task. In such cases, the task should be able to be stopped and then resumed when convenient.
4. **Minimal Input Parameters :** The method should not expect too much guidance from the data-mining analyst.
5. **Robustness :** Most data obtained from a variety of sources has errors. Therefore, the method should be able to deal with noise, outlier and missing values gracefully.
6. **Ability to Discover Different Cluster-Shapes :** Clusters appear in different shapes and not all clusters are spherical. So the method should be able to discover cluster-shapes other than spherical.
7. **Different Data Types :** Many problems have a mixture of data types, for e.g. numerical, categorical & even textual. Therefore, the method should be able to deal with numerical, boolean and categorical data.
8. **Result Independent of Data Input Order :** The method should not be sensitive to data input-order.

Q.27 What do you mean clustering ?

Ans. : Clustering : • Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. Clustering can be considered the most important unsupervised learning problem.

- A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Fig. Q.27.1 shows cluster.

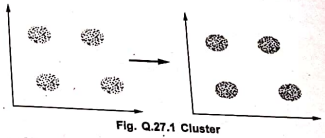


Fig. Q.27.1 Cluster

- Clustering means grouping of data or dividing a large data set into smaller data sets of some similarity.
- A clustering algorithm attempts to find natural groups of components or data based on some similarity. Also, the clustering algorithm finds the centroid of a group of data sets.
- The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering.
- Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.
- Clustering analysis helps construct meaningful partitioning of a large set of objects. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, etc.
- Clustering algorithms may be classified as listed below :
  1. Exclusive clustering
  2. Overlapping clustering
  3. Hierarchical clustering
  4. Probabilistic clustering
- A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation.

• June 2016 • Dec. 2016 • Dec. 2017

Scanned with CamScanner

- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Q.28 Write short note on clustering tree.

Ans. : • Clustering is an exploratory data analysis task. It aims to find the intrinsic structure of data by organizing data objects into similarity groups or clusters.

- Let P be a finite set of points in an om-space. If the distances between different pairs of points in P are of different orders of magnitude, then the om-space imposes a unique tree-like hierarchical structure on P.
- The points will naturally fall into clusters, each cluster C being a collection of points all of which are much closer to one another than to any point in P outside C.
- The collection of all the clusters over P forms a strict tree under the subset relation.
- Moreover, the structure of this tree and the comparative sizes of different clusters in the tree captures all of the order-of-magnitude relations between any pair of points in P.
- A cluster tree is a tree T such that :

1. Every leaf of T is a distinct symbol.
2. Every internal node of T has at least two children.
3. Each internal node of T is labelled with a non-negative value. Two or more nodes may given the same value.
4. Every leaf of the tree is labelled 0.
5. The label of every internal node in the tree is less than the label of its parent.

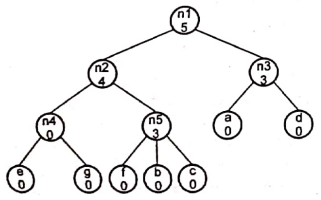
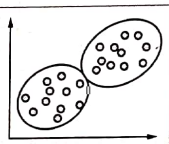
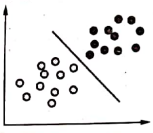


Fig. Q.28.1

- For any node N of T, the field "N.symbols" gives the set of symbols in the leaves in the subtree of T rooted at N and the field "N.label" gives the integer label on node N.
- **Cluster tree construction :** This step uses a modified decision tree algorithm with a new purity function to construct a cluster tree to capture the natural distribution of the data without making any prior assumptions.
- **Cluster tree pruning :** After the tree is built, an interactive pruning step is preformed to simplify the tree to find meaningful / useful clusters. The final clusters are expressed as a list of hyper - rectangular regions.

Q.29 Explain difference between clustering and classification.

Ans. :

Clustering	Classification
This function maps the data into one of several clusters which is the grouping of data items based on the similarities between them.	This model function classifies the data into one of several predefined categorical classes.
Involved in unsupervised learning.	Involved in supervised learning.
Training sample is not provided.	Training sample is provided.
The number of cluster is not known before clustering, These are identified after the completion of clustering.	The number of classes is known before classification as there is predefined output based on input data.
Data is not labeled.	Labeled data points.
Asks how can I group this set of items ?	Asks what class does this item belong to ?
Unknown number of classes.	Known number of classes.
Used to understand data.	Used to classify future observations.
	

Q.30 What is Hierarchical method of clustering ? Differentiate agglomerative and divisive hierarchical method of clustering.

[RGPV : June-14, Marks 7]

Ans. : • This method use distance matrix as clustering criteria. This method does not require the number of clusters K as an input, but needs a termination condition. Hierarchical clustering is a widely used data analysis tool.

- The idea is to build a binary tree of the data that successively merges similar groups of points. Visualizing this tree provides a useful summary of the data.
- Hierarchical clustering arranges items in a hierarchy with a treelike structure based on the distance or similarity between them. The graphical representation of the resulting hierarchy is a tree-structured graph called a **dendrogram.**
- Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion.

Agglomerative hierarchical clustering

- This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied. Most hierarchical clustering methods belong to this category.
- Initially, AGNES places each objects into a cluster of its own. The clusters are then merged step-by-step according to some criterion. For example, cluster C_1 and C_2 may be merged if an object in C_1 and object in C_2 form the minimum Euclidean distance between any two objects from different clusters.
- In the agglomerative hierarchical approach, we start by defining each data point to be a cluster and combine existing clusters at each step. Here are four different methods for doing this :

1. **Single linkage :** Smallest pairwise distance between elements from each cluster
2. **Complete linkage :** Largest distance between elements from each cluster
3. **Average linkage :** The average distance between elements from each cluster
4. **Centroid linkage :** Distance between cluster means

Divisive Hierarchical Clustering

- This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the clusters into smaller and smaller pieces, until each object form a cluster on its own or until it satisfies certain termination conditions, such as a desired number of cluster or the diameter of each cluster is within a certain threshold.

Agglomerative	Divisive Hierarchical Clustering
Initially each item in its own cluster.	Initially all items in one cluster.
Iteratively clusters are merged together.	Large clusters are successively divided.
Bottom up.	Top down.

Q.31 Compare KNN, Naïve Bayes and decision tree.

Ans. :

Parameter	KNN	Naive Bayes	Decision Tree
Deterministic/ Non-determin istic	Non-determin istic	Non-determin istic	Deterministic
Effectiveness on	Small data	Huge data	Large data
Speed	Slower for large data.	Faster than KNN.	Faster
Dataset	It can not deal with noisy data.	It can deal with noisy data.	It can deal with noisy data.
Accuracy	Provides high accuracy.	For obtaining good results it requires a very large number of records.	High accuracy

Scanned with CamScanner

5.6 : Types of Cluster Analysis Methods

Q.32 What is constraint-based cluster analysis ?

Ans. : *In this method the clustering is performed by incorporation of user or application oriented constraints. The constraint refers to the user expectation or the properties of desired clustering results. The constraint give us the interactive way of communication with the clustering process. The constraint can be specified by the user or the application requirement.*

Q.33 List the advantages of Grid based method.

Ans. : Advantages

1. *The major advantage of this method is fast processing time.*
2. *It is dependent only on the number of cells in each dimension in the quantized space.*

Q.34 How many types of clustering methods are there ? Explain any one partitioning clustering algorithm. **[RGPV : June-15, Marks 7]**

Ans. :

- *Clustering methods can be classified into the following categories :*

a) Partitioning method

b) Hierarchical method

c) Density-based method

d) Grid-based method

e) Model-based method

f) Constraint-based method.

Partitioning Method :

- *Given a database of "n" objects, a partitioning method constructs K partitions of the data, where each partition represents a cluster and $k \le n$.*
- *It classifies the data into K groups, which together satisfy the following requirements: each group must contain at least one object, and each object must belong to exactly one group.*
- The commonly used partitioning methods are K-means and K-medoids

K-means clustering method : K-means clustering is heuristic method. Here each cluster is represented by the center of the cluster. "K" stands for number of clusters, it is typically a user input to the algorithm; some criteria can be used to automatically estimate K.

- This method initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart.
- Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated everytime a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.
- Given K, the K-means algorithm consists of four steps :
  1. Select initial centroids at random.
  2. Assign each object to the cluster with the nearest centroid.
  3. Compute each centroid as the mean of the objects assigned to it.
  4. Repeat previous 2 steps until no change.
- The $x_1, \ldots, x_N$ are data points or vectors of observations. Each observation (vector x_i) will be assigned to one and only one cluster. The $C(i)$ denotes cluster number for the ith observation. K-means minimizes within-cluster point scatter :

$$W(C) = \frac{1}{2} \sum_{K=1}^{K} \sum_{C(i)=K} \sum_{C(j)=K} \| x_i - x_j \|^2$$

$$= \sum_{K=1}^{K} N_k \sum_{C(i)=K}^{K} \| x_i - m_K \|^2$$

Where

m_K is the mean vector of the Kth cluster.

N_K is the number of observations in K^{th} cluster.

K-Means Algorithm Properties

1. There are always K clusters.
2. There is always at least one item in each cluster.
3. The clusters are non-hierarchical and they do not overlap.

4. Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

The K-Means Algorithm Process

1. The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
2. For each data point.
   a. Calculate the distance from the data point to each cluster.
   b. If the data point is closest to its own cluster, leave it where it is.
   c. If the data point is not closest to its own cluster, move it into the closest cluster.
3. Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
4. The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion.

- K-means algorithm is iterative in nature. It converges, however only a local minimum is obtained. It works only for numerical data. This method easy to implement.

• Advantages of K-Means Algorithm :

1. Efficient in computation 2. Easy to implement

• Weaknesses

1. Applicable only when *mean* is defined.
2. Need to specify K, the *number* of clusters, in advance.
3. Trouble with noisy data and *outliers*.
4. Not suitable to discover clusters with *non-convex shapes*.

Q.35 Write short note on : density-based methods. **[RGPV : June-15, Marks 7]**

Ans. : • A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.

- Typically, for each data-point in a cluster, at least a minimum number of points must exist within a given radius. Data that is not within such high-density clusters is regarded as outliers or noise.
- A density-based cluster is a set of density-connected objects that is maximal with respect to density-reachability. Every object not contained in any cluster is considered to be noise.
- DBSCAN is a density based clustering algorithm that works with a number of different distance metrics. When DBSCAN has processed a set of data points, a point will either be in a cluster or will be classified as noise.
- DBSCAN is based on the concepts of a point being "density reachable" and "density connected." These concepts are explained in detail in.
- Data points fall into three classes :

  a) Core points : These are points that are at the interior of a cluster.

  b) Border points : A border point is a point that is not a core point, i.e., there are not enough points in its neighborhood, but it falls within the neighborhood of a core point.

  c) Noise points : A noise point is any point that is not a core point or a border point.
- DBSCAN requires two input parameters : Size of the neighborhood(R) and minimum points in the neighborhood(N).
- The point-parameter (N) determines the density of acceptable clusters and determines which objects will be labeled outliers or noise.
- The size-parameter R determines the size of the clusters found. If R is big enough, there would be one big cluster and no outliers. If R is small, there will be small dense clusters and there might be many outliers.
- The algorithm is as follows
  1. Select values of R and N.
  2. Arbitrarily select an object p.
  3. Retrieve all objects that are connected to p, given R and N.

Scanned with CamScanner

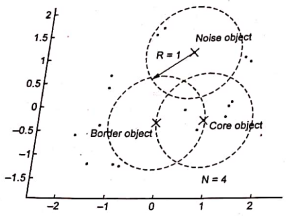


Fig. Q.35.1

4. If p is a core object, a cluster is formed.
5. If p is a border object, no objects are in its proximity. Choose another object. Go to step 3.
6. Continue the process until all of the objects have been processed.

Q.36 Explain applications of cluster analysis.

Ans. :

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer basis. And they can characterize their customer groups based on purchasing patterns.
- In field of biology it can be used to derive plant and animal taxonomies, categorize genes with similar functionality and gain insight into structures inherent in populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according house type, value, geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function cluster analysis serve as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

5.7 : Dealing with Large Databases, Quality and Validity of Cluster Analysis Methods

Q37 Explain K- means method for large databases.

Ans. :

- It first picks the number of clusters and their seed centriods and then attempts to classify each object to belong to one of the following three groups :

1. Those that is certain to belong to a cluster. These objects together are called discard-set. Some information about these objects is computed and saved. This includes number of objects (n), a vector sum of all attribute values of the n objects and a vector sum of squares of all attribute values of n objects.
2. Those that are sufficiently close to each other to be replaced by their summary. These objects together are called the compression-set. The objects are however sufficiently far away from each cluster's centroid that they cannot be put in the discard set.
3. The remaining objects that are too difficult to assign to either of the two groups above. These objects are called the retained-set.

- These are stored as individual objects. They cannot be replaced by a summary.

5.8 : Quality and Validity of Cluster Analysis

Q.38 Explain quality and validity of cluster analysis methods.

Ans. : Let number of clusters = k

Let clusters C_i, i = 1....., k

Let total number of objects = N

Let number of objects in cluster = M_i

so that $M_1 + M_2 + ... + M_k = N$

- The within-cluster variation between the objects in a cluster is defined as the average squared distance of each object from the centroid of the cluster.

- If m_i is the centroid of the cluster C_i then the mean of the cluster is given by $m_i = \{\sum x_j\}/M_i$ and the internal cluster variation is given by $I_i = \{\sum (x_j - m_i)^2\}/M_i$
- The average within-cluster variation is given by,

$$I = (l/k)(I_1 + I_2 + ... + I_k) = (\Sigma I_i)/k$$

- The between cluster distances E is the average sum of squares of pairwise distances between the centroids of the k clusters.
- We may write E as

$$E = (l/k^2) \sum_i \sum_j (\mu_i - \mu_j)^2$$

Important Points to Remember

- Classification is a technique used to predict group membership for data instances.
- Cluster is a group of objects that belong to the same class. In other words the similar object are grouped in one cluster and dissimilar are grouped in other cluster.
- Cluster analysis is process of grouping a set of data-objects into clusters.
- Supervised learning in which the network is trained by providing it with input and matching output patterns. These input-output pairs are usually provided by an external teacher.
- Unsupervised learning algorithms aim to learn rapidly and can be used in real-time.

END...

Scanned with CamScanner