K.S.R. COLLEGE OF ENGINEERING(Autonomous)

Vision of the Institution

- We envision to achieve status as an excellent educational institution in the global knowledge hub, making self-learners, experts, ethical and responsible engineers, technologists, scientists, managers, administrators and entrepreneurs who will significantly contribute to research and environment friendly sustainable growth of the nation and the world.

Mission of the Institution

- To inculcate in the students self-learning abilities that enable them to become competitive and considerate engineers, technologists, scientists, managers, administrators and entrepreneurs by diligently imparting the best of education, nurturing environmental and social needs.
- To foster and maintain a mutually beneficial partnership with global industries and Institutions through knowledge sharing, collaborative research and innovation.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Vision of the Department

- To create ever green professionals for software industry, academicians for knowledge cultivation and researchers for contemporary society modernization.

Mission of the Department

- To produce proficient design, code and system engineers for software development.
- To keep updated contemporary technology and fore coming challenges for welfare of the society.

Programme Educational Objectives (PEOs)

PEO1 : Figure out, formulate, analyze typical problems and develop effective solutions by imparting the idea and principles of science, mathematics, engineering fundamentals and computing.

PEO2 : Competent professionally and successful in their chosen career through life-long learning.

PEO3 : Excel individually or as member of a team in carrying out projects and exhibit social needs and follow professional ethics.

K.S.R. COLLEGE OF ENGINEERING(Autonomous)

Department of Computer Science and Engineering

Subject Name: Data Warehousing and DataMining
Subject Code: 16CS566 Year/Semester: III/V

Course Outcomes: On completion of this course, the student will be able to

CO1 To provide an exposure to fundamental concepts of data warehouse.
CO2 To learn the concept of extraction of knowledge and data preprocessing. .
CO3 To understand techniques, mechanism, algorithm for feature extraction and classification.
CO4 To learn relevant tools that supports decision making in data mining.
CO5 To acquire the knowledge of data mining tools.

Program Outcomes (POs) and Program Specific Outcomes (PSOs)

A. Program Outcomes (POs)

Engineering Graduates will be able to :

PO1 **Engineering knowledge:** Ability to exhibit the knowledge of mathematics, science, engineering fundamentals and programming skills to solve problems in computer science.

PO2 **Problem analysis:** Talent to identify, formulate, analyze and solve complex engineering problems with the knowledge of computer science. .

PO3 **Design/development of solutions:** Capability to design, implement, and evaluate a computer based system, process, component or program to meet desired needs.

PO4 **Conduct investigations of complex problems:** Potential to conduct investigation of complex problems by methods that include appropriate experiments, analysis and synthesis of information in order to reach valid conclusions.

PO5 **Modern tool Usage:** Ability to create, select, and apply appropriate techniques, resources and modern engineering tools to solve complex engineering problems.

PO6 **The engineer and society:** Skill to acquire the broad education necessary to understand the impact of engineering solutions on a global economic, environmental, social, political, ethical, health and safety.

PO7 **Environmental and sustainability:** Ability to understand the impact of the professional engineering solutions in societal and Environmental contexts and demonstrate the knowledge of, and need for sustainable development.

PO8 **Ethics:** Apply ethical principles and commit to professional ethics and responsibility and norms of the engineering practices.

PO9 **Individual and team work:** Ability to function individually as well as on multi-disciplinary teams.

PO10 **Communication:** Ability to communicate effectively in both verbal and written mode to excel in the career.

PO11 **Project management and finance:** Ability to integrate the knowledge of engineering and management principles to work as a member and leader in a team on diverse projects.

PO12 **Life-long learning:** Ability to recognize the need of technological change by independent and life-long learning.

B. Program Specific Outcomes (PSOs)

PSO1 Develop and Implement computer solutions that accomplish goals to the industry, government or research by exploring new technologies.

PSO2 Grow intellectually and professionally in the chosen field.

UNIT I

1. **Define data warehouse? (Understanding)**

A data warehouse is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making. (or)

A data warehouse is a subject-oriented, time-variant and nonvolatile collection of data in support of management's decision-making process.

2. **Define OLTP? (Understanding)**

If an on-line operational database systems is used for efficient retrieval, efficient storage and management of large amounts of data, then the system is said to be on-line transaction processing.

3. **Define OLAP? (Understanding)**

Data warehouse systems serves users (or) knowledge workers in the role of data analysis and decision-making. Such systems can organize and present data in various formats. These systems are known as on-line analytical processing systems.

4. **How a database design is represented in OLAP systems? (Understanding)**

Star schema
Snowflake schema
Fact constellation schema

5. **Write short notes on multidimensional data model? (Applying)**

Data warehouses and OLTP tools are based on a multidimensional data model. This model is used for the design of corporate data warehouses and department data marts. This model contains a Star schema, Snowflake schema and Fact constellation schemas. The core of the multidimensional model is the data cube.

6. **Define data cube? (Understanding)**

It consists of a large set of facts (or) measures and a number of dimensions.

7. **What are facts? (Understanding)**

Facts are numerical measures. Facts can also be considered as quantities by which we can analyze the relationship between dimensions.

8. **What are dimensions? (Understanding)**

Dimensions are the entities (or) perspectives with respect to an organization for keeping records and are hierarchical in nature.

9. **Define dimension table? (Understanding)**

A dimension table is used for describing the dimension.

(e.g.) A dimension table for item may contain the attributes item_ name, brand and type.

10. **Define fact table? (Understanding)**

Fact table contains the name of facts (or) measures as well as keys to each of the related dimensional tables.

11. **What are the functionalities of Sourcing, Acquisition, Cleanup and Transformation Tools? (Understanding)**

    a. Removing unwanted data from operational databases
    b. Converting to common data names and definitions

c. Calculating summaries and derived data
d. Establishing defaults for missing data
e. Accommodating source data definition changes

12. Define Metadata. (Understanding)

- Metadata is data about data that describes the data warehouse.
- It is used for building, maintaining, managing, and using the data warehouse.
- Metadata can be classified into the following:
  - Technical metadata
  - Business metadata

13. List out the information in the technical metadata documents. (Understanding)

a. Information about data sources
b. Transformation descriptions
c. Warehouse objects and data structure definition for data targets
d. The rules used to perform data cleanup and data enhancement
e. Data mapping operations
f. Access authorization, backup history, archive history, information delivery history and data access

14. List out the information in the business metadata(Understanding)

a. Subject areas
b. Internet home pages
c. Other information to support all data warehousing components
d. Data warehouse operational information

15. Give the types of access tools(Understanding)

a. Data query and reporting tools
b. Application development tools
c. Executive information system tools
d. OLAP tools
e. Data mining tools

16. Give the types of query and reporting tools(Understanding)

a. Reporting tools
   1. Production reporting tools
   2. Report writers
b. Managed query tools

17. Define data mart. (Understanding)

Data mart is a data store that is subsidiary to a data warehouse of integrated data. The data mart is directed at a partition of data that is created for use of a dedicated group of users.

What are the two approaches for data warehouse development? [Evaluating]

a. The **Top-down Approach**, meaning that 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. **(From Data warehouse to Data marts)**
b. The **Bottom-up Approach**, implying that the business priorities resulted in developing individual data marts, which are then integrated into enterprise data warehouse.(From Data marts to Data Warehouse)

18. Define Snowflake schema(Understanding)

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

19. **Define Fact Constellation(Understanding)**

Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

20. **What are the categories of measures[Evaluating]**
    - Distributive
    - Algebraic
    - holistic

21. **Define distributive measure(Understanding)**

Distributive: if the result derived by applying the function to *n* aggregate values is the same as that derived by applying the function on all the data without partitioning.
E.g., count(), sum(), min(), max().

22. **Define algebraic measure(Understanding)**

algebraic: if it can be computed by an algebraic function with *M* arguments (where *M* is a bounded integer), each of which is obtained by applying a distributive aggregate function.
E.g., avg(), min_N(), standard_deviation().

23. **Define holistic measure(Understanding)**

holistic: if there is no constant bound on the storage size needed to describe a subaggregate.
E.g., median(), mode(), rank().

24. **What are the factors that affect the data warehouse design? [Evaluating]**

a. Heterogeneity of Data sources, which affects data conversion, quality, timeliness
b. Use of historical data, while implies that data may be “old”.
c. Tendency of databases to grow very large

25. **Define Virtual warehouse (Understanding)**

Virtual warehouse

A set of views over operational databases
Only some of the possible summary views may be materialized

26. **What is star schema? (Understanding)**

The star schema architecture is the simplest data warehouse schema. It is called a star schema because the diagram resembles a star, with points radiating from a center. The center of the star consists of fact table and the points of the star are the dimension tables. Usually the fact tables in a star schema are in third normal form(3NF) whereas dimensional tables are de-normalized. Despite the fact that the star schema is the simplest architecture, it is most commonly used nowadays and is recommended by Oracle.

27. **Define fact table. (Understanding)**

A fact table is a table that contains summarized numerical and historical data (facts) and a multipart index composed of foreign keys from the primary keys of related dimension tables.

A fact table typically has two types of columns: foreign keys to dimension tables and measures those that contain numeric facts. A fact table can contain fact's data on detail or aggregated level.

28. **Define dimension table. (Understanding)**

Dimensions are categories by which summarized data can be viewed. E.g. a profit summary in a fact table can be viewed by a Time dimension (profit by month, quarter, year), Region dimension (profit by country, state, city), Product dimension (profit for product1, product2).

29. Give the characteristics of star schema. (Understanding)

The main characteristics of star schema:

- Simple structure -> easy to understand schema
- Great query effectives -> small number of tables to join
- Relatively long time of loading data into dimension tables -> de-normalization, redundancy data caused that size of the table could be large.
- The most commonly used in the data warehouse implementations -> widely supported by a large number of business intelligence tools

30. List out the components of star schema? [Evaluating]

A large central table (fact table) containing the bulk of data with no redundancy.

A set of smaller attendant tables (dimension tables), one for each dimension.

31. What are dependent and independent data marts? [Evaluating]

Dependent data marts are sourced directly from enterprise data warehouses.

Independent data marts are data captured from one (or) more operational systems (or) external information providers (or) data generated locally with in particular department (or) geographic area.

16 MARKS

1. **Discuss the components of data warehouse. (Applying)**

2. List out the differences between OLTP and OLAP. (Evaluating)

3. Discuss the various schematic representations in multidimensional model. (Applying)

4. Expalin the three-tier data warehouse architecture. (Understanding)

5. Write notes on metadata repository. (Understanding)

UNIT II

1. Define Data Mart(Understanding)

- From a data warehouse, data flows to various departments for their customized DSS usage. These individual department components are called data marts.
- A data mart is a body of DSS data for a department that has an architectural Foundation of a data warehouse.

Data mart is a subset of a data warehouse and is much popular than data warehouse.

2. List out the type of Data Marts. (Evaluating)

- Multidimensional (MDDB OLAP or MOLAP) data mart.
- Relational OLAP (ROLAP) data mart.

3. List out the factors that should be considered for loading a Data Mart. (Evaluating)

- Frequency and schedule
- Total or partial refreshment
- Re-sequencing and merging of data
- Aggregation of data, summarization, efficiency
- Integrity of data
- Data relationship and integrity of data domains
- Producing meta data for describing the loading process

4. Define metadata(Understanding)

Metadata (data about data) describes the details about the data in a data warehouse or in a data mart.

5. **Write down the components of metadata for a given data warehouse or data mart? (Understanding)**
   - Description of sources of the data
   - Description of customization that may have taken place as the data passes from data warehouse into data mart
   - Description information about data mart, its tables, attributes and relationships, etc.
   - Definitions of all types

6. **Write about maintenance of data mart. (Understanding)**
   - Periodic maintenance of a data mart
     - Loading, refreshing and purging the data in it.
   - Refreshing the data is performed in regular cycles as per the nature of the frequency of data update.
   - Purging – the data mart is read periodically and some (old) data is selected for purging or removing.

7. **Give the nature of data in a data mart. (Understanding)**
   - Detailed level
   - Summary level
   - Ad hoc data
   - Preprocessed or prepared data

8. **List out the software components for a data mart. (Evaluating)**

   The software that can be found with a data mart includes:
   - DBMS
   - Access and analysis software for automatic creation of data mart
   - Purging and archival software
   - Metadata management software

9. **What are the types of tables in the data mart? (Understanding)**
   - Summary tables
   - Detailed tables
   - Reference tables
   - Historical tables
   - Analytical tables

10. **Define multidimensional data model. (Understanding)**

The multidimensional data model is an integral part of On-Line Analytical Processing, or OLAP. Because OLAP is on-line, it must provide answers quickly; analysts pose iterative queries during interactive sessions, not in batch jobs that run overnight. And because OLAP is also analytic, the queries are complex. The multidimensional data model is designed to solve complex queries in real time.

11. **What are the operations of multidimensional data model? (Understanding)**

Operations in Multidimensional Data Model:

- Aggregation (roll-up)
  –dimension reduction: e.g., total sales by city

  –summarization over aggregate hierarchy: e.g., total sales by city and year -> total sales by region and by year

- Selection (slice) defines a subcube
  e.g., sales where city =Palo Altoand date =1/15/96

- Navigation to detailed data (drill-down)

–e.g., (sales -expense) by city, top 3% of cities by average income

- Visualization Operations (e.g., Pivot or dice)

12. Define catalog. (Understanding)

Catalog does not contain any data. It just contains information about connecting to the database and the fields that will be accessible for reports.

13. Define OLAP. (Understanding)

OLAP stands for Online Analytical Processing. It uses database tables (fact and dimension tables) to enable multidimensional viewing, analysis and querying of large amounts of data. E.g.

OLAP technology could provide management with fast answers to complex queries on their operational data or enable them to analyze their company's historical data for trends and patterns. Online Analytical Processing (OLAP) applications and tools are those that are designed to ask ―complex a query of large multidimensional collections of data.‖ Due to that OLAP is accompanied with data warehousing.

14. Why we need OLAP? (Analyzing)

Need the key driver of OLAP is the multidimensional nature of the business problem. These problems are characterized by retrieving a very large number of records that can reach gigabytes and terabytes and summarizing this data into a form information that can by used by business analysts.

15. Give the OLAP Guidelines. (Understanding)

1. Multidimensional conceptual view
2. Transparency
3. Accessibility
4. Consistent reporting performance
5. Client/server architecture
6. Generic dimensionality
7. Dynamic sparse matrix handling
8. Multiuser support
9. Unrestricted cross-dimensional operations
10. Interactive data manipulation
11. Flexible reporting
12. Unlimited dimensions and aggregation levels

16. What are the supports provided by OLAP systems? (Understanding)

- Comprehensive database management tools: This gives the database management to control distributed Businesses.
- The ability to drill down to detail source record level: Which requires that The OLAP tool should allow smooth transitions in the multidimensional database.
- Incremental database refresh: The OLAP tool should provide partial refresh.
- Structured Query Language (SQL interface): the OLAP system should be able to integrate effectively in the surrounding enterprise environment.

17. Differentiate OLTP and OLAP. (Evaluating)

OLTP vs OLAP

OLTP stands for On Line Transaction Processing and is a data modeling approach typically used to facilitate and manage usual business applications. Most of applications you see and use are OLTP based. OLTP technology used to perform updates on operational or transactional systems

(e.g., point of sale systems)

OLAP stands for On Line Analytic Processing and is an approach to answer multi-dimensional queries. OLAP was conceived for Management Information Systems and Decision Support Systems. OLAP technology used to perform complex analysis of the data in a data warehouse.

18. Define ROLAP. (Understanding)

This methodology relies on manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality. In essence, each action of slicing and dicing is equivalent to adding a "WHERE" clause in the SQL statement. Data stored in relational tables.

19. Give the advantages of ROLAP. (Understanding)

Advantages: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.

Can leverage functionalities inherent in the relational database: Often, relational database already comes with a host of functionalities. ROLAP technologies, since they sit on top of the relation al database, can therefore leverage these functionalities.

20. Give the disadvantages of ROLAP. (Understanding)

Disadvantages:

Performance can be slow: Because each ROLAP report is essentially a SQL query (or multiple SQL queries) 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 (for example, it is difficult to perform complex calculations using SQL), ROLAP technologies are therefore traditionally limited by what SQL can do. ROLAP vendors have mitigated this risk by building into the tool out-of-the-box complex functions as well as the ability to allow users to define their own functions.

21. Define MOLAP. (Understanding)

This is the more traditional way of OLAP analysis. In MOLAP, data is stored in a multidimensional cube. The storage is not in the relational database, but in proprietary formats. That is, data stored in array-based structures.

22. Give the advantages of MOAP. (Understanding)

Advantages:

Excellent performance: MOLAP cubes are built for fast data retrieval, and are optimal for slicing and dicing operations.

Can perform complex calculations: All calculations have been pre-generated when the cube is created. Hence, complex calculations are not only doable, but they return quickly.

23. Give disadvantages of MOAP. (Understanding)

Disadvantages:

Limited in the amount of data it can handle: Because all calculations are performed when the cube is built, it is not possible to include a large amount of data in the cube itself.

This is not to say that the data in the cube cannot be derived from a large amount of data. Indeed, this is possible. But in this case, only summary - level information will be included in the cube itself.

Requires additional investment: Cube technology are often proprietary and do not already exist in the organization. Therefore, to adopt MOLAP technology, chances are additional investments in human and capital resources are needed.

24. Define HOLAP? (Understanding)

The hybrid OLAP approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and the faster computation of MOLAP,(i.e.) a HOLAP server may allow large volumes of detail data to be stored in a relational database, while aggregations are kept in a separate MOLAP store.

25. List out the OLAP operations in multidimensional data model? (Evaluating)

- Roll-up
- Drill-down
- Slice and dice
- Pivot (or) rotate

26. What is roll-up operation? (Understanding)

The roll-up operation is also called drill-up operation which performs aggregation on a data cube either by climbing up a concept hierarchy for a dimension (or) by dimension reduction.

27. What is drill-down operation? (Understanding)

Drill-down is the reverse of roll-up operation. It navigates from less detailed data to more detailed data. Drill-down operation can be taken place by stepping down a concept hierarchy for a dimension.

28. What is slice operation? (Understanding)

The slice operation performs a selection on one dimension of the cube resulting in a sub cube.

29. What is dice operation? (Understanding)

The dice operation defines a sub cube by performing a selection on two (or) more dimensions.

30. What is pivot operation? (Understanding)

This is a visualization operation that rotates the data axes in an alternative presentation of the data.

31. What is enterprise warehouse? (Understanding)

An enterprise warehouse collects all the information’s about subjects spanning the entire organization. It provides corporate-wide data integration, usually from one (or) more operational systems (or) external information providers. It contains detailed data as well as summarized data and can range in size from a few giga bytes to hundreds of giga bytes, tera bytes (or) beyond.

16 MARKS

1. **Explain in detail about data mart.(Creating)**
2. **Explain OLAP guidelines in detail.(Creating)**
3. **Differentiate OLTP and OLAP. (Evaluating)**
4. **Explain MOLAP, ROLAP and HOLAP in detail. (Creating)**
5. **Explain about multidimensional data model in detail. (Creating)**

UNIT III

1. Define Data mining. (Understanding)

It refers to extracting or “mining” knowledge from large amount of data. Data mining is a process of discovering interesting knowledge from large amounts of data stored either, in database, data warehouse, or other information repositories

2. Give some alternative terms for data mining. (Understanding)

- Knowledge mining
- Knowledge extraction
- Data/pattern analysis.
- Data Archaeology
- Data dredging

3. What is KDD. (Understanding)

KDD-Knowledge Discovery in Databases.

4. What are the steps involved in KDD process. (Understanding)

- Data cleaning
- Data Mining
- Pattern Evaluation
- Knowledge Presentation
- Data Integration
- Data Selection
- Data Transformation

5. What is the use of the knowledge base? (Understanding)

Knowledge base is domain knowledge that is used to guide search or evaluate the interestingness of resulting pattern. Such knowledge can include concept hierarchies used to organize attribute /attribute values in to different levels of abstraction of Data Mining.

6.Arcitecture of a typical data mining system. (Understanding)

Knowledge base

7.Mention some of the data mining techniques. (Understanding)

- Statistics
- Machine learning
- Decision Tree
- Hidden markov models
- Artificial Intelligence
- Genetic Algorithm
- Meta learning

8.Give few statistical techniques. (Understanding)

- Point Estimation
- Data Summarization
- Bayesian Techniques
- Testing Hypothesis
- Correlation
- Regression

9.What is meta learning. (Understanding)

Concept of combining the predictions made from multiple models of data mining and analyzing those predictions to formulate a new and previously unknown prediction.

GUI
Pattern Evaluation
Database or Data warehouse
Server

10.Define Genetic algorithm. (Understanding)

• Search algorithm.

• Enables us to locate optimal binary string by processing an initial random population of binary strings by performing operations such as artificial mutation , crossover and selection.

11.What is the purpose of Data mining Technique? (Understanding)

It provides a way to use various data mining tasks.

12.Define Predictive model. (Understanding)

It is used to predict the values of data by making use of known results from a different set of sample data.

13.Data mining tasks that are belongs to predictive model (Evaluating)

- Classification
- Regression
- Time series analysis

14.Define descriptive model(Understanding)

• It is used to determine the patterns and relationships in a sample data. Data mining tasks that belongs to descriptive model:

- Clustering
- Summarization
- Association rules
- Sequence discovery

15. Define the term summarization (Understanding)

The summarization of a large chunk of data contained in a web page or a document.

Summarization = characterization=generalization

16. List out the advanced database systems. (Understanding)

- Extended-relational databases
- Object-oriented databases
- Deductive databases
- Spatial databases
- Temporal databases
- Multimedia databases
- Active databases
- Scientific databases
- Knowledge databases

17. Define cluster analysis (Understanding)

Cluster analyses data objects without consulting a known class label. The class labels are not present in the training data simply because they are not known to begin with.

18.Classifications of Data mining systems.(Evaluating)

- Based on the kinds of databases mined:
- Based on kinds of Knowledge mined
- Based on kinds of techniques utilized
- Based on applications adopted

19.Describe challenges to data mining regarding data mining methodology and user interaction issues. (Applying)

- Mining different kinds of knowledge in databases
- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages and ad hoc data mining
- Presentation and visualization of data mining results
- Handling noisy or incomplete data
- Pattern evaluation

20.Describe challenges to data mining regarding performance issues. (Applying)

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, and incremental mining algorithms

21.Describe issues relating to the diversity of database types. (Applying)

- Handling of relational and complex types of data
- Mining information from heterogeneous databases and global information Systems

22.What is meant by pattern? (Understanding)

Pattern represents knowledge if it is easily understood by humans; valid on test data with some degree of certainty; and potentially useful, novel,or validates a hunch about which the used was curious. Measures of pattern interestingness, either objective or subjective, can be used to guide the discovery process.

23.How is a data warehouse different from a database?

Data warehouse is a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision-making. Database consists of a collection of interrelated data.

24.What are the major task of data preprocessing? (Understanding)

a. Data cleaning

Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

b. Data integration

Integration of multiple databases, data cubes, or files

c. Data transformation

Normalization and aggregation

d. Data reduction

Obtains reduced representation in volume but produces the same or similar analytical results

e. Data discretization

Part of data reduction but with particular importance, especially for numerical data

25.Define noise(Understanding)

Noise: random error or variance in a measured variable

16 **MARKS**

1. Explain the evolution of Database technology? (Understanding)
2.Explain the steps of knowledge discovery in databases? (Understanding)
3. Explain the architecture of data mining system? (Understanding)
4.Explain various tasks in data mining?
(or) **(Understanding)**
Explain the taxonomy of data mining tasks? (Understanding)
5.Explain various techniques in data mining? (Understanding)
6.Explain about data preprocessing techniques in detail. (Understanding)

UNIT IV

1. Define Association Rule Mining. (Understanding)

Association rule mining searches for interesting relationships among items in a given data set.

2. When we can say the association rules are interesting? (Applying)

Association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Users or domain experts can set such thresholds.

3. Explain Association rule in mathematical notations. (Understanding)

Let I-{i1,i2,…..,im} be a set of items
Let D, the task relevant data be a set of database transaction T is a set of items
An association rule is an implication of the form A=>B where A C I, B C I, and An B=f. The rule A=>B contains in the transaction set D with support s, where s is the percentage of transactions in D that contain AUB. The Rule A=> B has confidence c in the transaction set D if c is the percentage of transactions in D containing A that also contain B.

4. Define support and confidence in Association rule mining. (Understanding)

Support S is the percentage of transactions in D that contain AUB. Confidence c is the percentage of transactions in D containing A that also contain B.
Support (A=>B)= P(AUB)
Confidence (A=>B)=P(B/A)

5. How are association rules mined from large databases? (Creating)

- I step: Find all frequent item sets:
- II step: Generate strong association rules from frequent item sets

6. Describe the different classifications of Association rule mining. (Applying)

- Based on types of values handled in the Rule
  - i. Boolean association rule
  - ii. Quantitative association rule
- Based on the dimensions of data involved
  - i. Single dimensional association rule
  - ii. Multidimensional association rule
- Based on the levels of abstraction involved
  - i. Multilevel association rule
  - ii. Single level association rule
- Based on various extensions
  - i. Correlation analysis
  - ii. Mining max patterns

7. What is the purpose of Apriori Algorithm? (Understanding)

Apriori algorithm is an influential algorithm for mining frequent item sets for

Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties.

8. Define anti-monotone property. (Understanding)

If a set cannot pass a test, all of its supersets will fail the same test as well.

9. How to generate association rules from frequent item sets? (Applying)

Association rules can be generated as follows

For each frequent item set1, generate all non empty subsets of 1.

For every non empty subsets s of 1, output the rule "S=>(1-s)"if

Support count(1)

=min_conf,

Support_count(s)

Where min_conf is the minimum confidence threshold.

10. Give few techniques to improve the efficiency of Apriori algorithm. (Applying)

• Hash based technique
• Transaction Reduction
• Portioning
• Sampling
• Dynamic item counting

11. What are the things suffering the performance of Apriori candidate generation technique. (Understanding)

• Need to generate a huge number of candidate sets

• Need to repeatedly scan the scan the database and check a large set of candidates by pattern matching

12. Describe the method of generating frequent item sets without candidate generation. (Applying)

Frequent-pattern growth(or FP Growth) adopts divide-and-conquer strategy.

Steps:

Compress the database representing frequent items into a frequent pattern tree or FP tree Divide the compressed database into a set of conditional database Mine each conditional database separately

13. Define Iceberg query. (Understanding)

It computes an aggregate function over an attribute or set of attributes in order to find aggregate values above some specified threshold. Given relation R with attributes a1,a2,…..,an and b, and an aggregate function,agg_f, an iceberg query is the form Select R.a1,R.a2,…..R.an,agg_f(R,b) From relation R Group by R.a1,R.a2,….,R.an Having agg_f(R.b)>=threshold

14. Mention few approaches to mining Multilevel Association Rules (Understanding)

• Uniform minimum support for all levels(or uniform support)
• Using reduced minimum support at lower levels(or reduced support)
• Level-by-level independent
• Level-cross filtering by single item
• Level-cross filtering by k-item set

15. What are multidimensional association rules? (Understanding)

Association rules that involve two or more dimensions or predicates

• Interdimension association rule: Multidimensional association rule with no repeated predicate or dimension

• Hybrid-dimension association rule: Multidimensional association rule with multiple occurrences of some predicates or dimensions.

16. Define constraint-Based Association Mining. (Understanding)

Mining is performed under the guidance of various kinds of constraints provided by the user.

The constraints include the following

- Knowledge type constraints
- Data constraints
- Dimension/level constraints
- Interestingness constraints
- Rule constraints.

17. Define the concept of classification. (Understanding) (Understanding)

Two step process

- A model is built describing a predefined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes.
- The model is used for classification.

18. What is Decision tree? (Understanding)

A decision tree is a flow chart like tree structures, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The top most in a tree is the root node.

19. What is Attribute Selection Measure? (Understanding)

The information Gain measure is used to select the test attribute at each node in the decision tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split.

20. Describe Tree pruning methods. (Applying)

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outlier. Tree pruning methods address this problem of over fitting the data.

Approaches:

- Pre pruning
- Post pruning

21. Define Pre Pruning (Understanding)

A tree is pruned by halting its construction early. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset samples.

22. Define Post Pruning. (Understanding)

Post pruning removes branches from a “Fully grown” tree. A tree node is pruned by removing its branches.

Eg: Cost Complexity Algorithm

24. Define the concept of prediction.

Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled sample or to assess the value or value ranges of an attribute that a given sample is likely to have.

16 MARKS

1. Explain the issues regarding classification and prediction? (Understanding)
2. Explain classification by Decision tree induction?(Applying)
3. Write short notes on patterns? (creating)
4. Explain mining single –dimensional Boolean associated rules from transactional databases? (Understanding)
5. Explain apriori algorithm?
6. Explain how the efficiency of apriori is improved? (Understanding)
7. Explain frequent item set without candidate without candidate generation?
8. Explain mining Multi-dimensional Boolean association rules from transaction databases? (Understanding)
9. Explain constraint-based association mining? (Evaluating)
10. Explain regression in predictive modeling? (Understanding)
11. Explain statistical perspective in data mining? (Understanding)
12. Explain Bayesian classification. (Understanding)

UNIT V

1.Define Clustering? (Understanding)

Clustering is a process of grouping the physical or conceptual data object into clusters.

2. What do you mean by Cluster Analysis?(Applying)

A cluster analysis is the process of analyzing the various clusters to organize the different objects into meaningful and descriptive objects.

3. What are the fields in which clustering techniques are used? (Understanding)

- Clustering is used in biology to develop new plants and animal taxonomies.
- Clustering is used in business to enable marketers to develop new distinct groups of their customers and characterize the customer group on basis of purchasing.
- Clustering is used in the identification of groups of automobiles Insurance policy customer.
- Clustering is used in the identification of groups of house in a city on the basis of house type, their cost and geographical location.
- Clustering is used to classify the document on the web for information discovery.

4.What are the requirements of cluster analysis? (Understanding)

The basic requirements of cluster analysis are

- Dealing with different types of attributes.
- Dealing with noisy data.
- Constraints on clustering.
- Dealing with arbitrary shapes.
- High dimensionality
- Ordering of input data
- Interpretability and usability
- Determining input parameter and
- Scalability

5.What are the different types of data used for cluster analysis? (Understanding)

The different types of data used for cluster analysis are interval scaled, binary, nominal, ordinal and ratio scaled data.

6. What are interval scaled variables? (Understanding)

Interval scaled variables are continuous measurements of linear scale.

For example, height and weight, weather temperature or coordinates for any cluster. These measurements can be calculated using Euclidean distance or Minkowski distance.

7. Define Binary variables? And what are the two types of binary variables? (Understanding)

Binary variables are understood by two states 0 and 1, when state is 0, variable is absent and when state is 1, variable is present. There are two types of binary variables, symmetric and asymmetric binary variables. Symmetric variables are those variables that have same state values and weights. Asymmetric variables are those variables that have not same state values and weights.

8. Define nominal, ordinal and ratio scaled variables? (Understanding)

A nominal variable is a generalization of the binary variable. Nominal variable has more than two states, For example, a nominal variable, color consists of four states, red, green, yellow, or black. In Nominal variables the total number of states is N and it is denoted by letters, symbols or integers.

An ordinal variable also has more than two states but all these states are ordered in a meaningful sequence.

A ratio scaled variable makes positive measurements on a non-linear scale, such as exponential scale, using the formula AeBt or Ae-Bt Where A and B are constants.

9. What do u mean by partitioning method? (Creating)

In partitioning method a partitioning algorithm arranges all the objects into various partitions, where the total number of partitions is less than the total number of objects. Here each partition represents a cluster. The two types of partitioning method are k-means and k-medoids.

10. Define CLARA and CLARANS? (Understanding)

Clustering in LARge Applications is called as CLARA. The efficiency of CLARA depends upon the size of the representative data set. CLARA does not work properly if any representative data set from the selected representative data sets does not find best k-medoids.

To recover this drawback a new algorithm, Clustering Large Applications based upon RANdomized search (CLARANS) is introduced. The CLARANS works like CLARA, the only difference between CLARA and CLARANS is the clustering process that is done after selecting the representative data sets.

11. What is Hierarchical method? (Understanding)

Hierarchical method groups all the objects into a tree of clusters that are arranged in a hierarchical order. This method works on bottom-up or top-down approaches.

12. Differentiate Agglomerative and Divisive Hierarchical Clustering? (Applying)

Agglomerative Hierarchical clustering method works on the bottom-up approach. In Agglomerative hierarchical method, each object creates its own clusters. The single Clusters are merged to make larger clusters and the process of merging continues until all the singular clusters are merged into one big cluster that consists of all the objects. Divisive Hierarchical clustering method works on the top-down approach. In this method all the objects are arranged within a big singular cluster and the large cluster is continuously divided into smaller clusters until each cluster has a single object.

13. What is CURE? (Understanding)

Clustering Using Representatives is called as CURE. The clustering algorithms generally work on spherical and similar size clusters. CURE overcomes the problem of spherical and similar size cluster and is more robust with respect to outliers.

14. Define Chameleon method? (Understanding)

Chameleon is another hierarchical clustering method that uses dynamic modeling. Chameleon is introduced to recover the drawbacks of CURE method. In this method two clusters are merged, if the interconnectivity between two clusters is greater than the interconnectivity between the objects within a cluster.

15. Define Density based method? (Understanding)

Density based method deals with arbitrary shaped clusters. In density-based method, clusters are formed on the basis of the region where the density of the objects is high.

16. What is a DBSCAN? (Understanding)

Density Based Spatial Clustering of Application Noise is called as DBSCAN. DBSCAN is a density based clustering method that converts the high-density objects regions into clusters with arbitrary shapes and sizes. DBSCAN defines the cluster as a maximal set of density connected points.

17. What do you mean by Grid Based Method? (Understanding)

In this method objects are represented by the multi resolution grid data structure. All the objects are quantized into a finite number of cells and the collection of cells build the grid structure of objects. The clustering operations are performed on that grid structure. This method is widely used because its processing time is very fast and that is independent of number of objects.

18. What is a STING? (Understanding)

Statistical Information Grid is called as STING; it is a grid based multi resolution clustering method. In STING method, all the objects are contained into rectangular cells, these cells are kept into various levels of resolutions and these levels are arranged in a hierarchical structure.

19. Define Wave Cluster? (Understanding)

It is a grid based multi resolution clustering method. In this method all the objects are represented by a multidimensional grid structure and a wavelet transformation is applied for finding the dense region. Each grid cell contains the information of the group of objects that map into a cell. A wavelet transformation is a process of signaling that produces the signal of various frequency sub bands.

20. What is Model based method? (Understanding)

For optimizing a fit between a given data set and a mathematical model based methods are used. This method uses an assumption that the data are distributed by probability distributions. There are two basic approaches in this method that are 1. Statistical Approach 2. Neural Network Approach.

21. What is the use of Regression? (Understanding)

Regression can be used to solve the classification problems but it can also be used for applications such as forecasting. Regression can be performed using many different types of techniques; in actually regression takes a set of data and fits the data to a formula.

22. What are the reasons for not using the linear regression model to estimate the output data?(Applying)

There are many reasons for that, One is that the data do not fit a linear model, It is possible however that the data generally do actually represent a linear model, but the linear model generated is poor because noise or outliers exist in the data. Noise is erroneous data and outliers are data values that are exceptions to the usual and expected data.

23. What are the two approaches used by regression to perform classification?(Evaluating)

Regression can be used to perform classification using the following approaches

1. Division: The data are divided into regions based on class.

2. Prediction: Formulas are generated to predict the output class value.

24. What do u mean by logistic regression? (Understanding)

Instead of fitting a data into a straight line logistic regression uses a logistic curve.
The formula for the univariate logistic curve is
P= e (C0+C1X1)
1+e (C0+C1X1)
The logistic curve gives a value between 0 and 1 so it can be interpreted as the probability of class membership.

25. What is Time Series Analysis? (Understanding)

A time series is a set of attribute values over a period of time. Time Series Analysis may be viewed as finding patterns in the data and predicting future values.

26. What are the various detected patterns? (Creating)

Detected patterns may include:
¨ *Trends :* It may be viewed as systematic non-repetitive changes to the values over time.
¨ *Cycles :* The observed behavior is cyclic.
¨ *Seasonal :* The detected patterns may be based on time of year or month or day.
¨ *Outliers :* To assist in pattern detection , techniques may be needed to remove or reduce the impact of outliers.

27. What is Smoothing? (Understanding)

Smoothing is an approach that is used to remove the nonsystematic behaviors found in time series. It usually takes the form of finding moving averages of attribute values. It is used to filter out noise and outliers.

28. What is Auto regression? (Understanding)

Auto regression is a method of predicting a future time series value by looking at previous values. Given a time series X = (x1,x2,….xn) a future value, x n+1, can be found using
x n+1 = x + j nx n + j n-1x n-1 +……+ e n+1 Here e n+1 represents a random error, at time n+1.In addition, each element in the time series can be viewed as a combination of a random error and a linear combination of previous values.

16 **MARKS**

1. Discuss the requirements of clustering in data mining. (Analyzing)
2. Explain the partitioning method of clustering. (Understanding)
3. Explain PAM algorithm in detail. (Understanding)
4. Describe the working of DBSCAN algorithm and explain the concept of a clusters used in DBSCAN.(Applying)
5. Explain hierarchical clustering method in detail. (Understanding)
6. Explain outlier analysis in detail. (Understanding)
7. Explain data mining applications in detail. (Understanding)