KARPAGAM ACADEMY OF HIGHER EDUCATION

KARPAGAM UNIVERSITY

(Deemed University Under Section 3 of UGC Act 1956)

Coimbatore -21.

DEPARTEMENT OF COMPUTER SCIENCE

SYLLABUS

Semester-I

17CSP103 DATA MINING AND WAREHOUSING 4H – 4C

Instruction Hours / week: L: 4 T: 0 P: 0 **Marks:** Int : **40** Ext : **60** Total: **100**

COURSE OBJECTIVE:

This course introduce students to the basic concepts and techniques of Data Mining develop skills of using recent data mining software for solving practical problems gain experience of doing independent study and research.

COURSE OUTCOME:

- Developing skills of using recent data mining software for solving practical problems.
- Gaining experience of doing independent study and research.
- Possessing some knowledge of the concepts and terminology associated with database systems statistics and machine learning

UNIT-I

Introduction: Fundamentals of data mining - Data Mining Functionalities - Classification of Data Mining systems - Major issues in Data Mining.

Data Warehouse and OLAP Technology: An Overview - Data Warehouse - Multidimensional Data Model - Data Warehouse Architecture - Data Warehouse Implementation - From Data Warehousing to Data Mining.

UNIT-II

Data Preprocessing: Needs Preprocessing the Data - Data Cleaning - Data Integration and Transformation - Data Reduction - Discretization and Concept Hierarchy Generation - Online Data Storage.

UNIT-III

Mining Frequent Patterns Associations and Correlations: Basic Concepts - Efficient and Scalable Frequent item set Mining Methods - Mining various kinds of Association rules – From Association Mining to Correlation Analysis - Constraint-Based Association Mining.

UNIT-IV

Classification and Prediction: Issues Regarding Classification and Prediction -Classification by Decision Tree Induction - Rule-based Classification – Prediction - Accuracy and Error Measures - Evaluating the Accuracy of a classifier or Predictor -Ensemble Methods - incrvveases the Accuracy - Model Selection.

UNIT-V

Cluster Analysis Introduction : Types of Data in Cluster Analysis - A Categorization of Major Clustering Methods - Partitioning Methods - Hierarchical Methods – Density-Based Methods Grid-Based Methods - Model-Based Clustering Methods - Clustering High-Dimensional Data – Constraint-Based Cluster Analysis - Outlier Analysis.

Applications and Trends in Data mining: Text Mining - Web Mining - Multimedia Mining- Spatial Mining - Visual data mining.

SUGGESTED READINGS

TEXT BOOK

1. Jiawei Han & Micheline Kamber. (2013). Data Mining – Concepts and Techniques(#3rd ed.). Mumbai: Morgan Kaufmann Publishers.
   (Page Nos: 1-36 47 -94 105-148 227 -267 289 -306 318- 322 354-372 386-458 600-640)

REFERENCES

1. Michael, J.A. , Berry Gordon, S. & Linoff. (2011). Data mining Techniques (3rd ed.). Wiley Publishing Inc.
2. Arun, K. Pujari. (2010). Data Mining Techniques (2nd ed.). New Delhi:University Press
3. Gupta, G..K. (2000). Introduction to Data mining with case studies (1st ed.). New Delhi: Prentice Hall of India.
4. Hillol Kargupta ,Anupam Joshi, Krishnamoorthy Sivakumar & Yelena Yesha. (2005). Data Mining Next Generation Challenges and Future Directions(1st ed.). New Delhi: Prentice Hall of India .
5. Inmon, W. H. (2005). Building the DataWarehouse (4th ed.). New Delhi: Wiley Dreamtech India.
6. Michael, J.A., Berry Gordon, S., & Linoff. (2008). Mastering Data Mining (3rd ed.). New Delhi: John Wiley & Sons Inc.
7. Margaret, H. Dunham. (2008). Data Mining Introductory and advanced topics (4th ed.). New Delhi:Pearson Education.
8. Paulraj Ponnaiah. (2011). Data Warehousing Fundamentals (2nd ed.). New Delhi: Wiley Student ed.
9. Ralph Kimball. (2011).The Data Warehouse Life cycle Tool kit (2nd ed.). New Delhi: Wiley Student ed.
10. Sam Anahory, & Dennis Murray. (2009). Data Warehousing in the Real World (3rd ed.). Pearson Education Asia.
11. Soman, K.P., Shyam Diwakar, & Ajay,V. (2006). Insight into Data Mining Theory and Practice (1st ed.). New Delhi: Prentice Hall of India.

WEB SITES

1. Thedacs.Com
2. Dwreview.Com
3. Pcai.Com
4. Eruditionhome.Com

KARPAGAM ACADEMY OF HIGHER EDUCATION
DEPARTMENT OF COMPUTER SCIENCE
LESSON PLAN

Facultly Name: K.Banuroopa **Batch: 2017-2019**
Subject : DATA MINING AND WAREHOUSING **Sub.Code : 17CSP103**
CLASS : I - M.Sc (CS) **Semester: I**

UNIT-I			
S.No	**Lecture Duration (Period)**	**Topics to be Covered**	**Support Materials**
1	1	**Introduction:** Fundamentals of data mining	T1 : 1-20
2	1	Data Mining Functionalities	T1 : 21-27
3	1	Classification of Data Mining systems	T1 : 29-31
4	1	Major issues in Data Mining.	T1 : 36-38
5	1	**Data Warehouse and OLAP Technology**- An Overview	W1
6	1	Data Warehouse	T1 : 105-109
7	1	Multidimensional Data Model	T1 : 110-126
8	1	Data Warehouse Architecture	T1 : 127-135 , R1 : 484-494
9	1	Data Warehouse Implementation	T1 : 137-145
10	1	From Data Warehousing to Data Mining	T1 : 146-149
11	1	Recapitulation and Discussion of Important Questions	
Total No. of Hours Planned for Unit I			11 Hours

TEXT BOOK:

1. Jiawei Han & Micheline Kamber. (2013). Data Mining – Concepts and Techniques(#3rd ed.). Mumbai: Morgan Kaufmann Publishers.

(Page Nos: 1-36 47 -94 105-148 227 -267 289 -306 318- 322 354-372 386-458 600-640)

REFERENCES

1. Michael, J.A. , Berry Gordon, S. & Linoff. (2011). Data mining Techniques (3rd ed.). Wiley Publishing Inc.

2. Inmon, W. H. (2005). Building the DataWarehouse (4^{th} ed.). New Delhi: Wiley Dreamtech India.
3. Paulraj Ponnaiah. (2011). Data Warehousing Fundamentals (2nd ed.). New Delhi: Wiley Student ed.
4. Ralph Kimball. (2011).The Data Warehouse Life cycle Tool kit (2nd ed.). New Delhi: Wiley Student ed.
5. Sam Anahory, & Dennis Murray. (2009). Data Warehousing in the Real World (3rd ed.). Pearson Education Asia.

UNIT-II			
1	1	**Data Preprocessing**	T1 : 47-48
2	1	Needs Preprocessing the Data	
3	1	Data Cleaning	T1 : 61-66 / R1 : 590-594
4	1	Data Integration and	T1 : 67-71
5	1	Transformation	
6	1	Data Reduction- Data Cube Aggregation	T1 : 72-73
7	1	Data Reduction- Attribute Subset Selection	T1 : 73-75
8	1	Data Reduction- Dimensionality Reduction	T1 : 75-77
9	1	Data Reduction- Numerosity Reduction	T1 : 80-85
10	1	Discretization and Concept Hierarchy Generation for Numerical Data	T1 : 86-89
11	1	Concept Hierarchy Generation for Categorical Data	T1 : 90-96
12	1	Online Data Storage	W2
13	1	Recapitulation and Discussion of Important Questions	
Total No. of Hours Planned for -Unit II			13Hours

TEXT BOOK

1. Jiawei Han & Micheline Kamber. (2013). Data Mining – Concepts and Techniques(#3rd ed.). Mumbai: Morgan Kaufmann Publishers.

(Page Nos: 1-36 47 -94 105-148 227 -267 289 -306 318- 322 354-372 386-458 600-640)

REFERENCES

1. Michael, J.A. , Berry Gordon, S. & Linoff. (2011). Data mining Techniques (3rd ed.). Wiley Publishing Inc.

2. Michael, J.A., Berry Gordon, S., & Linoff. (2008). Mastering Data Mining (3rd ed.). New Delhi: John Wiley & Sons Inc.

UNIT-III			
1	1	**Mining Frequent Patterns, Associations and Correlations:** Basic Concepts	T1 : 227-233
2	1	Efficient and Scalable Frequent item set Mining Methods- Apriori Algorithm	T1 : 234-240
3	1	Efficient and Scalable Frequent item set Mining Methods- Mining Frequent Itemsets Using Vertical Data Format	T1 : 245-247
4	1	FP growth algorithm	T1 : 247-250
5	1	Mining various kinds of Association rules – Mining Multilevel Association Rules	T1 : 250-254
6	1	Mining various kinds of Association rules- Mining Multidimensional Association Rules	T1 : 254-256
7	1	Mining various kinds of Association rules- Mining Quantitative Association Rules	T1 : 257-258
8	1	From Association Mining to Correlation Analysis	T1 : 259-264
9	1	Constraint-Based Association Mining	T1 : 265-271
10	1	Recapitulation and Discussion of Important Questions	
Total No. of Hours Planned for -Unit III: 10 Hours			

TEXT BOOK:

1. Jiawei Han & Micheline Kamber. (2013). Data Mining – Concepts and Techniques(#3rd ed.). Mumbai: Morgan Kaufmann Publishers.

(Page Nos: 1-36 47 -94 105-148 227 -267 289 -306 318- 322 354-372 386-458 600-640)

REFERENCES:

1. Michael, J.A. , Berry Gordon, S. & Linoff. (2011). Data mining Techniques (3rd ed.). Wiley Publishing Inc.

UNIT-IV			
1	1	**Classification and Prediction:**	T1 : 289-290
2	1	Issues Regarding Classification and Prediction	T1 : 285-288
3	1	Classification by Decision Tree Induction	T1 : 289-290
4	1	Rule-based Classification	T1 : 291-299 R1 : 165-176
5	1	Rule-based Classification	T1 : 300-309
6	1	Prediction	T1 : 318-326
7	1	Accuracy and Error Measures	T1 : 354-358
8	1	Evaluating the Accuracy of a classifier or Predictor	T1 : 363-365
9	1	Ensemble Methods- increases the Accuracy	T1 : 366-369
10	1	Model Selection	T1 : 370-372
11	1	Recapitulation and Discussion of Important Questions	
Total No. of Hours Planned for -Unit IV			11 Hours

TEXT BOOK:

1. Jiawei Han & Micheline Kamber. (2013). Data Mining – Concepts and Techniques(#3rd ed.). Mumbai: Morgan Kaufmann Publishers.

(Page Nos: 1-36 47 -94 105-148 227 -267 289 -306 318- 322 354-372 386-458 600-640)

REFERENCES

1. Michael, J.A. , Berry Gordon, S. & Linoff. (2011). Data mining Techniques (3rd ed.). Wiley Publishing Inc.

UNIT-V			
1	1	**Cluster Analysis Introduction:** Types of Data in Cluster Analysis	T1 : 383-387
2	1	A Categorization of Major Clustering Methods	T1 : 398-400
3	1	Partitioning Methods	T1 : 401-407
4	1	Hierarchical Methods	T1 : 408-417
5	1	Density-Based Methods, Grid-Based Methods	T1 : 418-423
6	1	Model-Based Clustering Methods	T1 : 429-433
7	1	Clustering High-Dimensional Data	T1 : 434-443
8	1	Constraint-Based Cluster Analysis	T1 : 444-450
9	1	Outlier Analysis	T1 : 451-459
10	1	Applications and Trends in Data mining: Text Mining	W1 / T1 : 649-660
11	1	Web Mining - Multimedia Mining	T1 : 607-614
		Spatial Mining - Visual data mining	T1 : 600-607
12	1	Recapitulation and Discussion of Important Questions	
13	1	Discussion of Previous ESE Question Papers	
14	1	Discussion of Previous ESE Question Papers	
15	1	Discussion of Previous ESE Question Papers	
Total No. of Hours Planned for Unit V			15 Hours
Total No. of Periods: 60			

TEXT BOOK:

1. Jiawei Han & Micheline Kamber. (2013). Data Mining – Concepts and Techniques(#3rd ed.). Mumbai: Morgan Kaufmann Publishers.

(Page Nos: 1-36 47 -94 105-148 227 -267 289 -306 318- 322 354-372 386-458 600-640)

REFERENCES

1. Michael, J.A. , Berry Gordon, S. & Linoff. (2011). Data mining Techniques (3rd ed.). Wiley Publishing Inc.

2. Hillol Kargupta ,Anupam Joshi, Krishnamoorthy Sivakumar & Yelena Yesha. (2005). Data Mining Next Generation Challenges and Future Directions(1st ed.). New Delhi: Prentice Hall of India

UNIT I

SYLLABUS:

Introduction: Fundamentals of data mining - Data Mining Functionalities - Classification of Data Mining systems - Major issues in Data Mining.

Data Warehouse and OLAP Technology: An Overview - Data Warehouse - Multidimensional Data Model - Data Warehouse Architecture - Data Warehouse Implementation - From Data Warehousing to Data Mining.

What is data mining?

Data mining refers to extracting or mining" knowledge from large amounts of data. There are many other terms related to data mining, such as knowledge mining, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases", or KDD

Essential step in the process of knowledge discovery in databases

Knowledge discovery as a process is depicted in following figure and consists of an iterative sequence of the following steps:

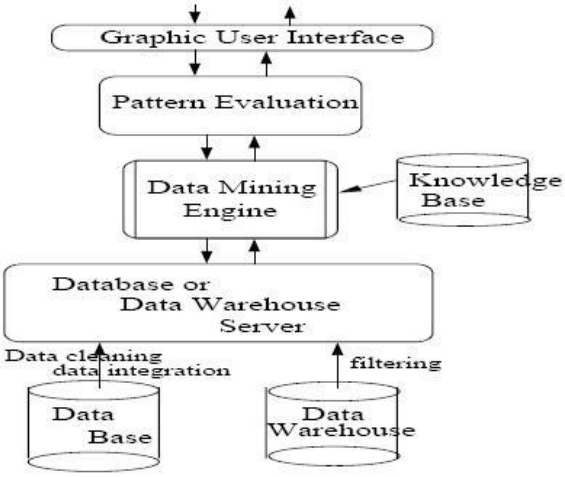
- data cleaning: to remove noise or irrelevant data
- data integration: where multiple data sources may be combined
- data selection: where data relevant to the analysis task are retrieved from the database
- data transformation: where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations
- data mining :an essential process where intelligent methods are applied in order to extract data patterns
- pattern evaluation: to identify the truly interesting patterns representing knowledge based on some interestingness measures
- knowledge presentation: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Architecture of a typical data mining system/Major Components

Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Based on this view, the architecture of a typical data mining system may have the following major components:

- A database, data warehouse, or other information repository, which consists of the set of databases, data warehouses, spreadsheets, or other kinds of information repositories containing the student and course information.
- A database or data warehouse server which fetches the relevant data based on users' data mining requests.
- A knowledge base that contains the domain knowledge used to guide the search or to evaluate the interestingness of resulting patterns. For example, the knowledge base may contain metadata which describes data from multiple heterogeneous sources.
- A data mining engine, which consists of a set of functional modules for tasks such as classification, association, classification, cluster analysis, and evolution and deviation analysis.
- A pattern evaluation module that works in tandem with the data mining modules by employing interestingness measures to help focus the search towards interestingness patterns.

- A graphical user interface that allows the user an interactive approach to the data mining system.



Architecture of a typical data mining system.

Data mining: on what kind of data?

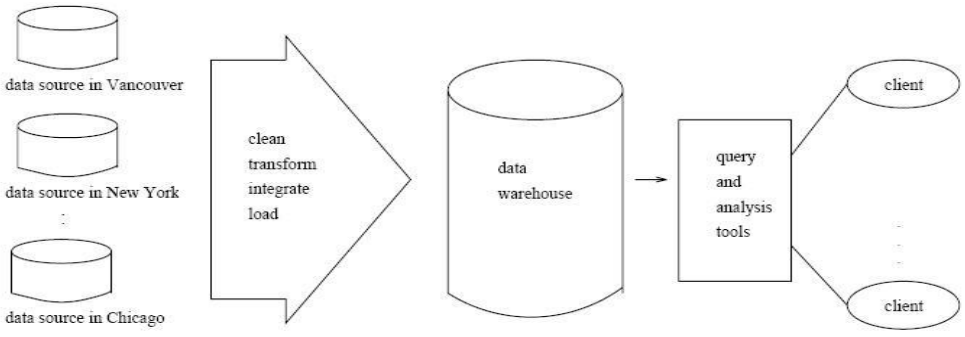
In principle, data mining should be applicable to any kind of information repository. This includes relational databases, data warehouses, transactional databases, advanced database systems, flat files, and the World-Wide Web. Advanced database systems include object-oriented and object-relational databases, and special application-oriented databases, such as spatial databases, time-series databases, text databases, and multimedia databases.

Flat files: Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific measurements, etc.

Relational Databases: a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key.

Data warehouses

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleansing, data transformation, data integration, data loading, and periodic data refreshing. The figure shows the basic architecture of a data warehouse.



Architecture of a typical data warehouse.

In order to facilitate decision making, the data in a data warehouse are organized around major subjects, such as customer, item, supplier, and activity. The data are stored to provide information from a historical perspective and are typically summarized.

A data warehouse is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or sales amount. The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube. It provides a multidimensional view of data and allows the precomputation and fast accessing of summarized data.

The data cube structure that stores the primitive or lowest level of information is called a base cuboid. Its corresponding higher level multidimensional (cube) structures are called (non-base) cuboids. A base cuboid together with all of its corresponding higher level cuboids form a data cube. By providing multidimensional data views and the precomputation of summarized data, data warehouse systems are well suited for On-Line Analytical Processing, or OLAP. OLAP operations make use of background knowledge regarding the domain of the data being studied in order to allow the presentation of data at different levels of abstraction. Such operations accommodate different user viewpoints. Examples of OLAP operations include drill-down and roll-up, which allow the user to view the data at differing degrees of summarization, as illustrated in above figure.

Transactional databases

In general, a transactional database consists of a flat file where each record represents a transaction. A transaction typically includes a unique transaction identity number (trans ID), and a list of the items making up the transaction (such as items purchased in a store) as shown below:

SALES

Trans-ID	List of item_ID's
T100 ……..	I1,I3,I8 ……

Advanced database systems and advanced database applications

An **objected-oriented database** is designed based on the object-oriented programming paradigm where data are a large number of objects organized into classes and class hierarchies.Each entity in the database is considered as an object. The object contains a set of variables that describe the object, a set of messages that the object can use to communicate with other objects or with the rest of the database system and a set of methods where each method holds the code to implement a message.

A spatial database contains spatial-related data, which may be represented in the form of raster or vector data. Raster data consists of n-dimensional bit maps or pixel maps, and vector data are represented by lines, points, polygons or other kinds of processed primitives, Some examples of spatial databases include geographical (map) databases, VLSI chip designs, and medical and satellite images databases.

Time-Series Databases: Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations

between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.

A text database is a database that contains text documents or other word descriptions in the form of long sentences or paragraphs, such as product specifications, error or bug reports, warning messages, summary reports, notes, or other documents.

A multimedia database stores images, audio, and video data, and is used in applications such as picture content-based retrieval, voice-mail systems, video-on-demand systems, the World Wide Web, and speech-based user interfaces.

The **World-Wide Web** provides rich, world-wide, on-line information services, where data objects are linked together to facilitate interactive access. Some examples of distributed information services associated with the World-Wide Web include America Online, Yahoo!, AltaVista, and Prodigy.

Data mining functionalities/Data mining tasks: what kinds of patterns can be mined?

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories:

1. Descriptive
2. predictive

Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions**.**

Concept/class description: characterization and discrimination

Data can be associated with classes or concepts. It describes a given set of data in a concise and summarative manner, presenting interesting general properties of the data. These descriptions can be derived via data characterization, by summarizing the data of the class under study (often called the target class) data discrimination, by comparison of the target class with one or a set of comparative classes both data characterization and discrimination

Data characterization

It is a summarization of the general characteristics or features of a target class of data.

Example:

A data mining system should be able to produce a description summarizing the characteristics of a student who has obtained more than 75% in every semester; the result could be a general profile of the student.

Data Discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.

Example

The general features of students with high GPA's may be compared with the general features of students with low GPA's. The resulting description could be a general comparative profile of the students such as 75% of the students with high GPA's are fourth-year computing science students while 65% of the students with low GPA's are not.

The output of data characterization can be presented in various forms. Examples include pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables, including crosstabs. The resulting descriptions can also be presented as generalized relations, or in rule form called characteristic rules.

Discrimination descriptions expressed in rule form are referred to as discriminant rules.

Mining Frequent Patterns, Association and Correlations

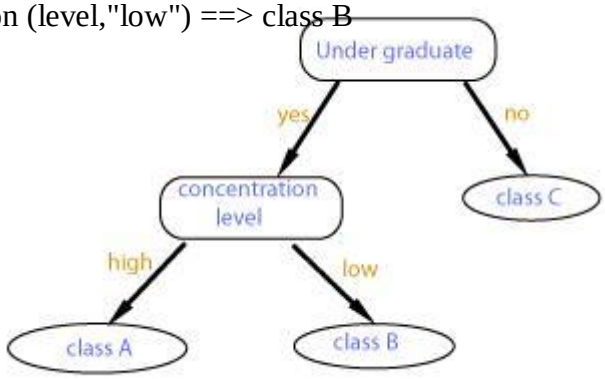
It is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. For example, a data mining system may find association rules like

major(X, “computing science””) ⇒ owns(X, “personal computer”)
[support = 12%, confidence = 98%]
where X is a variable representing a student. The rule indicates that of the students under study, 12% (support) major in computing science and own a personal computer. There is a 98% probability (confidence, or certainty) that a student in this group owns a personal computer.
Example:
A grocery store retailer to decide whether to but bread on sale. To help determine the impact of this decision, the retailer generates association rules that show what other products are frequently purchased with bread. He finds 60% of the times that bread is sold so are pretzels and that 70% of the time jelly is also sold. Based on these facts, he tries to capitalize on the association between bread, pretzels, and jelly by placing some pretzels and jelly at the end of the aisle where the bread is placed. In addition, he decides not to place either of these items on sale at the same time.

Classification and prediction
Classification:

- ✓ It predicts categorical class labels
- ✓ It classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- ✓ Typical Applications
  - ✓ credit approval
  - ✓ target marketing
  - ✓ medical diagnosis
  - ✓ treatment effectiveness analysis

Classification can be defined as the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).
Example:
An airport security screening station is used to deter mine if passengers are potential terrorist or criminals. To do this, the face of each passenger is scanned and its basic pattern(distance between eyes, size, and shape of mouth, head etc) is identified. This pattern is compared to entries in a database to see if it matches any patterns that are associated with known offenders
A classification model can be represented in various forms, such as
1) IF-THEN rules,
student (class , "undergraduate") AND concentration (level, "high") ==> class A
student (class ,"undergraduate") AND concentrtion (level,"low") ==> class B
student (class , "post graduate") ==> class C
2) Decision tree



Prediction:

Find some missing or unavailable data values rather than class labels referred to as prediction. Although prediction may refer to both data value prediction and class label prediction, it is usually confined to data value prediction and thus is distinct from classification. Prediction also encompasses the identification of distribution trends based on the available data.

Example:

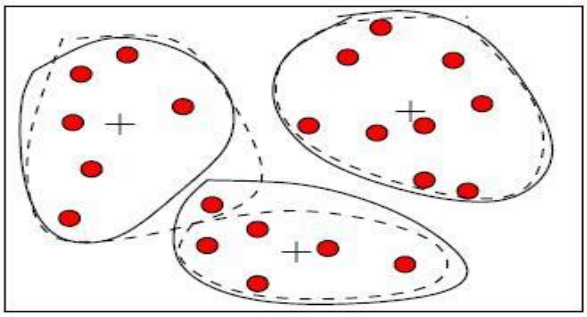
Predicting flooding is difficult problem. One approach is uses monitors placed at various points in the river. These monitors collect data relevant to flood prediction: water level, rain amount, time, humidity etc. These water levels at a potential flooding point in the river can be predicted based on the data collected by the sensors upriver from this point. The prediction must be made with respect to the time the data were collected.

Classification vs. Prediction

Classification differs from prediction in that the former is to construct a set of models (or functions) that describe and distinguish data class or concepts, whereas the latter is to predict some missing or unavailable, and often numerical, data values. Their similarity is that they are both tools for prediction: Classification is used for predicting the class label of data objects and prediction is typically used for predicting missing numerical data values.

Clustering analysis

Clustering analyzes data objects without consulting a known class label.The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.Each cluster that is formed can be viewed as a class of objects. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together as shown below:



customer data with respect to customer locations in a city, showing three data clusters. Each cluster 'center' is marked with a '+'.

Example:

A certain national department store chain creates special catalogs targeted to various demographic groups based on attributes such as income, location and physical characteristics of potential customers (age, height, weight, etc). To determine the target mailings of the various catalogs and to assist in the creation of new, more specific catalogs, the company performs a clustering of potential customers based on the determined attribute values. The results of the clustering exercise are the used by management to create special catalogs and distribute them to the correct target population based on the cluster for that catalog.

Classification vs. Clustering

- ✓ In general, in classification you have a set of predefined classes and want to know which class a new object belongs to.
- ✓ Clustering tries to group a set of objects and find whether there is *some* relationship between the objects.
- ✓ In the context of machine learning, classification is *supervised learning* and clustering is *unsupervised learning*.

Outlier analysis: A database may contain data objects that do not comply with general model of data. These data objects are outliers. In other words, the data objects which do not fall within the cluster will be called as outlier data objects. Noisy data or exceptional data are also called as outlier data. The analysis of outlier data is referred to as outlier mining.

Example

Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the location and type of purchase, or the purchase frequency.

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time.

Example:

The data of result the last several years of a college would give an idea if quality of graduated produced by it

Correlation analysis

Correlation analysis is a technique use to measure the association between two variables. A **correlation coefficient** (**r**) is a statistic used for measuring the strength of **a** supposed linear association between two variables. Correlations range from -1.0 to +1.0 in value.

- ✓ A correlation coefficient of 1.0 indicates a perfect positive relationship in which high values of one variable are related perfectly to high values in the other variable, and conversely, low values on one variable are perfectly related to low values on the other variable.
- ✓ A correlation coefficient of 0.0 indicates no relationship between the two variables. That is, one cannot use the scores on one variable to tell anything about the scores on the second variable.
- ✓ A correlation coefficient of -1.0 indicates a perfect negative relationship in which high values of one variable are related perfectly to low values in the other variables, and conversely, low values in one variable are perfectly related to high values on the other variable.

Are all of the patterns interesting? / What makes a pattern interesting?

A pattern is interesting if,

- ✓ It is easily understood by humans,
- ✓ Valid on new or test data with some degree of certainty,
- ✓ Potentially useful, and
- ✓ Novel.

A pattern is also interesting if it validates a hypothesis that the user sought to confirm. An interesting pattern represents knowledge.

- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Classification of data mining systems

There are many data mining systems available or being developed. Some are specialized systems dedicated to a given data source or are confined to limited data mining functionalities, other are more versatile and comprehensive. Data mining systems can be categorized according to various criteria among other classification are the following:

Classification according to the type of data source mined: this classification categorizes data mining systems according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

Classification according to the data model drawn on: this classification categorizes data mining systems based on the data model involved such as relational database, object-oriented database, data warehouse, transactional, etc.

Classification according to the king of knowledge discovered: this classification categorizes data mining systems based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

Classification according to mining techniques used: Data mining systems **employ** and provide different techniques. This classification categorizes data mining systems according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc. The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

Five primitives for specifying a data mining task

Task-relevant data: This primitive specifies the data upon which mining **is** to be performed. It involves specifying the database and tables or data warehouse containing the relevant data, conditions for selecting the relevant data, the relevant attributes or dimensions for exploration, and instructions regarding the ordering or grouping of the data retrieved.

Knowledge type to be mined: This primitive specifies the specific data mining function to be performed, such as characterization, discrimination, association, classification, clustering, or evolution analysis. As well, the user can be more specific and provide pattern templates that all discovered patterns must match. These templates or meta patterns (also called meta rules or meta queries), can be used to guide the discovery process.

Background knowledge: This primitive allows users to specify knowledge they have about the domain to be mined. Such knowledge can be used to guide the knowledge discovery process and evaluate the patterns that are found. Of the several kinds of background knowledge, this chapter focuses on concept hierarchies.

Pattern interestingness measure: This primitive allows users to specify functions that are used to separate uninteresting patterns from knowledge and may be used to guide

the mining process, as well as to evaluate the discovered patterns. This allows the user to confine the number of uninteresting patterns returned by the process, as a data mining process may generate a large number of patterns. Interestingness measures can be specified for such pattern characteristics as simplicity, certainty, utility and novelty.

Visualization of discovered patterns: This primitive refers to the form in which discovered patterns are to be displayed. In order for data mining to be effective in conveying knowledge to users, data mining systems should be able to display the discovered patterns in multiple forms such as rules, tables, cross tabs (cross-tabulations), pie or bar charts, decision trees, cubes or other visual representations.

Integration of a Data Mining System with a Database or Data Warehouse System

The differences between the following architectures for the integration of a data mining system with a database or data warehouse system are as follows.

No coupling: The data mining system uses sources such as flat files to obtain the initial data set to be mined since no database system or data warehouse system functions are implemented as part of the process. Thus, this architecture represents a poor design choice.

Loose coupling:

The data mining system is not integrated with the database or data warehouse system beyond their use as the source of the initial data set to be mined, and possible use in storage of the results. Thus, this architecture can take advantage of the flexibility, efficiency and features such as indexing that the database and data warehousing systems may provide. However, it is difficult for loose coupling to achieve high scalability and good performance with large data sets as many such systems are memory-based.

Semitight coupling:

Some of the data mining primitives such as aggregation, sorting or pre computation of statistical functions are efficiently implemented in the database or data warehouse system, for use by the data mining system during mining-query processing. Also, some frequently used inter mediate mining results can be pre computed and stored in the database or data warehouse system, thereby enhancing the performance of the data mining system.

Tight coupling:

The database or data warehouse system is fully integrated as part of the data mining system and thereby provides optimized data mining query processing. Thus, the datamining sub system is treated as one functional component of an information system. This is a highly desirable architecture as it facilitates efficient implementations of data mining functions, high system performance, and an integrated information processing environment

From the descriptions of the architectures provided above, it can be seen that tight coupling is the best alternative without respect to technical or implementation issues. However, as much of the technical infrastructure needed in a tightly coupled system is still evolving, implementation of such a system is non-trivial. Therefore, the most popular architecture is currently semi tight coupling as it provides a compromise between loose and tight coupling.

Major issues in data mining

Major issues in data mining is regarding mining methodology, user interaction, performance, and diverse data types

Mining methodology and user-interaction issues:

_ Mining different kinds of knowledge in databases: Since different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of

data analysis and knowledge discovery tasks, including data characterization, discrimination, association, classification, clustering, trend and deviation analysis, and similarity analysis. These tasks may use the same database in different ways and require the development of numerous data mining techniques.

_ **Interactive mining of knowledge at multiple levels of abstraction**: Since it is difficult to know exactly what can be discovered within a database, the data mining process should be interactive.

_ **Incorporation of background knowledge:** Background knowledge, or information regarding the domain under study, may be used to guide the discovery patterns. Domain knowledge related to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.

_ Data mining query languages and ad-hoc data mining: Knowledge in Relational query languages (such as SQL) required since it allow users to pose ad-hoc queries for data retrieval.

Presentation and visualization of data mining results: Discovered knowledge should be expressed in high-level languages, visual representations, so that the knowledge can be easily understood and directly usable by humans

_ Handling outlier or incomplete data: The data stored in a database may reflect outliers: noise, exceptional cases, or incomplete data objects. These objects may confuse the analysis process, causing over fitting of the data to the knowledge model constructed. As a result, the accuracy of the discovered patterns can be poor. Data cleaning methods and data analysis methods which can handle outliers are required.

_ Pattern evaluation: refers to interestingness of pattern: A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, representing common knowledge or lacking novelty. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns,

Performance issues. These include efficiency, scalability, and parallelization **of** data mining algorithms.

_ Efficiency and scalability of data mining algorithms: To effectively extract **information** from a huge amount of data in databases, data mining algorithms must be efficient and scalable.

_ Parallel, distributed, and incremental updating algorithms: Such **algorithms** divide the data into partitions, which are processed in parallel. The results from the partitions are then merged.

Issues relating to the diversity of database types

_ Handling of relational and complex types of data: Since relational databases **and** data warehouses are widely used, the development of efficient and effective data mining systems for such data is important.

_ Mining information from heterogeneous databases and global information systems: Local and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi-structured, or unstructured data with diverse data semantics poses great challenges to data mining.

What is Data Warehouse?

Data Warehouse Introduction

A data warehouse is a collection of data marts representing historical data from different operations in the company. This data is stored in a structure optimized for querying and

data analysis as a data warehouse. Table design, dimensions and organization should be consistent throughout a data warehouse so that reports or queries across the data warehouse are consistent. A data warehouse can also be viewed as a database for historical data from different functions within a company.

The term Data Warehouse was coined by Bill Inmon in 1990, which he defined in the following way: "A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process". He defined the terms in the sentence as follows:

Subject Oriented: Data that gives information about a particular subject instead of about a company's ongoing operations.

Integrated: Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.

Time-variant: All data in the data warehouse is identified with a particular time period.

Non-volatile: Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business. It is a single, complete and consistent store of data obtained from a variety of different sources made available to end users in what they can understand and use in a business context. It can be

- Used for decision Support
- Used to manage and control business
- Used by managers and end-users to understand the business and make judgments

Data Warehousing is an architectural construct of information systems that provides users with current and historical decision support information that is hard to access or present in traditional operational data stores.

Other important terminology

Enterprise Data warehouse: It collects all information about subjects (*customers, products, sales, assets, personnel*) that span the entire organization

Data Mart: Departmental subsets that focus on selected subjects. A data mart is a segment of a data warehouse that can provide data for reporting and analysis on a section, unit, department or operation in the company, e.g. sales, payroll, production. Data marts are sometimes complete individual data warehouses which are usually smaller than the corporate data warehouse.

Decision Support System (DSS):Information technology to help the knowledge worker (executive, manager, and analyst) makes faster & better decisions

Drill-down: Traversing the summarization levels from highly summarized data to the underlying current or old detail

Metadata: Data about data. Containing location and description of warehouse system components: names, definition, structure…

Benefits of data warehousing

- Data warehouses are designed to perform well with aggregate queries running on large amounts of data.
- The structure of data warehouses is easier for end users to navigate, understand and query against unlike the relational databases primarily designed to handle lots of transactions.
- Data warehouses enable queries that cut across different segments of a company's operation. E.g. production data could be compared against inventory data even if they were originally stored in different databases with different structures.

- Queries that would be complex in very normalized databases could be easier to build and maintain in data warehouses, decreasing the workload on transaction systems.
- Data warehousing is an efficient way to manage and report on data that is from a variety of sources, non uniform and scattered throughout a company.
- Data warehousing is an efficient way to manage demand for lots of information from lots of users.
- Data warehousing provides the capability to analyze large amounts of historical data for nuggets of wisdom that can provide an organization with competitive advantage.

Data Warehouse Characteristics

A data warehouse can be viewed as an information system with the following attributes:

- It is a database designed for analytical tasks
- It‘s content is periodically updated
- It contains current and historical data to provide a historical perspective of information

Features of OLTP and OLAP

The major distinguishing features between OLTP and OLAP are summarized as follows.

Users and system orientation: An OLTP system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals. An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts.

Data contents: An OLTP system manages current data that, typically, are too detailed to be easily used for decision making. An OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity. These features make the data easier for use in informed decision making.

Database design: An OLTP system usually adopts an entity-relationship (ER) data model and an application oriented database design. An OLAP system typically adopts either a star or snowflake model and a subject-oriented database design.

View: An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations. In contrast, an OLAP system often spans multiple versions of a database schema. OLAP systems also deal with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, OLAP data are stored on multiple storage media.

Access patterns: The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly read-only operations although many could be complex queries.

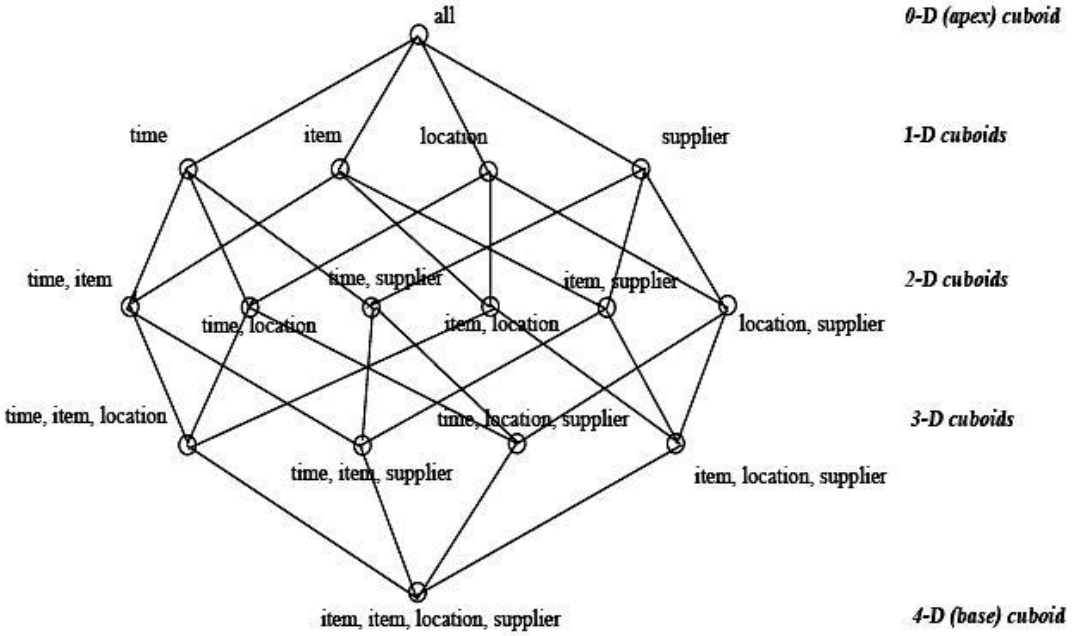
Comparison between OLTP and OLAP systems.

Feature	OLTP	OLAP
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, anal
Function	day-to-day operations	long term informational requirements, decision support
DB design	E-R based, application-oriented	star/snowflake, subject-oriented
Data	current; guaranteed up-to-date	historical; accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
# of records accessed	tens	millions
# of users	thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

A Multidimensional Data Model.

The most popular data model for data warehouses is a multidimensional model. This model can exist in the form of a star schema, a snowflake schema, or a fact constellation schema. Let's have a look at each of these schema types.

From Tables and Spreadsheets to Data Cubes



Stars,Snowflakes,and Fact Constellations:
Schemas for Multidimensional Databases

Star schema: The star schema is a modeling paradigm in which the data warehouse contains (1) a large central table (fact table), and (2) a set of smaller attendant tables (dimension tables), one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

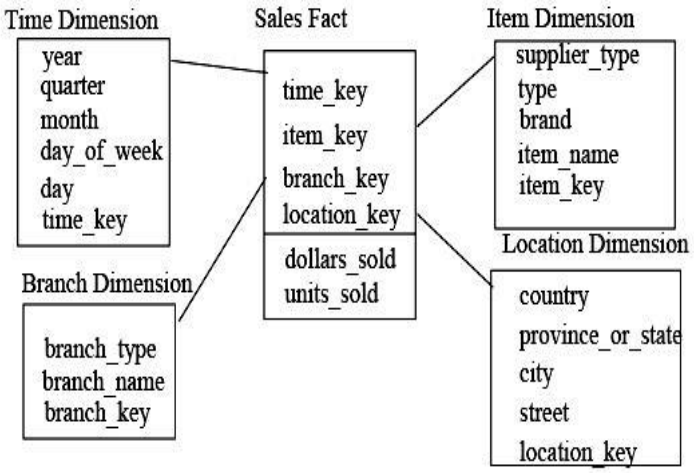
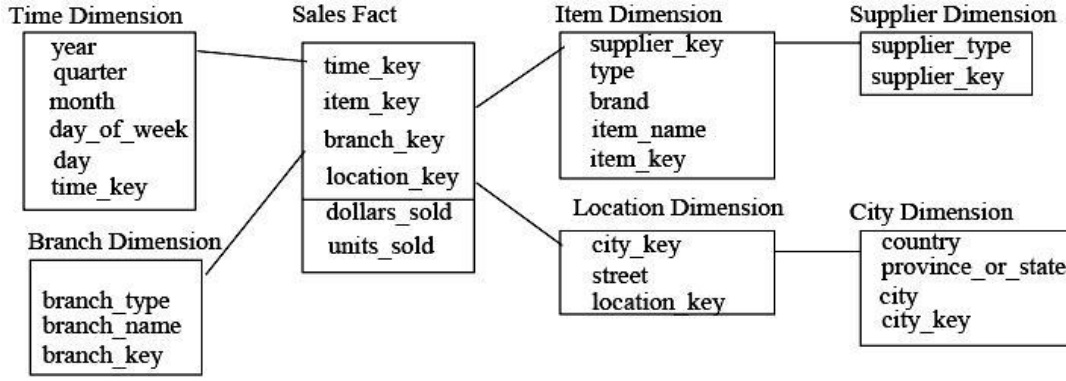


Figure Star schema of a data warehouse for sales.

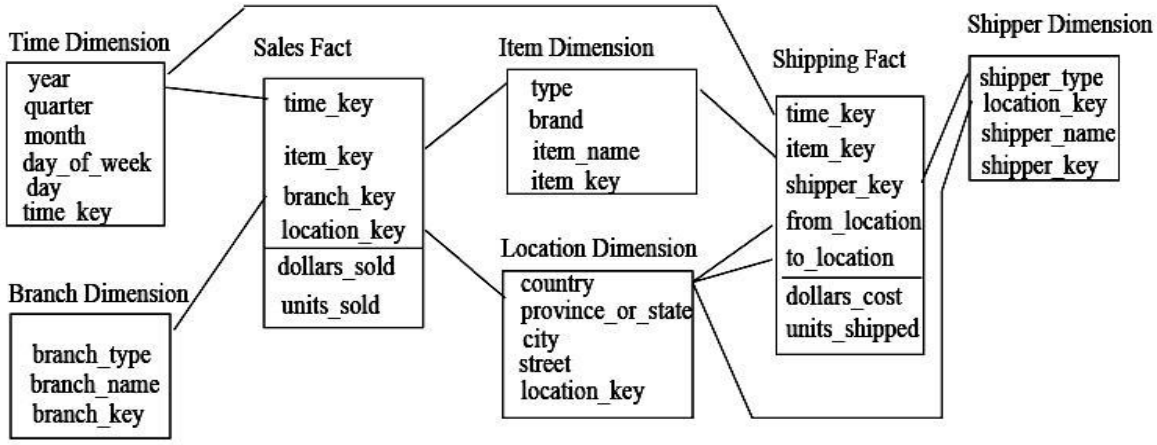
Snowflake schema: The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake. The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form. Such a table is easy to maintain and also saves storage space because a large dimension table can be extremely large when the dimensional structure is included as columns.

Figure Snowflake schema of a data warehouse for sales.



Fact constellation: Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.

Figure Fact constellation schema of a data warehouse for sales and shipping.



Measures: Three Categories

Measure: a function evaluated on aggregated data corresponding to given dimension-value pairs. Measures can be:

- distributive: if the measure can be calculated in a distributive manner.
  - E.g., count(), sum(), min(), max().
- algebraic: if it can be computed from arguments obtained by applying distributive aggregate functions.
  - E.g., avg()=sum()/count(), min_N(), standard_deviation().
- holistic: if it is not algebraic.
  - E.g., median(), mode(), rank().

A Concept Hierarchy

A Concept hierarchy defines a sequence of mappings from a set of low level Concepts to higher level, more general Concepts. Concept hierarchies allow data to be handled at varying levels of abstraction

OLAP operations on multidimensional data.

Roll-up: The roll-up operation performs aggregation on a data cube, either by climbing-up a concept hierarchy for a dimension or by dimension reduction. Figure shows the result of a roll-up operation performed on the central cube by climbing up the concept hierarchy for location. This hierarchy was defined as the total order street < city < province or state <country.

Drill-down: Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping-down a concept hierarchy for a dimension or introducing additional dimensions. Figure shows the result of a drill-down operation performed on the central cube by stepping down a concept hierarchy for time defined as day < month < quarter < year. Drill-down occurs by descending the time hierarchy from the level of quarter to the more detailed level of month.

Slice and dice: The slice operation performs a selection on one dimension of the given cube, resulting in a sub cube. Figure shows a slice operation where the sales data are selected from the central cube for the dimension time using the criteria time=”Q2". The dice operation defines a sub cube by performing a selection on two or more dimensions.

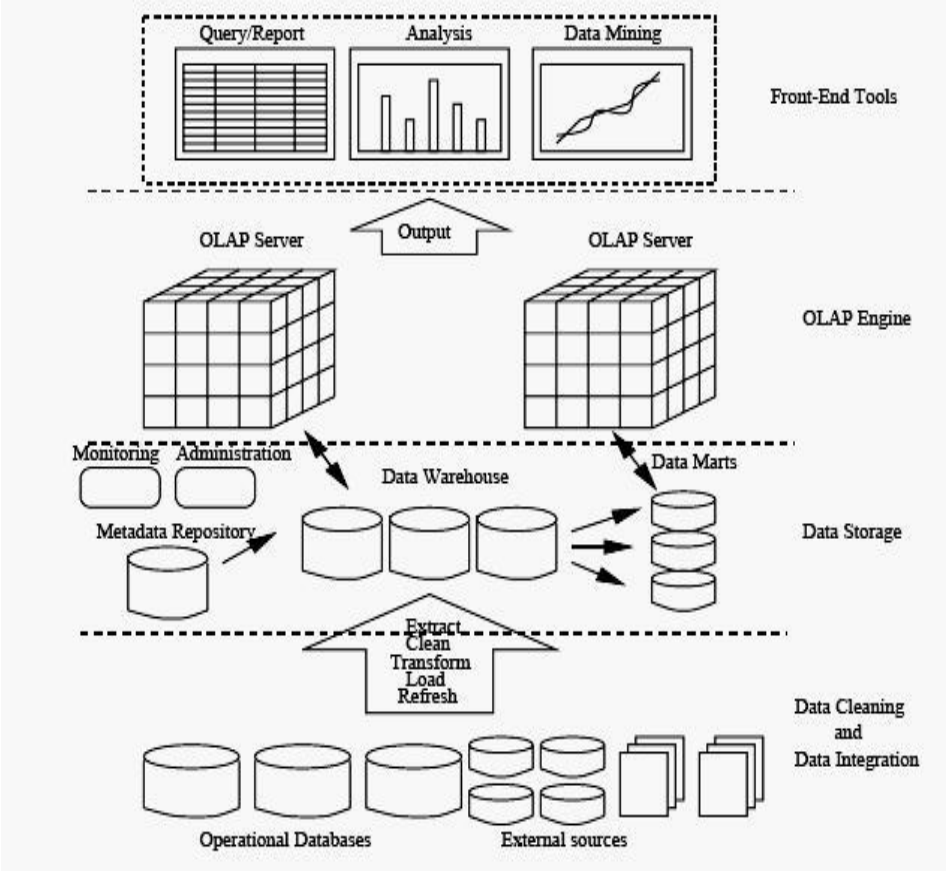
Pivot (rotate): Pivot is a visualization operation which rotates the data axes in view in order to provide an alternative presentation of the data. Figure shows a pivot operation where the item and location axes in a 2-D slice are rotated.

Data warehouse architecture

The Design of a Data Warehouse: A Business Analysis Framework Four different views regarding the design of a data warehouse must be considered: the top-down view, the data source view, the data warehouse view, the business query view.

- The top-down view allows the selection of relevant information necessary for the data warehouse.
- The data source view exposes the information being captured, stored and managed by operational systems.
- The data warehouse view includes fact tables and dimension tables
- Finally the business query view is the Perspective of data in the data warehouse from the viewpoint of the end user.

Three-tier Data warehouse architecture



The bottom tier is ware-house database server which is almost always a relational database system. The middle tier is an OLAP server which is typically implemented using either (1) a Relational OLAP (ROLAP) model, (2) a Multidimensional OLAP (MOLAP) model. The top tier is a client, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

From the architecture point of view, there are three data warehouse models: the enterprise warehouse, the data mart, and the virtual warehouse.

Enterprise warehouse: An enterprise warehouse collects all of the information about subjects spanning the entire organization. It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope. It typically contains detailed data as well as summarized data, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond.

Data mart: A data mart contains a subset of corporate-wide data that is of value to a specific group of users. The scope is connected to specific, selected subjects. For example, a marketing data mart may connect its subjects to customer, item, and sales. The data contained in data marts tend to be summarized. Depending on the source of data, data marts can be categorized into the following two classes:

(i).Independent data marts are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area.

(ii).Dependent data marts are sourced directly from enterprise data warehouses.

Virtual warehouse: A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized. A virtual warehouse is easy to build but requires excess capacity on operational database servers.

From Data Warehousing to Data mining

Data Warehouse Usage:

Three kinds of data warehouse applications

- Information processing
  - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
- Analytical processing
  - multidimensional analysis of data warehouse data
  - supports basic OLAP operations, slice-dice, drilling, pivoting

- Data mining
  - knowledge discovery from hidden patterns
  - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

Differences among the three tasks

Note:

From On-Line Analytical Processing to On Line Analytical Mining (OLAM) called from data warehousing to data mining

From on-line analytical processing to on-line analytical mining.

On-Line Analytical Mining (OLAM) (also called OLAP mining), which integrates on-line analytical processing (OLAP) with data mining and mining knowledge in multidimensional databases, is particularly important for the following reasons.

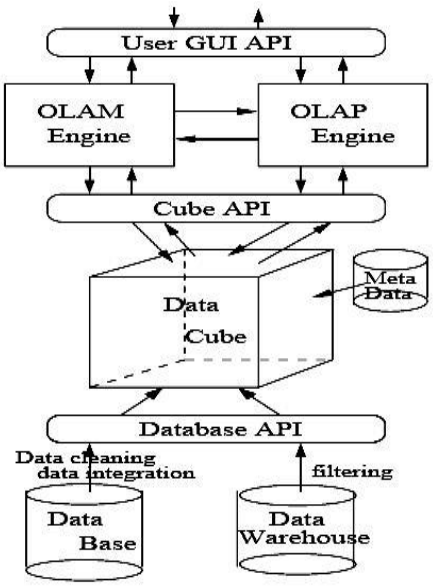
1. High quality of data in data warehouses.

Most data mining tools need to work on integrated, consistent, and cleaned data, which requires costly data cleaning, data transformation and data integration as preprocessing steps. A data warehouse constructed by such preprocessing serves as a valuable source of high quality data for OLAP as well as for data mining.

2. Available information processing infrastructure surrounding data warehouses.
Comprehensive information processing and data analysis infrastructures have been or will be systematically constructed surrounding data warehouses, which include accessing, integration, consolidation, and transformation of multiple, heterogeneous databases, ODBC/OLEDB connections, Web-accessing and service facilities, reporting and OLAP analysis tools.
3. OLAP-based exploratory data analysis.
Effective data mining needs exploratory data analysis. A user will often want to traverse through a database, select portions of relevant data, analyze them at different granularities, and present knowledge/results in different forms. On-line analytical mining provides facilities for data mining on different subsets of data and at different levels of abstraction, by drilling, pivoting, filtering, dicing and slicing on a data cube and on some intermediate data mining results.
4. On-line selection of data mining functions.
By integrating OLAP with multiple data mining functions, on-line analytical mining provides users with the ability to select desired data mining functions and swap data mining tasks dynamically.

Architecture for on-line analytical mining

An OLAM engine performs analytical mining in data cubes in a similar manner as an OLAP engine performs on-line analytical processing. An integrated OLAM and OLAP architecture is shown in Figure, where the OLAM and OLAP engines both accept users' on-line queries via a User GUI API and work with the data cube in the data analysis via a Cube API.
A metadata directory is used to guide the access of the data cube. The data cube can be constructed by accessing and/or integrating multiple databases and/or by filtering a data warehouse via a Database API which may support OLEDB or ODBC connections. Since an OLAM engine may perform multiple data mining tasks, such as concept description, association, classification, prediction, clustering, time-series analysis ,etc., it usually consists of multiple, integrated data mining modules and is more sophisticated than an OLAP engine.
Figure: An integrated OLAM and OLAP architecture.



Possible Questions:

2 Marks Questions:

1. What is data mining?
2. List out the classification of data mining systems
3. What is KDD?
4. Differentiate OLAP and OLTP
5. What kind data is stored in Spatiotemporal Databases?
6. What is data ware house?
7. Define data cubes
8. List the schemas for multi dimensional data bases.
9. What is a data mart?
10. Define OLAM.

Part B:

1. With a neat sketch enlighten the architecture of a data warehouse
2. Give the classification of Data Mining Systems?
3. Describe multidimensional data model. How is it used in data warehousing?
4. Explain the data mining functionalities in detail.
5. Describe the various issues in data mining systems.
6. Describe the architecture of typical data mining system with a neat sketch.
7. Explain the distinction between:
   (i) Measures and Dimensions
   (ii) Fact tables and Dimension tables
   (iii) Star and Snowflake data warehousing Schemes

Part-C: Compulsory Questions (10 marks)

1. Briefly compare the following concepts. You may use an example to explain your point(s).
   (a) Snowflake schema, fact constellation,
   (b) Data cleaning, data transformation,
   (c) Enterprise warehouse, data mart, virtual warehouse
2. Describe various OLAP operations on multidimensional data and illustrate those using examples?
3. "Data Mining is the multi-disciplinary field." Discuss.

UNIT-II

SYLLABUS:

Data Preprocessing: Needs Preprocessing the Data - Data Cleaning - Data Integration and Transformation - Data Reduction - Discretization and Concept Hierarchy Generation - Online Data Storage.

Data preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user

Why Data Preprocessing?

Data in the real world is dirty. It can be in incomplete, noisy and inconsistent from. These data needs to be preprocessed in order to help improve the quality of the data, and quality of the mining results.

- If no quality data , then no quality mining results. The quality decision is always based on the quality data.
- If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult

Incomplete data: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. e.g., occupation=“ ”.

Noisy data: containing errors or outliers data. e.g., Salary=“-10”

Inconsistent data: containing discrepancies in codes or names. e.g., Age=“42” Birthday=“03/07/1997”

Incomplete data may come from

- “Not applicable” data value when collected
- Different considerations between the time when the data was collected and when it is analyzed.
- Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection by instruments
  - Human or computer error at data entry
  - Errors in data transmission

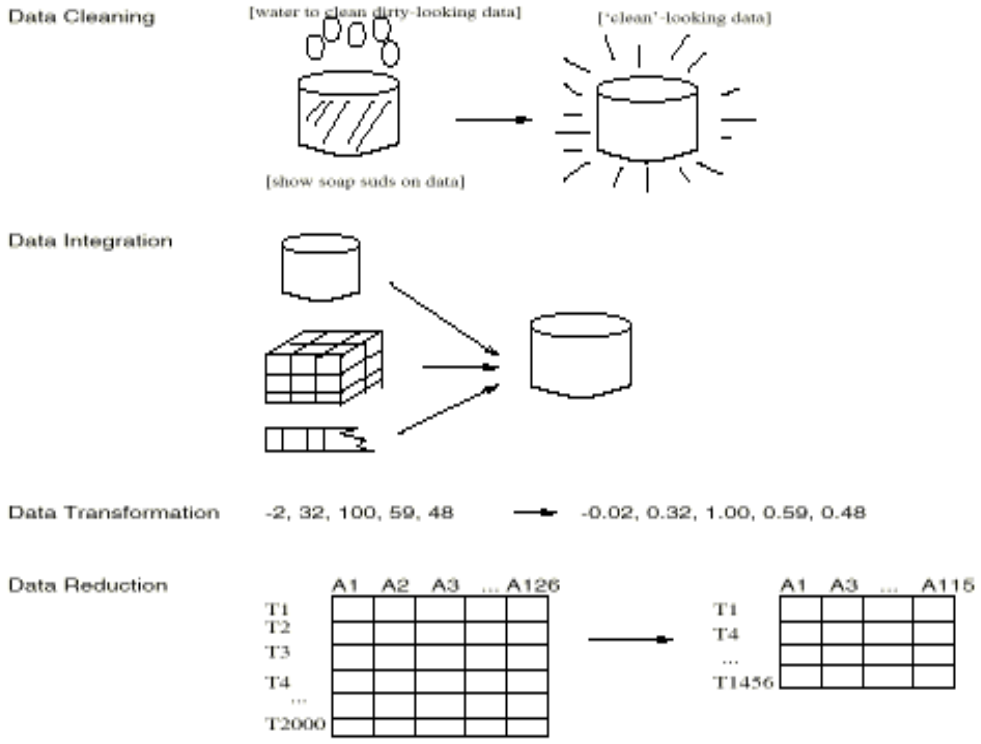
- Inconsistent data may come from
- Different data sources
- Functional dependency violation (e.g., modify some linked data)

Major Tasks in Data Preprocessing

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

- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
- Part of data reduction but with particular importance, especially for numerical data

Forms of Data Preprocessing



Descriptive Data Summarization

Categorize the measures

- A measure is distributive, if we can partition the dataset into smaller subsets, compute the measure on the individual subsets, and then combine the partial results in order to arrive at the measure's value on the entire (original) dataset
- A measure is algebraic if it can be computed by applying an algebraic function to one or more distributive measures
- A measure is holistic if it must be computed on the entire dataset as a whole

Measure the Central Tendency

- A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location.
- In other words, in many real-life situations, it is helpful to describe data by a single number that is most representative of the entire collection of numbers. Such a number is called a measure of central tendency. The most commonly used measures are as follows. **Mean, Median, and Mode**

Mean: mean, or average, of numbers is the sum of the numbers divided by n. That is:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n} \quad \text{i.e.,} \quad \text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

shortly,

$$\bar{x} = \frac{\sum x}{n}$$

where $\bar{x}$ (read as 'x bar') is the mean of the set of x values,
$\sum x$ is the sum of all the x values, and
n is the number of x values.

Example 1

The marks of seven students in a mathematics test with a maximum possible mark of 20 are given below:

15 13 18 16 14 17 12

Find the mean of this set of data values.

Solution:

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$
$$= \frac{15+13+18+16+14+17+12}{7}$$
$$= \frac{105}{7}$$
$$= 15$$

So, the mean mark is 15.

Midrange

The midrange of a data set is the average of the minimum and maximum values.

Median: median of numbers is the middle number when the numbers are written in order.

If is even, the median is the average of the two middle numbers.

In general:

$$\text{Median} = \frac{1}{2}(n+1)\text{ th value, where } n \text{ is the number of data values in the sample}$$

If the number of values in the data set is even, then the **median** is the average of the two middle values.

Mode of numbers is the number that occurs most frequently. If two numbers tie for most frequent occurrence, the collection has two modes and is called bimodal.

The mode has applications in printing . For example, it is important to print more of the most popular books; because printing different books in equal numbers would cause a shortage of some books and an oversupply of others.

Likewise, the mode has applications in manufacturing. For example, it is important to manufacture more of the most popular shoes; because manufacturing different shoes in equal numbers would cause a shortage of some shoes and an oversupply of others.

Example

Find the mode of the following data set:

48 44 48 45 42 49 48

Solution:

The mode is 48 since it occurs most often.

It is possible for a set of data values to have more than one mode.

- If there are two data values that occur most frequently, we say that the set of data values is **bimodal**.
- If there is three data values that occur most frequently, we say that the set of data values is **trimodal**
- If two or more data values that occur most frequently, we say that the set of data values is **multimodal**
- If there is no data value or data values that occur most frequently, we say that the set of data values has no mode.
- The mean, median and mode of a data set are collectively known as measures of **central tendency** as these three measures focus on where the data is centered or clustered. To analyze data using the mean, median and mode, we need to use the most appropriate measure of central tendency. The following points should be remembered:

The mean is useful for predicting future results when there are no extreme values in the data set. However, the impact of extreme values on the mean may be important and should be considered. E.g. the impact of a stock market crash on average investment returns.
The median may be more useful than the mean when there are extreme values in the data set as it is not affected by the extreme values.
The mode is useful when the most common item, characteristic or value of a data set is required.

Measures of Dispersion

Measures of dispersion measure how spread out a set of data is. The two most commonly used measures of dispersion are the variance and the standard deviation. Rather than showing how data are similar, they show how data differs from its variation, spread, or dispersion.
Other measures of dispersion that may be encountered include the Quartiles, Inter quartile range (IQR), Five number summary, range and box plots

Variance and Standard Deviation

Very different sets of numbers can have the same mean. You will now study two measures of dispersion, which give you an idea of how much the numbers in a set differ from the mean of the set. These two measures are called the variance of the set and the standard deviation of the set

Consider a set of numbers $\{x_1, x_2, \ldots, x_n\}$ with a mean of $\bar{x}$. The **variance** of the set is

$$v = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

and the **standard deviation** of the set is $\sigma = \sqrt{v}$ (σ is the lowercase Greek letter *sigma*).

The standard deviation of a set is a measure of how much a typical number in the set differs from the mean. The greater the standard deviation, the more the numbers in the set *vary* from the mean. For instance, each of the following sets has a mean of 5.

$\{5, 5, 5, 5\}$, $\{4, 4, 6, 6\}$, and $\{3, 3, 7, 7\}$

The standard deviations of the sets are 0, 1, and 2.

$$\sigma_1 = \sqrt{\frac{(5-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2}{4}} = 0$$

$$\sigma_2 = \sqrt{\frac{(4-5)^2 + (4-5)^2 + (6-5)^2 + (6-5)^2}{4}} = 1$$

$$\sigma_3 = \sqrt{\frac{(3-5)^2 + (3-5)^2 + (7-5)^2 + (7-5)^2}{4}} = 2$$

Percentile

- Percentiles are values that divide a sample of data into one hundred groups containing (as far as possible) equal numbers of observations.
- The pth percentile of a distribution is the value such that p percent of the observations fall at or below it.
- The most commonly used percentiles other than the median are the 25th percentile and the 75th percentile.
- The 25th percentile demarcates the first quartile, the median or 50th percentile demarcates the second quartile, the 75th percentile demarcates the third quartile, and the 100th percentile demarcates the fourth quartile.

Quartiles

Quartiles are numbers that divide an ordered data set into four portions, each containing approximately one-fourth of the data. Twenty-five percent of the data values come before the first quartile (Q1). The median is the second quartile (Q2); 50% of the data values come before the median. Seventy-five percent of the data values come before the third quartile (Q3).

Q1=25th percentile=(n*25/100), where n is total number of data in the given data set
Q2=median=50th percentile=(n*50/100)
Q3=75th percentile=(n*75/100)

Inter quartile range (IQR)

The inter quartile range is the length of the interval between the lower quartile (Q1) and the upper quartile (Q3). This interval indicates the central, or middle, 50% of a data set.

I Q R = Q 3 - Q 1

Range

The range of a set of data is the difference between its largest (maximum) and smallest (minimum) values. In the statistical world, the range is reported as a single number, the difference between maximum and minimum. Sometimes, the range is often reported as "from (the minimum) to (the maximum)," i.e., two numbers.

Example:

Given data set: 3, 4, 4, 5, 6, 8

The range of data set is 3–8. The range gives only minimal information about the spread of the data, by defining the two extremes. It says nothing about how the data are distributed between those two endpoints.

Box plots

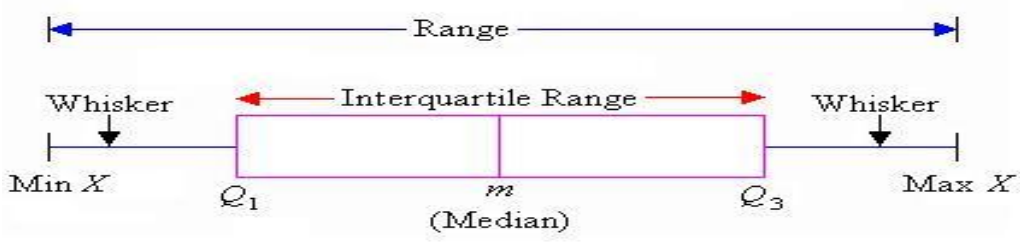
A box plot is a graph used to represent the range, median, quartiles and inter quartile range of a set of data values.

Constructing a Box plot: To construct a box plot:

Draw a box to represent the middle 50% of the observations of the data set.

Show the median by drawing a vertical line within the box.

Draw the lines (called **whiskers**) from the lower and upper ends of the box to the minimum and maximum values of the data set respectively, as shown in the following diagram.



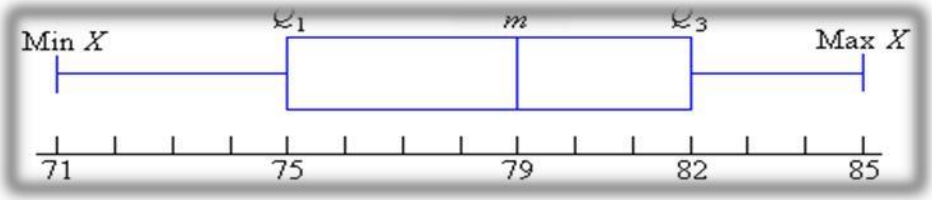
X is the set of data values.
Min X is the minimum value in the data set.
Max X is the maximum value in the data set.
Example: Draw a boxplot for the following data set of scores:
76 79 76 74 75 71 85 82 82 79 81
Step 1: Arrange the score values in ascending order of magnitude:
71 74 75 76 76 79 79 81 82 82 85
There are 11 values in the data set.
Step 2: Q1=25th percentile value in the given data set

Q1=11*(25/100) th value
=2.75 =>3rd value
=75

Step 3: Q2=median=50th percentile value
=11 * (50/100) th value
=5.5th value => 6th value
=79

Step 4: Q3=75th percentile value
=11*(75/100)th value
=8.25th value=>9th value
= 82

Step 5: Min X= 71
Step 6: Max X=85
Step 7: Range= 85-71 = 14
Step 5: IQR=height of the box=Q3-Q1=9-3=6th value=79



Since the medians represent the middle points, they split the data into four equal parts. In other words:

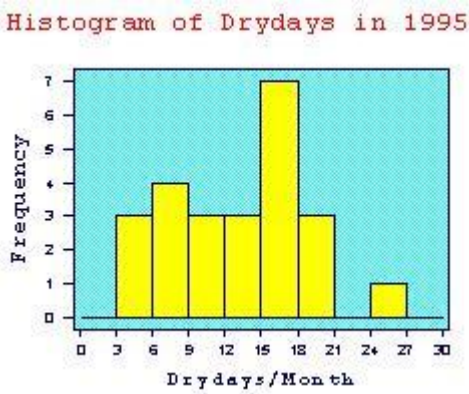
- one quarter of the data numbers are less than 75
- one quarter of the data numbers are between 75 and 79
- one quarter of the data numbers are between 79 and 82
- one quarter of the data numbers are greater than 82

Outliers

Outlier data is a data that falls outside the range. Outliers will be any points below Q1 – 1.5×IQR or above Q3 + 1.5×IQR.

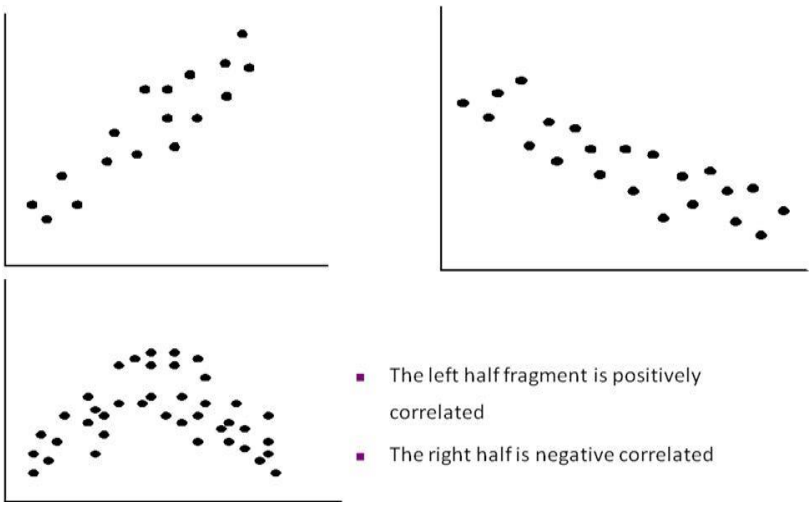
Example:
Find the outliers, if any, for the following data set:
10.2, 14.1, 14.4, **14.4**, 14.4, 14.5, 14.5, **14.6,** 14.7, 14.7, 14.7, **14.9,** 15.1, 15.9, 16.4
To find out if there are any outliers, I first have to find the IQR. There are fifteen data points, so the median will be at position (15/2) = 7.5=8th value=14.6. That is, Q2 = 14.6.
Q1 is the fourth value in the list and Q3 is the twelfth: Q1 = 14.4 and Q3 = 14.9.
Then IQR = 14.9 – 14.4 = 0.5.
Outliers will be any points below:
Q1 – 1.5×IQR = 14.4 – 0.75 = 13.65 or above Q3 + 1.5×IQR = 14.9 + 0.75 = 15.65.
Then the outliers are at 10.2, 15.9, and 16.4.
The values for Q1 – 1.5×IQR and Q3 + 1.5×IQR are the "fences" that mark off the "reasonable" values from the outlier values. Outliers lie outside the fences.

Graphic Displays of Basic Descriptive Data Summaries
1 Histogram
A histogram is a way of summarizing data that are measured on an interval scale (either discrete or continuous). It is often used in exploratory data analysis to illustrate the major features of the distribution of the data in a convenient form. It divides up the range of possible values in a data set into classes or groups. For each group, a rectangle is constructed with a base length equal to the range of values in that specific group, and an area proportional to the number of observations falling into that group. This means that the rectangles might be drawn of non-uniform height.



The histogram is only appropriate for variables whose values are numerical and measured on an interval scale. It is generally used when dealing with large data sets (>100 observations)
A histogram can also help detect any unusual observations (outliers), or any gaps in the data set.
2 Scatter Plot
A scatter plot is a useful summary of a set of bivariate data (two variables), usually drawn before working out a linear correlation coefficient or fitting a regression line. It gives a good visual picture of the relationship between the two variables, and aids the interpretation of the correlation coefficient or regression model.
Each unit contributes one point to the scatter plot, on which points are plotted but not joined. The resulting pattern indicates the type and strength of the relationship between the two variables.

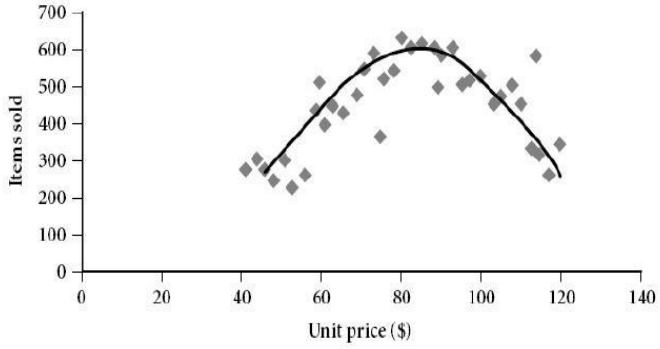
Positively and Negatively Correlated Data



A scatter plot will also show up a non-linear relationship between the two variables and whether or not there exist any outliers in the data.

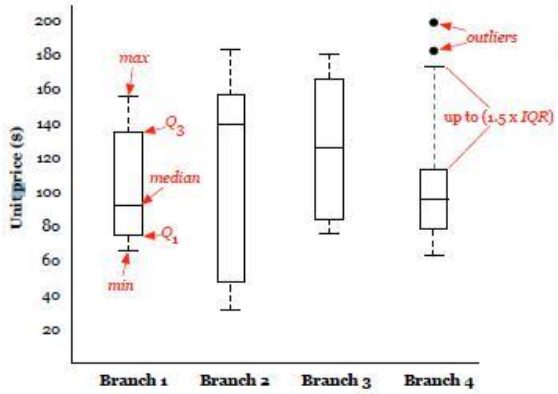
3 Loess curve

It is another important exploratory graphic aid that adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence. The word loess is short for "local regression."



4 Box plot

The picture produced consists of the most extreme values in the data set (maximum and minimum values), the lower and upper quartiles, and the median.

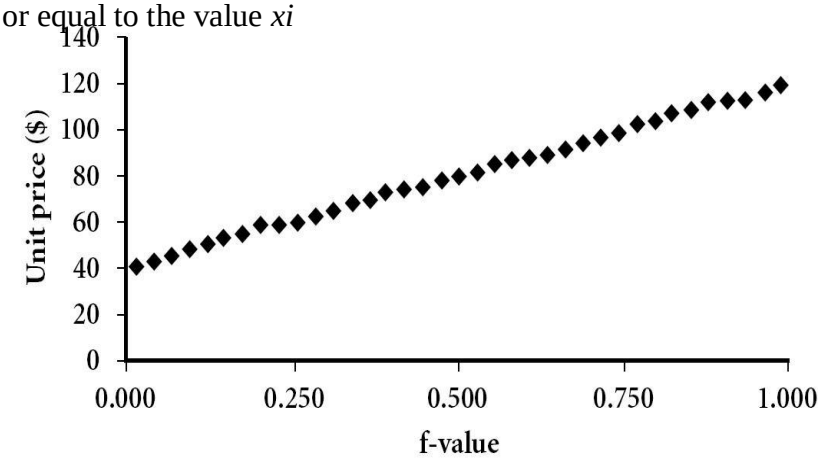


5 Quantile plot

Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)

Plots quantile information

For a data *xi* data sorted in increasing order, *fi* indicates that approximately 100 *fi*% of the data are below or equal to the value *xi*



The f quantile is the data value below which approximately a decimal fraction f of the data is found. That data value is denoted q(f). Each data point can be assigned an f-value. Let a time series x of length n be sorted from smallest to largest values, such that the sorted values have rank. The f-value for each observation is computed as . 1,2,..., n . The f-value for each observation is computed as,

$$f_i = \frac{i - 0.5}{n}$$

6 Quantile-Quantile plots (Q-Q plot)

Quantile-quantile plots allow us to compare the quantiles of two sets of numbers.

This kind of comparison is much more detailed than a simple comparison of means or medians.

A normal distribution is often a reasonable model for the data. Without inspecting the data, however, it is risky to assume a normal distribution. There are a number of graphs that can be used to check the deviations of the data from the normal distribution. The most useful tool for assessing normality is a quantile or QQ plot. This is a scatter plot with the quantiles of the scores on the horizontal axis and the expected normal scores on the vertical axis.

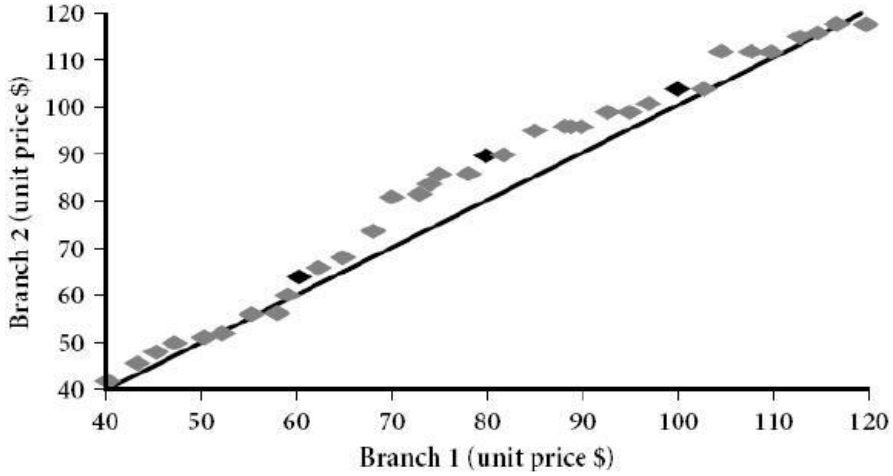
In other words, it is a graph that shows the quantiles of one univariate distribution against the corresponding quantiles of another. It is a powerful visualization tool in that it allows the user to view whether there is a shift in going from one distribution to another.

The steps in constructing a QQ plot are as follows:

First, we sort the data from smallest to largest. A plot of these scores against the expected normal scores should reveal a straight line.

The expected normal scores are calculated by taking the z-scores of (I - ½)/n where I is the rank in increasing order.

Curvature of the points indicates departures of normality. This plot is also useful for detecting outliers. The outliers appear as points that are far away from the overall pattern op points



How is a quantile-quantile plot different from a quantile plot?

A quantile plot is a graphical method used to show the approximate percentage of values below or equal to the independent variable in a univariate distribution. Thus, it displays quantile information for all the data, where the values measured for the independent variable are plotted against their corresponding quantile.

A quantile-quantile plot however, graphs the quantiles of one univariate distribution against the corresponding quantiles of another univariate distribution. Both axes display the range of values measured for their corresponding distribution, and points are plotted that correspond to the quantile values of the two distributions. A line (y = x) can be added to the graph along with points representing where the first, second and third quantiles lie, in order to increase the graph's informational value. Points that lie above such a line indicate a correspondingly higher value for the distribution plotted on the y-axis, than for the distribution plotted on the x-axis at the same quantile. The opposite effect is true for points lying below this line.

Data Cleaning

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.
Various methods for handling this problem:

Missing Values

The various methods for handling the problem of missing values in data tuples include:
Ignoring the tuple: This is usually done when the class label is missing (assuming the mining task involves classification or description). 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.
Manually filling in the missing value: In general, 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.
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.

Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values, for all samples belonging to the same class as the given tuple: For example, if classifying customers according to credit risk, replace the missing value with

the average income value for customers in the same credit risk category as that of the given tuple.

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. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

Noisy data:

Noise is a random error or variance in a measured variable. Data smoothing tech is used for removing such noisy data.

Several Data smoothing techniques:

1 Binning methods: Binning methods smooth a sorted data value by consulting the neighborhood", or values around it. The sorted values are distributed into a number of 'buckets', or bins. Because binning methods consult the neighborhood of values, they perform local smoothing.

In this technique,

- The data for first sorted
- Then the sorted list partitioned into equi-depth of bins.
- Then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Smoothing by bin means: Each value in the bin is replaced by the mean value of the bin.

Smoothing by bin medians: Each value in the bin is replaced by the bin median.

Smoothing by boundaries: The min and max values of a bin are identified as the bin boundaries. Each bin value is replaced by the closest boundary value.

Suppose that the data for analysis include the attribute age. The age values for the data tuples are (in increasing order): 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps.

Comment on the effect of this technique for the given data.

The following steps are required to smooth the above data using smoothing by bin means with a bin depth of 3.

Step 1: Sort the data. (This step is not required here as the data are already sorted.)

Step 2: Partition the data into equi-depth bins of depth

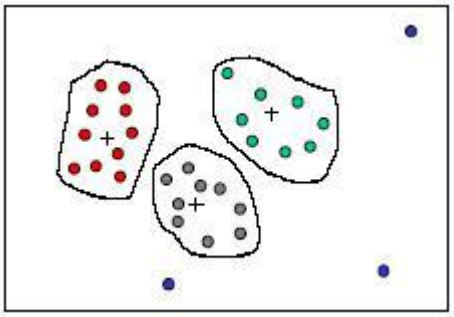
Bin 1: 13, 15, 16 Bin 2: 16, 19, 20 Bin 3: 20, 21, 22 Bin 4: 22, 25, 25 Bin 5: 25, 25, 30
Bin 6: 33, 33, 35 Bin 7: 35, 35, 35 Bin 8: 36, 40, 45 Bin 9: 46, 52, 70

Step 3: Calculate the arithmetic mean of each bin.

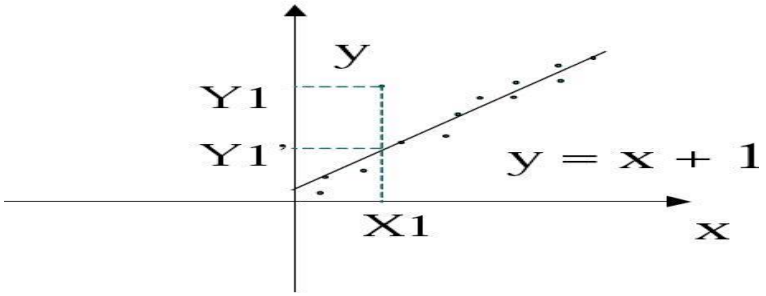
Step 4: Replace each of the values in each bin by the arithmetic mean calculated for the bin.

Bin 1: 14, 14, 14 Bin 2: 18, 18, 18 Bin 3: 21, 21, 21
Bin 4: 24, 24, 24 Bin 5: 26, 26, 26 Bin 6: 33, 33, 33
Bin 7: 35, 35, 35 Bin 8: 40, 40, 40 Bin 9: 56, 56, 56

2 Clustering: Outliers in the data may be detected by clustering, where similar values are organized into groups, or 'clusters'. Values that fall outside of the set of clusters may be considered outliers.



3 Regression : smooth by fitting the data into regression functions.
Linear regression involves finding the best of line to fit two variables, so that one variable can be used to predict the other.



- Multiple linear regression is an extension of linear regression, where more than two variables are involved and the data are fit to a multidimensional surface.
- Using regression to find a mathematical equation to fit the data helps smooth out the noise.
- Field overloading: is a kind of source of errors that typically occurs when developers compress new attribute definitions into unused portions of already defined attributes.
- Unique rule is a rule says that each value of the given attribute must be different from all other values of that attribute
- Consecutive rule is a rule says that there can be no missing values between the lowest and highest values of the attribute and that all values must also be unique.
- Null rule specifies the use of blanks, question marks, special characters or other strings that may indicate the null condition and how such values should be handled.

Data Integration and Transformation

Data Integration

It combines data from multiple sources into a coherent store. There are number of issues to consider during data integration.

Issues:

Schema integration: refers integration of metadata from different sources.

Entity identification problem: Identifying entity in one data source similar to entity in another table. For example, customer_id in one db and customer_no in another db refer to the same entity

Detecting and resolving data value conflicts: Attribute values from different sources can be different due to different representations, different scales. E.g. metric vs. British units

Redundancy: is another issue while performing data integration. Redundancy can occur due to the following reasons:

- Object identification: The same attribute may have different names in different db
- Derived Data: one attribute may be derived from another attribute.

Data Transformation

Data transformation can involve the following:
Smoothing: which works to remove noise from the data
Aggregation: where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute weekly and annual total scores.
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.
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.
□□□□□□Attribute construction (feature construction): this is where new attributes are constructed and added from the given set of attributes to help the mining process.
Normalization
In which data are scaled to fall within a small, specified range, useful for classification algorithms involving neural networks, distance measurements such as nearest neighbor classification and clustering. There are 3 methods for data normalization. They are:
min-max normalization □
z-score normalization
normalization by decimal scaling
Min-max normalization: performs linear transformation on the original data values. It can be defined as,

$$v' \square\square_{maxA \square minA} (new\_maxA \square new\_minA) \square new\_minA$$

v is the value to be normalized
minA,maxA are minimum and maximum values of an attribute A new_ maxA, new_ minA are the normalization range.

Z-score normalization / zero-mean normalization: In which values of an attribute A are normalized based on the mean and standard deviation of A. It can be defined as,

$$v' \square \frac{v \square meanA}{stand\_devA}$$

This method is useful when min and max value of attribute A are unknown or when outliers that are dominate min-max normalization.
Normalization by decimal scaling: normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value v of A is normalized to v' by computing,

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \text{Max}(|v'|)<1$$

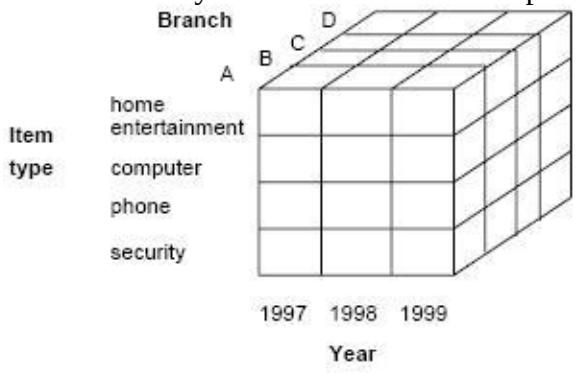
Data Reduction techniques
These techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. Data reduction includes,
Data cube aggregation, where aggregation operations are applied to the data in the construction of a data cube.
Attribute subset selection, where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.

Dimensionality reduction, where encoding mechanisms are used to reduce the data set size. Examples: Wavelet Transforms Principal Components Analysis

Numerosity reduction, where 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.

Discretization and concept hierarchy generation, where raw data values for attributes are replaced by ranges or higher conceptual levels. Data Discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies.

Data cube aggregation: Reduce the data to the concept level needed in the analysis. Queries regarding aggregated information should be answered using data cube when possible. Data cubes store multidimensional aggregated information. The following figure shows a data cube for multidimensional analysis of sales data with respect to annual sales per item type for each branch.

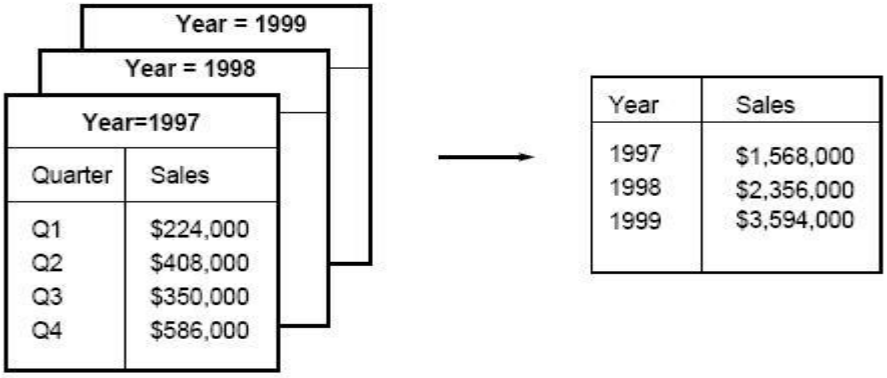


Each cells holds an aggregate data value, corresponding to the data point in multidimensional space.

Data cubes provide fast access to pre computed, summarized data, thereby benefiting on-line analytical processing as well as data mining.

The cube created at the lowest level of abstraction is referred to as the base cuboid. A cube for the highest level of abstraction is the apex cuboid. The lowest level of a data cube (base cuboid). Data cubes created for varying levels of abstraction are sometimes referred to as cuboids, so that a “data cube" may instead refer to a lattice of cuboids. Each higher level of abstraction further reduces the resulting data size.

The following database consists of sales per quarter for the years 1997-1999.



Suppose, the analyzer interested in the annual sales rather than sales per quarter, the above data can be aggregated so that the resulting data summarizes the total sales per year instead of

per quarter. The resulting data in smaller in volume, without loss of information necessary for the analysis task.

Dimensionality Reduction

It reduces the data set size by removing irrelevant attributes. This is a method of attribute subset selection are applied. A heuristic method of attribute of sub set selection is explained here:

Attribute sub selection / Feature selection

Feature selection is a must for any data mining product. That is because, when you build a data mining model, the dataset frequently contains more information than is needed to build the model. For example, a dataset may contain 500 columns that describe characteristics of customers, but perhaps only 50 of those columns are used to build a particular model. If you keep the unneeded columns while building the model, more CPU and memory are required during the training process, and more storage space is required for the completed model.

In which select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features

Basic heuristic methods of attribute subset selection include the following techniques, some of which are illustrated below:

- **Step-wise forward selection:** The procedure starts with an empty set of attributes. The best of the original attributes is determined and added to the set. At each subsequent iteration or step, the best of the remaining original attributes is added to the set.
- **Step-wise backward elimination:** The procedure starts with the full set of attributes. At each step, it removes the worst attribute remaining in the set.
- **Combination forward selection and backward elimination**: The step-wise forward selection and backward elimination methods can be combined, where at each step one selects the best attribute and removes the worst from among the remaining attributes.
- **Decision tree induction:** Decision tree induction constructs a flow-chart-like structure where each internal (non-leaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external (leaf) node denotes a class prediction. At each node, the algorithm chooses the “best" attribute to partition the data into individual classes. When decision tree induction is used for attribute subset selection, a tree is constructed from the given data. All attributes that do not appear in the tree are assumed to be irrelevant. The set of attributes appearing in the tree form the reduced subset of attributes.

Forward Selection

Initial attribute set:
{A1, A2, A3, A4, A5, A6}

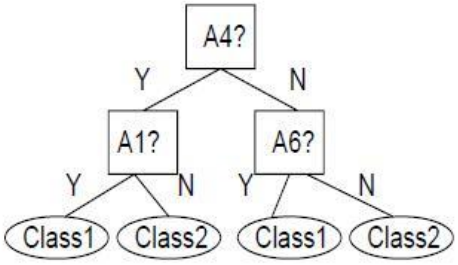
Initial reduced set:
{}
-> {A1}
--> {A1, A4}
---> Reduced attribute set:
{A1, A4, A6}

Backward Elimination

Initial attribute set:
{A1, A2, A3, A4, A5, A6}
-> {A1, A3, A4, A5, A6}
--> {A1, A4, A5, A6}
---> Reduced attribute set:
{A1, A4, A6}

Decision Tree Induction

Initial attribute set:
{A1, A2, A3, A4, A5, A6}



---> Reduced attribute set:
{A1, A4, A6}

Greedy (heuristic) methods for attribute subset selection.

Discretization and concept hierarchies

Discretization:

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. Interval labels can then be used to replace actual data values.

Concept Hierarchy

A concept hierarchy for a given numeric attribute defines a Discretization of the attribute. Concept hierarchies can be used to reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).

Discretization and Concept hierarchy for numerical data:

Three types of attributes:
Nominal — values from an unordered set, e.g., color, profession
Ordinal — values from an ordered set, e.g., military or academic rank
Continuous — real numbers, e.g., integer or real numbers
There are five methods for numeric concept hierarchy generation. These include:
binning,
histogram analysis,
clustering analysis,
entropy-based Discretization, and
data segmentation by "natural partitioning".

An information-based measure called "entropy" can be used to recursively partition the values of a numeric attribute A, resulting in a hierarchical Discretization.

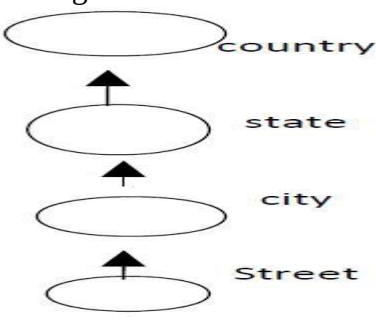
Procedure:

- Given a set of samples S, if S is partitioned into two intervals S_1 and S_2 using boundary T, the information gain after partitioning is
$$I(S,T) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2)$$
- Entropy is calculated based on class distribution of the samples in the set. Given m classes, the entropy of S_1 is
$$Entropy(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$
where p_i is the probability of class i in S_1
- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization
- The process is recursively applied to partitions obtained until some stopping criterion is met
- Such a boundary may reduce data size and improve classification accuracy

Concept hierarchy generation for category data

A concept hierarchy defines a sequence of mappings from set of low-level concepts to higher-level, more general concepts.

It organizes the values of attributes or dimension into gradual levels of abstraction. They are useful in mining at multiple levels of abstraction



Possible Questions:

2-Mark Questions:

1. What is data preprocessing?
2. What is data cleaning?
3. What is binning?
4. Give the benefits of online storage.
5. What are the four types of methods to handle noisy data?
6. Why is data pre-processing important?
7. Define data transformation
8. Define data reduction
9. What do mean by Data discretization?
10. What are types of data in concept hierarchy?

Part-B:

1. What do you mean by 'Data Processing'? Why processing is required? Describe the various forms of data Processing.
2. Describe how concept hierarchies are useful in data mining.
3. Discuss issues to consider during data integration.
4. Explain in detail about the Data cleaning in Data preprocessing.
5. Elucidate the needs and steps involved in Data Pre-processing?
6. Describe how the concept hierarchies are useful in data mining.
7. What are the various issues addressed during data integration.
8. Write in detail about:
    i) Data Integration
    ii) Data Transformation

Part-C: Compulsory Questions (10 marks)

1. How can the data be preprocessed so as to improve the efficiency and ease of mining process? Discuss.
2. Explain in detail about Data Reduction.
3. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the mean of the data? What is the median?

(b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

(c) What is the midrange of the data?

(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

UNIT III

SYLLABUS:

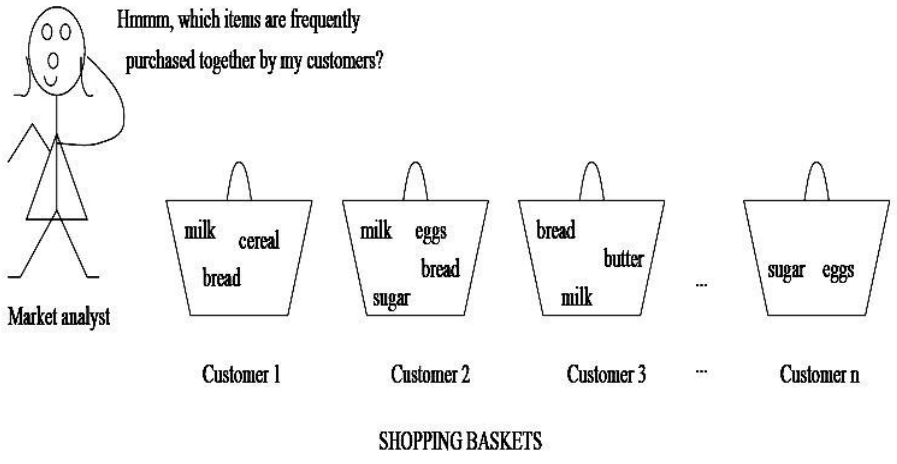
Mining Frequent Patterns Associations and Correlations: Basic Concepts - Efficient and Scalable Frequent item set Mining Methods - Mining various kinds of Association rules – From Association Mining to Correlation Analysis - Constraint-Based Association Mining.

Basic Concepts and a Road Map

Market – Basket analysis

A market basket is a collection of items purchased by a customer in a single transaction, which is a well-defined business activity. For example, a customer's visits to a grocery store or an online purchase from a virtual store on the Web are typical customer transactions. Retailers accumulate huge collections of transactions by recording business activities over time. One common analysis run against a transactions database is to find sets of items, or *itemsets,* that appear together in many transactions. A business can use knowledge of these patterns to improve the Placement of these items in the store or the layout of mail- order catalog page and Web pages. An itemset containing *i* items is called an *i-itemset*. The percentage of transactions that contain an itemset is called the itemsets *support*. For an itemset to be interesting, its support must be higher than a user-specified minimum. Such itemsets are said to be frequent.

Figure : Market basket analysis.



$$computer \Rightarrow financial\_management\_software \quad [support = 2\%, confidence = 60\%]$$

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for association Rule means that 2% of all the transactions under analysis show that computer and financial management software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold.

Frequent Itemsets, Closed Itemsets, and Association Rules

Association rule mining:

– Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.

Applications:

– Basket data analysis, cross-marketing, catalog design, loss-leader analysis, clustering, classification, etc.

Examples.

– Rule form: "Body ® Head [support, confidence]".
– buys(x, "diapers") ® buys(x, "beers") [0.5%, 60%]
– major(x, "CS") ^ takes(x, "DB") ® grade(x, "A") [1%, 75%]

Association Rule: Basic Concepts

Given: (1) database of transactions, (2) each transaction is a list of items (purchased by a customer in a visit)

Find: <u>all</u> rules that correlate the presence of one set of items with that of another set of items

– E.g., *98% of people who purchase tires and auto accessories also get automotive services done*

Applications

– * □ *Maintenance Agreement* (What the store should do to boost Maintenance Agreement sales)
– *Home Electronics* □ * (What other products should the store stocks up?)
– Attached mailing in direct marketing

– Detecting “ping-pong”ing of patients, faulty “collisions”

Rule Measures: Support and Confidence

Find all the rules *X & Y* *Z* with minimum confidence and support

– support, *s*, probability that a transaction contains {X Y Z}

– confidence, *c*, conditional probability that a transaction having {X Y} also contains *Z*

Efficient and Scalable Frequent Itemset Mining Methods

Mining Frequent Patterns

The method that mines the complete set of frequent itemsets with candidate generation.

Apriori property & The Apriori Algorithm.

Apriori property

All nonempty subsets of a frequent item set most also be frequent.

– An item set I does not satisfy the minimum support threshold, min-sup, then I is not frequent, i.e., support(I) < min-sup

– If an item A is added to the item set I then the resulting item set (I U A) can not occur more frequently than I.

Monotonic functions are functions that move in only one direction.

This property is called anti-monotonic.

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

This property is monotonic in failing the test.

The Apriori Algorithm

Join Step: C_k is generated by joining $L_{k-1}$with itself

Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

Input: Database, D, of transactions; minimum support threshold, *min_sup*.

Output: L, frequent itemsets in D.

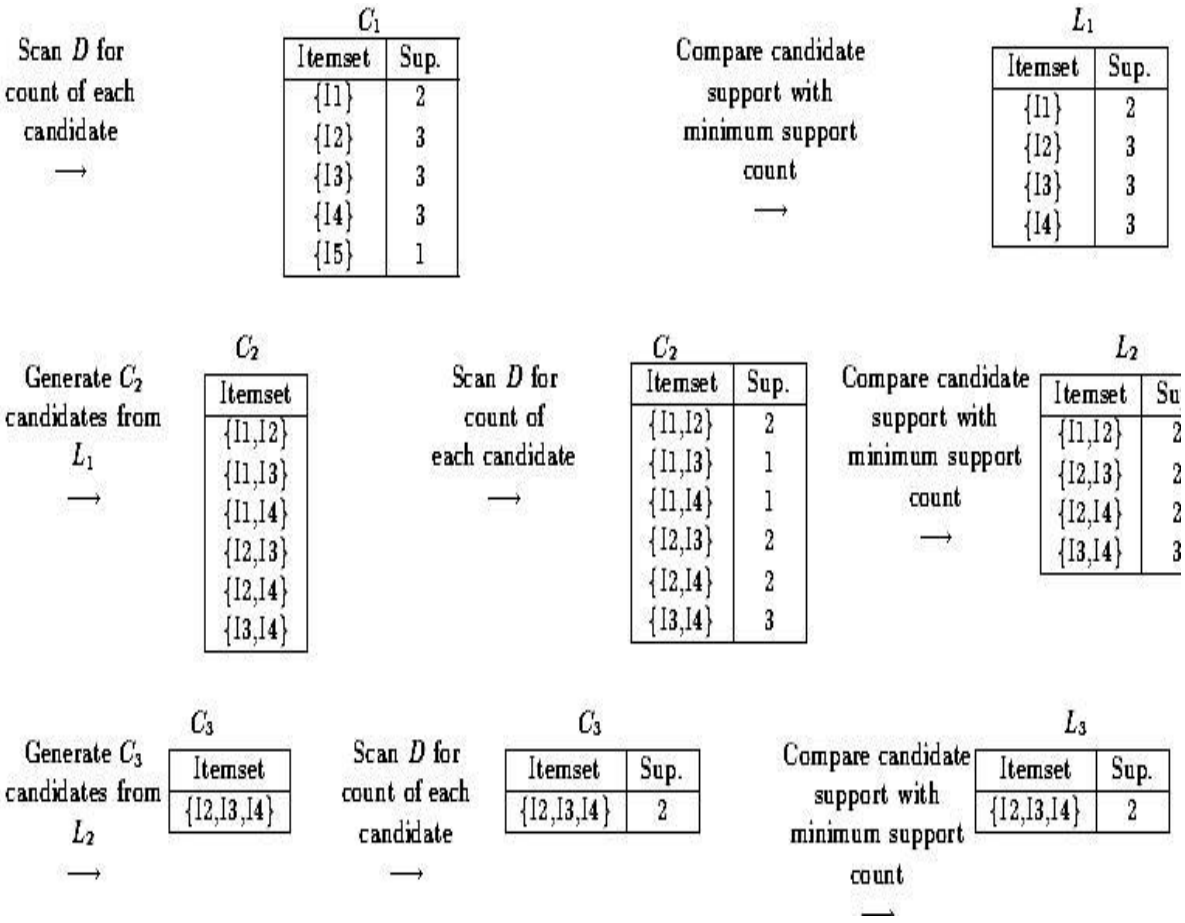
Method:

```
1)  L1 = find_frequent_1-itemsets(D);
2)  for (k = 2; Lk−1 ≠ ∅; k++) {
3)      Ck = apriori_gen(Lk−1, min_sup);
4)      for each transaction t ∈ D { // scan D for counts
5)          Ct = subset(Ck, t); // get the subsets of t that are candidates
6)          for each candidate c ∈ Ct
7)              c.count++;
8)      }
9)      Lk = {c ∈ Ck | c.count ≥ min_sup}
10) }
11) return L = ∪k Lk;
```

```
procedure apriori_gen(Lk−1:frequent (k-1) itemsets; min_sup: minimum support)
1)  for each itemset l1 ∈ Lk−1
2)      for each itemset l2 ∈ Lk−1
3)          if (l1[1] = l2[1]) ∧ (l1[2] = l2[2]) ∧ ... ∧ (l1[k − 2] = l2[k − 2]) ∧ (l1[k − 1] < l2[k − 1]) then {
4)              c = l1 ⋈ l2; // join step: generate candidates
5)              if has_infrequent_subset(c, Lk−1) then
6)                  delete c; // prune step: remove unfruitful candidate
7)              else add c to Ck;
8)          }
9)  return Ck;
```

```
procedure has_infrequent_subset(c: candidate k itemset; Lk−1: frequent (k − 1) itemsets); // use prior knowledge
1)  for each (k − 1) subset s of c
2)      if s ∉ Lk−1 then
3)          return TRUE;
4)  return FALSE;
```

Example



The method that mines the complete set of frequent itemsets without generation.

Compress a large database into a compact, Frequent-Pattern tree (FP-tree) structure

– highly condensed, but complete for frequent pattern mining

– avoid costly database scans

Develop an efficient, FP-tree-based frequent pattern mining method

– A divide-and-conquer methodology: decompose mining tasks into smaller ones

– Avoid candidate generation: sub-database test only!

Construct FP-tree from a Transaction DB

TID	*Items bought (ordered) frequent items*	
100	**{*f, a, c, d, g, i, m, p*}**	**{*f, c, a, m, p*}**
200	**{*a, b, c, f, l, m, o*}**	**{*f, c, a, b, m*}**
min_support = 0.5		
300	**{*b, f, h, j, o*}**	**{*f, b*}**
400	**{*b, c, k, s, p*}**	**{*c, b, p*}**
500	**{*a, f, c, e, l, p, m, n*}**	**{*f, c, a, m, p*}**

Steps:

Scan DB once, find frequent 1-itemset (single item pattern)

Order frequent items in frequency descending order

Scan DB again, construct FP-tree

Header Table

Item frequency head

f	*4*
c	*4*

3

3

3

p3

Benefits of the FP-tree Structure

Completeness:

- never breaks a long pattern of any transaction
- preserves complete information for frequent pattern mining

Compactness

- reduce irrelevant information—infrequent items are gone

- frequency descending ordering: more frequent items are more likely to be shared
- never be larger than the original database (if not count node-links and counts)
- Example: For Connect-4 DB, compression ratio could be over 100

Mining Frequent Patterns Using FP-tree

General idea (divide-and-conquer)

- Recursively grow frequent pattern path using the FP-tree Method
- For each item, construct its conditional pattern-base, and then its conditional FP-tree
- Repeat the process on each newly created conditional FP-tree
- Until the resulting FP-tree is empty, or it contains only one path (single path will generate all the combinations of its sub-paths, each of which is a frequent pattern)

Major Steps to Mine FP-tree

Construct conditional pattern base for each node in the FP-tree

Construct conditional FP-tree from each conditional pattern-base

Recursively mine conditional FP-trees and grow frequent patterns obtained so far

If the conditional FP-tree contains a single path, simply enumerate all the patterns

Why Is Frequent Pattern Growth Fast?

Our performance study shows

– FP-growth is an order of magnitude faster than Apriori, and is also faster than tree-projection

Reasoning

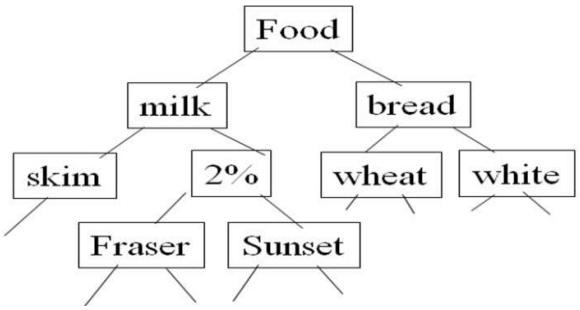
- No candidate generation, no candidate test
- Use compact data structure
- Eliminate repeated database scan

Basic operation is counting and FP-tree building

Mining Various Kinds of Association Rules

Mining multilevel association rules from transactional databases

Multilevel association rule: Multilevel association rules can be defined as applying association rules over different levels of data abstraction



Steps to perform multilevel association rules from transactional database are:

Step1:consider frequent item sets

Step2: arrange the items in hierarchy form

Step3: find the Items at the lower level (expected to have lower support)

Step4: Apply association rules on frequent item sets

Step5: Use some methods and identify frequent itemsets

Note: Support is categorized into two types

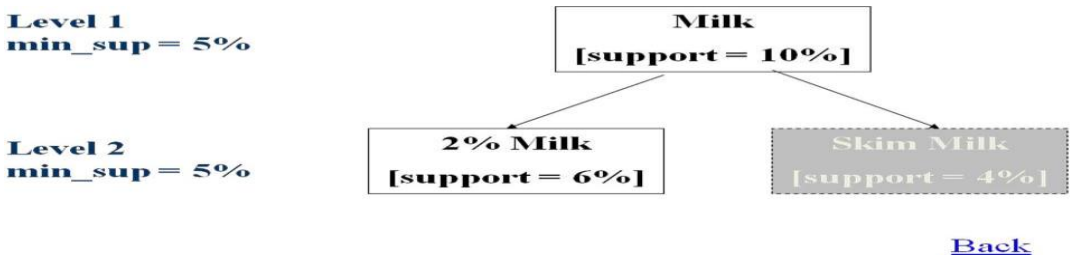
Uniform Support: the same minimum support for all levels

Reduced Support: reduced minimum support at lower levels

*multilevel association rules can be applied on both the supports

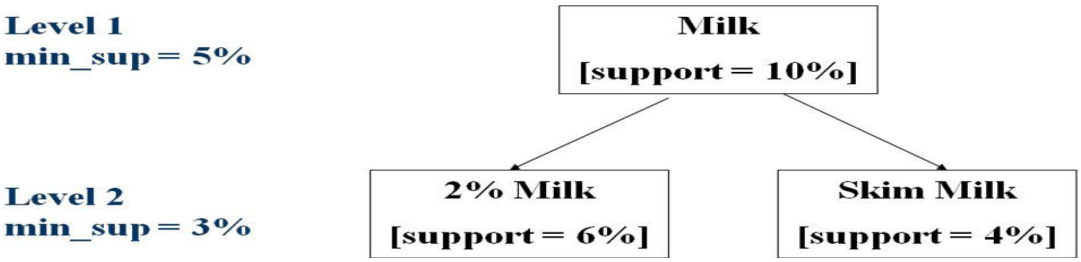
Example figure for Uniform Support:

Multi-level mining with uniform support



Reduced Support:

*Multi-level mining with reduced support



Mining multidimensional association rules from Relational databases and data warehouse

Multi dimensional association rule: Multi dimensional association rule can be defined as the statement which contains only two (or) more predicates/dimensions

Multi dimensional association rule also called as Inter dimensional association rule

We can perform the following association rules on relational database and data warehouse

1)Boolean dimensional association rule: Boolean dimensional association rule can be defined as comparing existing predicates/dimensions with non existing predicates/dimensions..

Single dimensional association rule: Single dimensional association rule can be defined as the statement which contains only single predicate/dimension

Single dimensional association rule also called as Intra dimensional association rule as usually multi dimensional association rule..

Multi dimensional association rule can be applied on different types of attributes ..here the attributes are

Categorical Attributes

– finite number of possible values, no ordering among values

Quantitative Attributes

– numeric, implicit ordering among values

Note1:Relational database can be viewed in the form of tables…so on tables we are performing the concept hierarchy

In relational database we are using the concept hierarchy that means generalization in order to find out the frequent item sets

Generalization: replacing low level attributes with high level attributes called generalization

Note2: data warehouse can be viewed in the form of multidimensional data model (uses data cubes) in order to find out the frequent patterns.

From association mining to correlation analysis:

*Here Association mining to correlation analysis can be performed Based on interesting measures of frequent items.

*among frequent items we are performing correlation analysis

*correlation analysis means one frequent item is dependent on other frequent item..

Consider Two popular measurements:

support; and

confidence

for each frequent item we are considering the above mentioned two measures to perform mining and correlation analysis..

Note: **Association mining:** Association mining can be defined as finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories

As we have seen above, the support and confidence measures are insufficient at filtering out uninteresting association rules. To tackle this weakness, a correlation measure can be used to augment the support-confidence framework for association rules. This leads to correlation rules of the form $A \Rightarrow B$ [support, confidence. correlation].

That is, a correlation rule is measured not only by its support and confidence but also by the correlation between itemsets A and B. There are many different correlation measures from which to choose. In this section, we study various correlation measures to determine which would be good for mining large data sets. Lift is a simple correlation measure that is given as follows. The occurrence of itemset A is independent of the occurrence of itemset B if

$P(A \cup B) = P(A)P(B)$;

otherwise, itemsets A and B are dependent and correlated as events.

This definition can easily be extended to more than two itemsets. The lift between the occurrence of A and B can be measured by computing

$lift(A, B) = P(A \cup B)\ P(A)P(B)$.

If the resulting value of Equation (5.23) is less than 1, then the occurrence of A is negatively correlated with the occurrence of B. If the resulting value is greater than 1, then A and B are positively correlated, meaning that the occurrence of one implies the occurrence of the other. If the resulting value is equal to 1, then A and B are independent and there is no correlation between them. Equation is equivalent to

$P(B|A)/P(B)$, or con $f(A \Rightarrow B)/sup(B)$,

which is also referred as the lift of the association (or correlation) rule $A \Rightarrow B$.

In other words, it assesses the degree to which the occurrence of one "lifts" the occurrence of the other. For example, if A corresponds to the sale of computer games and B corresponds to the sale of videos, then given the current market conditions, the sale of games is said to increase or "lift" the likelihood of the sale of videos by a factor of the value returned.

The second correlation measure is the χ 2 measure. To compute the χ 2 value, we take the squared difference between the observed and expected value for a slot (A and B pair) in the contingency table, divided by the expected value. This amount is summed for all slots of the contingency table

Constraint-based association mining:

Constraint based association mining can be defined as applying different types of constraints on different types of knowledge so the kinds of constraints used in the mining are

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

Knowledge type constraint: Based on classification, association, we are applying the knowledge type constraints.

Data constraints: SQL-like queries

Ex: Find product pairs sold together in Vancouver in Dec.'98.

Dimension/level constraints: in relevance to region, price, brand, customer category.

Rule constraints: On the form of the rules to be mined (e.g., # of predicates, etc) small sales (price < $10) triggers big sales (sum > $200).

Interestingness constraints:

1.Thresholds on measures of interestingness

2. strong rules (min_support □ 3%, min_confidence □ 60%).

Constraint Pushing: Mining Guided by Rule Constraints

- Rule constraints specify expected set/subsetν relationships of the variables in the mined rules, constant initiation of variables, and aggregate functions
- Rule constraints can be classified into the following five categories with respect to frequent itemset mining:

(1) antimonotonic

(2) monotonic

(3) succinct

(4) convertible

(5)inconvertible

The first category of constraints is **antimonotonic**. Consider the rule constraint "sum(I.price) ≤ 100" . Suppose we are using the Apriori framework, which at each iteration k explores itemsets of size k. If the price summation of the items is no less than 100, this itemset can be pruned from the search space, since adding more items into the set will only make it more expensive and thus will never satisfy the constraint. In other words, if an itemset does not satisfy this rule constraint, none of its supersets can satisfy the constraint. If a rule constraint obeys this property, it is antimonotonic.

The second category of constraints is **monotonic**. If the rule constraint in "sum(I.price) ≥ 100," the constraint-based processing method would be quite different. If an itemset I satisfies the constraint, that is, the sum of the prices in the set is no less than 100, further addition of more items to I will increase cost and will always satisfy the constraint. Therefore, further testing of this constraint on itemset I becomes redundant. In other words, if an itemset satisfies this rule constraint, so do all of its supersets. If a rule constraint obeys this property, it is monotonic. Similar rule monotonic constraints include "min(I.price) ≤ 10," "count(I) ≥ 10," and so on.

The third category is **succinct constraints**. For this category of constraints, we can enumerate all and only those sets that are guaranteed to satisfy the constraint. That is, if a rule constraint is succinct, we can directly generate precisely the sets that satisfy it, even before support counting begins. This avoids the substantial overhead of the generate-and-test paradigm. In other words, such constraints are precounting prunable. For example, the constraint "min(J.price) ≥ 500" is succinct, because we can explicitly and precisely generate all the sets of items satisfying the constraint. Specifically, such a set must contain at least one item whose price is no less than $500. It is of the form S1 ∪S2, where S1 6= 0/ is a subset of the set of all those items with prices no less than $500, and S2, possibly empty, is a subset of the set of all those items with prices no greater than $500. Because there is a precise "formula" for generating all of the sets satisfying a succinct constraint, there is no need to iteratively check the rule constraint during the mining process.

The fourth category is **convertible constraints.** Some constraints belong to none of the above three categories. However, if the items in the itemset are arranged in a particular order, the constraint may become monotonic or antimonotonic with regard to the frequent itemset

mining process. For example, the constraint "avg(I.price) ≤ 100" is neither antimonotonic nor monotonic. However, if items in a transaction are added to an itemset in price-ascending order, the constraint becomes antimonotonic, because if an itemset I violates the constraint (i.e., with an average price greater than $100), then further addition of more expensive items into the itemset will never make it satisfy the constraint.

Fifth category of constraints is called **inconvertible constraints**. The good news is that although there still exist some tough constraints that are not convertible, most simple SQL expressions with built-in SQL aggregates belong to one of the first four categories to which efficient constraint mining methods can be applied.

Possible Questions:

2-Mark Questions:

1. What is support?
2. Define confidence
3. What is Monotonic rule
4. Give the formula for chi square to calculate correlation
5. What do mean by lift?
6. What is advantage of FP growth over Apriori algorithm?
7. Which constraints are convertible?
8. Differentiate multi level and multiple association rule mining?

Part-B:

1. Explain how the efficiency of apriori is improved?
2. Explain constraint-based association mining?
3. Write and explain the algorithm for mining frequent item sets. Give relevant example.
4. Explain frequent item set without candidate generation?
5. Write and explain the algorithm for mining frequent item sets with example.
6. Explain constraint-based association mining?
7. What is meant by 'Association Rule Mining'? Explain by presenting an example of Market basket Analysis?
8. Explain Apriori algorithm?

Part-C: Compulsory Questions (10 marks)

1. A data base has five transaction. Let minimum support=60% and minimum confidence=75%.

TID	Items bought
T_{100}	{B, C, E, J}
T_{200}	{B, C, J}
T_{300}	{B, M, Y}
T_{400}	{B, J, M}
T_{500}	{C, J, M}

Find all frequent item sets using Apriori and FP-growth.

2. Discuss mining of multilevel association rules from transactional databases

UNIT-IV

Syllabus:

Classification and Prediction: Issues Regarding Classification and Prediction - Classification by Decision Tree Induction - Rule-based Classification – Prediction - Accuracy and Error Measures - Evaluating the Accuracy of a classifier or Predictor - Ensemble Methods - increases the Accuracy - Model Selection.

Classification and Prediction

Classification:

- used for prediction(future analysis) to know the unknown attributes with their values by using classifier algorithms and decision tree.(in data mining)
- Which constructs some models(like decision trees) then which classifies the attributes..
- already we know the types of attributes are
  1.categorical attribute and 2.numerical attribute
- These classification can work on both the above mentioned attributes.

Prediction: prediction also used for to know the unknown or missing values, which also uses some models in order to predict the attributes models like neural networks, if else rules and other mechanisms Classification and prediction are used in the Applications like

*credit approval

*target marketing

*medical diagnosis

Classification—A Two-Step Process

Model construction: describing a set of predetermined classes

– Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute

– The set of tuples used for model construction: training set

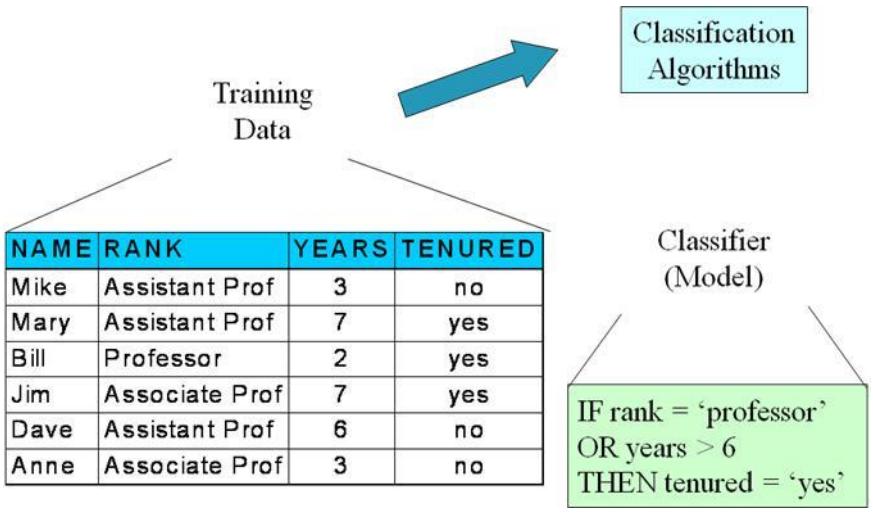
– The model is represented as classification rules, decision trees, or mathematical formulae

Model usage: for classifying future or unknown objects

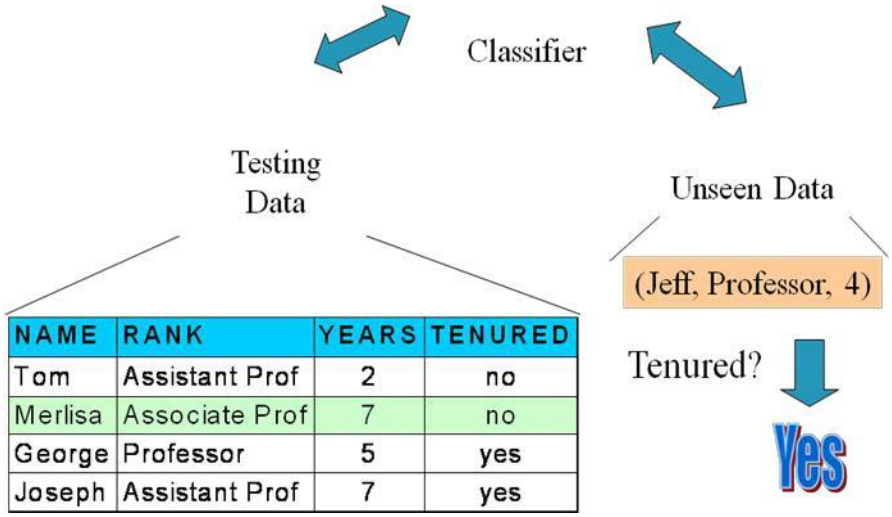
– Estimate accuracy of the model

- The known label of test sample is compared with the classified result from the model
- Accuracy rate is the percentage of test set samples that are correctly classified by the model
- Test set is independent of training set, otherwise over-fitting will occur

Process (1): Model Construction



Process (2): Using the Model in Prediction



Supervised vs. Unsupervised Learning

Supervised learning (classification)

Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations

New data is classified based on the training set

Unsupervised learning (clustering)

The class labels of training data is unknown

Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Issues regarding classification and prediction:

There are two issues regarding classification and prediction they are

Issues (1): Data Preparation

Issues (2): Evaluating Classification Methods

Issues (1): Data Preparation: Issues of data preparation includes the following 1)

1) Data cleaning

*Preprocess data in order to reduce noise and handle missing values (refer preprocessing techniques i.e. data cleaning notes)

2) Relevance analysis (feature selection)

Remove the irrelevant or redundant attributes (refer unit-iv AOI Relevance analysis)

Data transformation (refer preprocessing techniques i.e data cleaning notes) Generalize and/or normalize data

Issues (2): Evaluating Classification Methods: considering classification methods should satisfy the following properties

Predictive accuracy: time to construct the model

Speed and scalability: time to use the model

3. Robustness: handling noise and missing values

4. Scalability: efficiency in disk-resident databases

5. Interpretability: understanding and insight provided by the model

Goodness of rules :

- decision tree size
- compactness of classification rules

Classification by Decision Tree Induction

Decision tree

- A flow-chart-like tree structure
- Internal node denotes a test on an attribute
- Branch represents an outcome of the test
- Leaf nodes represent class labels or class distribution

Decision tree generation consists of two phases

– Tree construction

- At start, all the training examples are at the root

- Partition examples recursively based on selected attributes

– Tree pruning

- Identify and remove branches that reflect noise or outliers
- Use of decision tree: Classifying an unknown sample

– Test the attribute values of the sample against the decision tree

Training Dataset

This follows an example from Quinlan's ID3

age	income	student	credit_rating
<=30	high	no	fair
<=30	high	no	excellent
31…40	high	no	fair
>40	medium	no	fair
>40	low	yes	fair
>40	low	yes	excellent
31…40	low	yes	excellent
<=30	medium	no	fair
<=30	low	yes	fair
>40	medium	yes	fair
<=30	medium	yes	excellent
31…40	medium	no	excellent
31…40	high	yes	fair
>40	medium	no	excellent

Algorithm for Decision Tree Induction

Basic algorithm (a greedy algorithm)

– Tree is constructed in a top-down recursive divide-and-conquer manner

– At start, all the training examples are at the root

– Attributes are categorical (if continuous-valued, they are discretized in advance)

– Examples are partitioned recursively based on selected attributes

– Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)

Conditions for stopping partitioning

– All samples for a given node belong to the same class

– There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf

– There are no samples left

Extracting Classification Rules from Trees

- Represent the knowledge in the form of IF-THEN rules
- One rule is created for each path from the root to a leaf
- Each attribute-value pair along a path forms a conjunction
- The leaf node holds the class prediction
- Rules are easier for humans to understand

Example

IF *age* = "<=30" AND *student* = "*no*" THEN *buys_computer* = "*no*"

IF *age* = "<=30" AND *student* = "*yes*" THEN *buys_computer* = "*yes*"

IF *age* = "31…40" THEN *buys_computer* = "*yes*"

IF *age* = ">40" AND *credit_rating* = "*excellent*" THEN *buys_computer* = "*yes*"

IF *age* = ">40" AND *credit_rating* = "*fair*" THEN *buys_computer* = "*no*"

Avoid Overfitting in Classification

The generated tree may overfit the training data

– Too many branches, some may reflect anomalies due to noise or outliers

– Result is in poor accuracy for unseen samples

Two approaches to avoid over fitting

Prepruning:

- Halt tree construction early—do not split a node if this would result in the goodness measure falling below a threshold
- Difficult to choose an appropriate threshold

Post pruning:

- Remove branches from a "fully grown" tree—get a sequence of progressively pruned trees
- Use a set of data different from the training data to decide which is the "best pruned tree"

Tree Mining in Weka.

Tree Mining in Weka

Example:

– Weather problem: build a decision tree to guide the decision about whether or not to play tennis.

– Dataset (weather.nominal.arff)

Validation:

– Using training set as a test set will provide optimal classification accuracy.

– Expected accuracy on a different test set will always be less.

– 10-fold cross validation is more robust than using the training set as a test set.

Divide data into 10 sets with about same proportion of class label values as in original set.

Run classification 10 times independently with the remaining 9/10 of the set as the training set.

Average accuracy.

– Ratio validation: 67% training set / 33% test set.

– Best: having a separate training set and test set.

Results:

– Classification accuracy (correctly classified instances).

– Errors (absolute mean, root squared mean, …)

– Kappa statistic (measures agreement between predicted and observed classification; -100%-100% is the proportion of agreements after chance agreement has been excluded; 0% means complete agreement by chance)

Results:

– TP (True Positive) rate per class label

– FP (False Positive) rate

– Precision = TP rate = TP / (TP + FN)) * 100%

– Recall = TP / (TP + FP)) * 100%

– F-measure = 2* recall * precision / recall + precision

ID3 characteristics:

– Requires nominal values

– Improved into C4.5

- Dealing with numeric attributes

- Dealing with missing values
- Dealing with noisy data
- Generating rules from trees

Attribute Selection Measures

Information Gain

Gain ratio

Gini Index

Pruning of decision trees

Discarding one or more sub trees and replacing them with leaves simplify a decision tree, and that is the main task in decision-tree pruning. In replacing the sub tree with a leaf, the algorithm expects to lower the *predicted error rate and* increase the quality of a classification model. But computation of error rate is not simple. An error rate based only on a training data set does not provide a suitable estimate. One possibility to estimate the predicted error rate is to use a new, additional set of test samples if they are available, or to use the cross-validation techniques. This technique divides initially available samples into equal sized blocks and, for each block, the tree is constructed from all samples except this block and tested with a given block of samples. With the available training and testing samples, the basic idea of decision tree-pruning is to remove parts of the tree (sub trees) that do not contribute to the classification accuracy of unseen testing samples, producing a less complex and thus more comprehensible tree. There are two ways in which the recursive-partitioning method can be modified:

Deciding not to divide a set of samples any further under some conditions. The stopping criterion is usually based on some statistical tests, such as the χ^2 test: *If there are no significant differences in classification accuracy before and after division, then represent a current node as a leaf*. The decision is made in advance, before splitting, and therefore this approach is called *pre pruning*.

Removing restrospectively some of the tree structure using selected accuracy criteria. The decision in this process of *post pruning* is made after the tree has been built.

C4.5 follows the *post pruning* approach, but it uses a specific technique to estimate the predicted error rate. This method is called *pessimistic pruning*. For every node in a tree, the estimation of the upper confidence limit u_{cf} is computed using the statistical tables

for binomial distribution (given in most textbooks on statistics). Parameter U_{cf} is a function of $|T_i|$ and E for a given node. C4.5 uses the default confidence level of 25%, and compares $U_{25\%}$ ($|T_i|$ /E) for a given node T_i with a weighted confidence of its leaves. Weights are the total number of cases for every leaf. If the predicted error for a root node in a sub tree is less than weighted sum of $U_{25\%}$ for the leaves (predicted error for the sub tree), then a sub tree will be replaced with its root node, which becomes a new leaf in a pruned tree.

Let us illustrate this procedure with one simple example. A sub tree of a decision tree is given in Figure, where the root node is the test x_1 on three possible values {1, 2, 3} of the attribute A. The children of the root node are leaves denoted with corresponding classes and ($|T_i|$ /E) parameters. The question is to estimate the possibility of pruning the sub tree and replacing it with its root node as a new, generalized leaf node.

To analyze the possibility of replacing the sub tree with a leaf node it is necessary to compute a predicted error PE for the initial tree and for a replaced node. Using default confidence of 25%, the upper confidence limits for all nodes are collected from statistical tables: $U_{25\%}$ (6, 0) = 0.206, $U_{25\%}$(9, 0) = 0.143, $U_{25\%}$(1, 0) = 0.750, and $U_{25\%}$(16, 1) = 0.157. Using these values, the predicted errors for the initial tree and the replaced node are

PE_{tree} = 6.0.206 + 9.0.143 + 1.0.750 = 3.257

PE_{node} = 16.0.157 = 2.512

Since the existing subtree has a higher value of predicted error than the replaced node, it is recommended that the decision tree be pruned and the subtree replaced with the new leaf node.

Rule Based Classification

Using IF-THEN Rules for Classification

Represent the knowledge in the form of IF-THEN rules

IF *age* = youth AND *student* = yes THEN *buys_computer* = yes

Rule antecedent/precondition vs. rule consequent

Assessment of a rule: *coverage* and *accuracy*

n_{covers} = # of tuples covered by R

$n_{correct}$ = # of tuples correctly classified by R

coverage(R) = n_{covers} /|D| /* D: training data set */ accuracy(R) = $n_{correct}$ / n_{covers}

If more than one rule is triggered, need conflict resolution

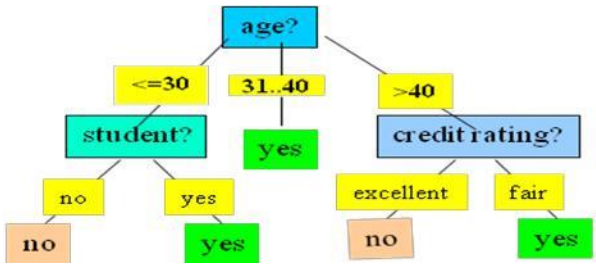
Size ordering: assign the highest priority to the triggering rules that has the "toughest" requirement (i.e., with the *most attribute test*)

Class-based ordering: decreasing order of *prevalence or misclassification cost per class*

Rule-based ordering (decision list): rules are organized into one long priority list, according to some measure of rule quality or by experts

Rule Extraction from a Decision Tree

- Rules are easier to understand than large trees
- One rule is created for each path from the root to a leaf
- Each attribute-value pair along a path forms a conjunction: the leaf holds the class prediction
- Rules are mutually exclusive and exhaustive



Example: Rule extraction from our *buys_computer* decision-tree

IF *age* = young AND *student* = *no* THEN *buys_computer* = *no*

IF *age* = young AND *student* = *yes* THEN *buys_computer* = *yes*

IF *age* = mid-age THEN *buys_computer* = *yes*

IF *age* = old AND *credit_rating* = *excellent* THEN *buys_computer* = *yes* IF *age* = young AND *credit_rating* = *fair* THEN *buys_computer* = *no*

Rule Extraction from the Training Data

Sequential covering algorithm: Extracts rules directly from training data

Typical sequential covering algorithms: FOIL, AQ, CN2, RIPPER

Rules are learned *sequentially*, each for a given class C_i will cover many tuples of C_i but none (or few) of the tuples of other classes

Steps:

1. Rules are learned one at a time

2. Each time a rule is learned, the tuples covered by the rules are removed

3. The process repeats on the remaining tuples unless *termination condition*, e.g., when no more training examples or when the quality of a rule returned is below a user-specified threshold
4. Comp. w. decision-tree induction: learning a set of rules *simultaneously*

What Is Prediction?

- (Numerical) prediction is similar to classification
  - construct a model
  - use model to predict continuous or ordered value for a given input
- Prediction is different from classification
  - Classification refers to predict categorical class label
  - Prediction models continuous-valued functions
- Major method for prediction: regression
  - model the relationship between one or more *independent* or **predictor** variables and a *dependent* or **response** variable
- Regression analysis
  - Linear and multiple regression
  - Non-linear regression
  - Other regression methods: generalized linear model, Poisson regression, log-linear models, regression trees

Linear Regression

- Linear regression: involves a response variable y and a single predictor variable x

$y = w_0 + w_1 x$

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2} \qquad w_0 = \bar{y} - w_1 \bar{x}$$

where w_0 (y-intercept) and w_1 (slope) are regression coefficients

- Method of least squares: estimates the best-fitting straight line

- Multiple linear regression: involves more than one predictor variable
- Training data is of the form $(\mathbf{X_1}, y_1)$, $(\mathbf{X_2}, y_2)$,…, $(\mathbf{X_{|D|}}, y_{|D|})$
- Ex. For 2-D data, we may have: $y = w_0 + w_1 x_1 + w_2 x_2$
- Solvable by extension of least square method or using SAS, S-Plus
- Many nonlinear functions can be transformed into the above

Nonlinear Regression

- Some nonlinear models can be modeled by a polynomial function
- A polynomial regression model can be transformed into linear regression model. For example,

$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$

convertible to linear with new variables: $x_2 = x^2$, $x_3 = x^3$

$y = w_0 + w_1 x + w_2 x_2 + w_3 x_3$

- Other functions, such as power function, can also be transformed to linear model
- Some models are intractable nonlinear (e.g., sum of exponential terms)
  - possible to obtain least square estimates through extensive calculation on more complex formulae

Other Regression-Based Models

- Generalized linear model:
  - Foundation on which linear regression can be applied to modeling categorical response variables
  - Variance of y is a function of the mean value of y, not a constant
  - Logistic regression: models the prob. of some event occurring as a linear function of a set of predictor variables
  - Poisson regression: models the data that exhibit a Poisson distribution
- Log-linear models: (for categorical data)
  - Approximate discrete multidimensional prob. distributions
  - Also useful for data compression and smoothing
- Regression trees and model trees
  - Trees to predict continuous values rather than class labels

Regression Trees and Model Trees

- Regression tree: proposed in CART system
  - CART: Classification And Regression Trees
  - Each leaf stores a *continuous-valued prediction*
  - It is the *average value of the predicted attribute* for the training tuples that reach the leaf
  - Model tree:
  - Each leaf holds a regression model—a multivariate linear equation for the predicted attribute
  - A more general case than regression tree
- Regression and model trees tend to be more accurate than linear regression when the data are not represented well by a simple linear model

Classifier Accuracy Measures

- Accuracy of a classifier M, acc(M): percentage of test set tuples that are correctly classified by the model M
  - Error rate (misclassification rate) of M = 1 – acc(M)
  - Given *m* classes, $CM_{i,j}$, an entry in a **confusion matrix**, indicates # of tuples in class *i* that are labeled by the classifier as class *j*
- Alternative accuracy measures (e.g., for cancer diagnosis)

sensitivity = t-pos/pos /* true positive recognition rate */

specificity = t-neg/neg /* true negative recognition rate */

precision = t-pos/(t-pos + f-pos)

accuracy = sensitivity * pos/(pos + neg) + specificity * neg/(pos + neg)

  - This model can also be used for cost-benefit analysis

classes	buy_computer = yes	buy_computer = no	total	recognition(%)
buy_computer = yes	6954	46	7000	99.34
buy_computer = no	412	2588	3000	86.27
total	7366	2634	10000	95.52

	C_1	C_2
C_1	True positive	False negative
C_2	False positive	True negative

Predictor Error Measures

- Measure predictor accuracy: measure how far off the predicted value is from the actual known value
- **Loss function**: measures the error betw. y_i and the predicted value y_i'
  - Absolute error: $| y_i – y_i' |$
  - Squared error: $(y_i – y_i')^2$
- Test error (generalization error): the average loss over the test set
  - Mean absolute error:
  - Mean squared error:
  - Relative absolute error:
  - Relative squared error:

The mean squared-error exaggerates the presence of outliers Popularly use (square) root mean-square error, similarly, root relative squared error\

Evaluating the Accuracy of a Classifier or Predictor

- Holdout method
  - Given data is randomly partitioned into two independent sets
    - Training set (e.g., 2/3) for model construction
    - Test set (e.g., 1/3) for accuracy estimation
  - Random sampling: a variation of holdout
    - Repeat holdout k times, accuracy = avg. of the accuracies obtained
- Cross-validation (*k*-fold, where k = 10 is most popular)
  - Randomly partition the data into *k mutually exclusive* subsets, each approximately equal size
  - At *i*-th iteration, use D_i as test set and others as training set

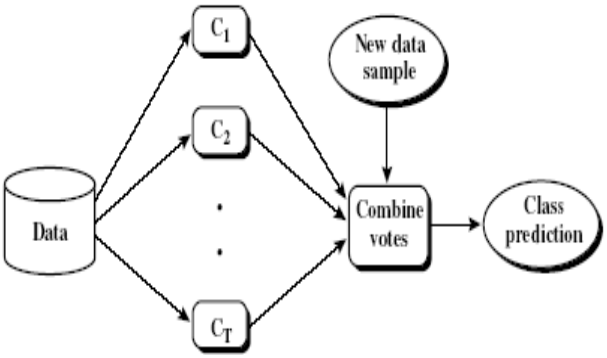
- Leave-one-out: k folds where k = # of tuples, for small sized data
- Stratified cross-validation: folds are stratified so that class dist. in each fold is approx. the same as that in the initial data

- Bootstrap
  - Works well with small data sets
  - Samples the given training tuples uniformly *with replacement*
    - i.e., each time a tuple is selected, it is equally likely to be selected again and re-added to the training set
- Several boostrap methods, and a common one is **.632 boostrap**
  - Suppose we are given a data set of d tuples. The data set is sampled d times, with replacement, resulting in a training set of d samples. The data tuples that did not make it into the training set end up forming the test set. About 63.2% of the original data will end up in the bootstrap, and the remaining 36.8% will form the test set (since $(1 - 1/d)^d \approx e^{-1} = 0.368$)
  - Repeat the sampling procedue k times, overall accuracy of the model:

$$acc(M) = \sum_{i=1}^{k} (0.632 \times acc(M_i)_{test\_set} + 0.368 \times acc(M_i)_{train\_set})$$

Ensemble Methods: Increasing the Accuracy

- Ensemble methods
  - Use a combination of models to increase accuracy
  - Combine a series of k learned models, M_1, M_2, …, M_k, with the aim of creating an improved model M*
- Popular ensemble methods
  - Bagging: averaging the prediction over a collection of classifiers
  - Boosting: weighted vote with a collection of classifiers
  - Ensemble: combining a set of heterogeneous classifiers



Bagging: Boostrap Aggregation

- Analogy: Diagnosis based on multiple doctors' majority vote
- Training
  - Given a set D of *d* tuples, at each iteration *i*, a training set D_i of *d* tuples is sampled with replacement from D (i.e., boostrap)
  - A classifier model M_i is learned for each training set D_i
- Classification: classify an unknown sample **X**
  - Each classifier M_i returns its class prediction
  - The bagged classifier M* counts the votes and assigns the class with the most votes to **X**
- Prediction: can be applied to the prediction of continuous values by taking the average value of each prediction for a given test tuple
- Accuracy
  - Often significant better than a single classifier derived from D
  - For noise data: not considerably worse, more robust

Proved improved accuracy in prediction

Boosting

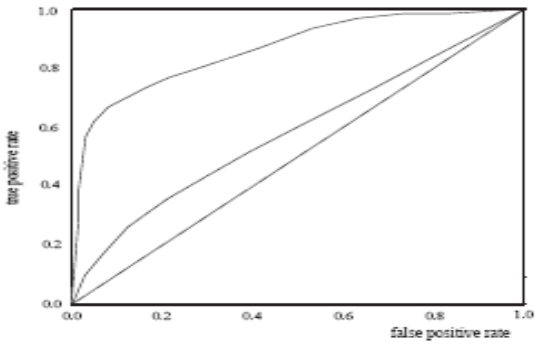
- Analogy: Consult several doctors, based on a combination of weighted diagnoses—weight assigned based on the previous diagnosis accuracy
- How boosting works?
  - Weights are assigned to each training tuple
  - A series of k classifiers is iteratively learned

- After a classifier M_i is learned, the weights are updated to allow the subsequent classifier, M_{i+1}, to pay more attention to the training tuples that were misclassified by M_i
- The final M* combines the votes of each individual classifier, where the weight of each classifier's vote is a function of its accuracy

- The boosting algorithm can be extended for the prediction of continuous values
- Comparing with bagging: boosting tends to achieve greater accuracy, but it also risks overfitting the model to misclassified data

Model Selection: ROC Curves

- ROC (Receiver Operating Characteristics) curves: for visual comparison of classification models
- Originated from signal detection theory
- Shows the trade-off between the true positive rate and the false positive rate
- The area under the ROC curve is a measure of the accuracy of the model
- Rank the test tuples in decreasing order: the one that is most likely to belong to the positive class appears at the top of the list
- The closer to the diagonal line (i.e., the closer the area is to 0.5), the less accurate is the model



- Vertical axis represents the true positive rate
- Horizontal axis rep. the false positive rate
- The plot also shows a diagonal line
- A model with perfect accuracy will have an area of 1.0

POSSIBLE QUESTIONS

2 MARKS:

1. What is classification?
2. Differentiate classification and prediction.
3. What is supervised learning?
4. Define decision trees.
5. What is the rule used in rule based classification?
6. What is a confusion matrix?
7. List any four error measures for prediction.
8. What is the purpose of ensemble methods.
9. What is boosting?
10. What is bagging?

Part B:

1. Briefly outline the major steps of decision tree classification.
2. Describe the various techniques for improving classifier accuracy.
3. Explain with an example the various steps in Decision tree induction.
4. Explain the issues regarding classification and prediction?
5. What are classification rules? How is regression related to classification?

Part C: 10 marks

1. Why tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?
2. Explain in detail about the following:
    1) Rule-based classification
    2) Predictions

UNIT V

SYLLABUS: Cluster Analysis Introduction : Types of Data in Cluster Analysis - A Categorization of Major Clustering Methods - Partitioning Methods - Hierarchical Methods – Density-Based Methods Grid-Based Methods - Model-Based Clustering Methods - Clustering High-Dimensional Data – Constraint-Based Cluster Analysis - Outlier Analysis.

Applications and Trends in Data mining: Text Mining - Web Mining - Multimedia Mining-Spatial Mining - Visual data mining.

Cluster Analysis

Cluster: a collection of data objects

– Similar to one another within the same cluster
– Dissimilar to the objects in other clusters

Cluster analysis

– Grouping a set of data objects into clusters

Clustering is unsupervised classification: no predefined classes

Typical applications

– As a stand-alone tool to get insight into data distribution
– As a preprocessing step for other algorithms

General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining, Image Processing Economic Science (especially market research),WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location

Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

What Is Good Clustering?

A good clustering method will produce high quality clusters with

– high intra-class similarity
– low inter-class similarity

The quality of a clustering result depends on both the similarity measure used by the method and its implementation.

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

Requirements of Clustering in Data Mining

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

Type of data in clustering analysis

- Interval-scaled variables:
- Binary variables:
- Nominal, ordinal, and ratio variables:
- Variables of mixed types:

A Categorization of Major Clustering Methods:

- **Partitioning algorithms**: Construct various partitions and then evaluate them by some criterion
- **Hierarchy algorithms**: Create a hierarchical decomposition of the set of data (or objects) using some criterion
- **Density-based**: based on connectivity and density functions
- **Grid-based**: based on a multiple-level granularity structure
- **Model-based**: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

Partitioning Algorithms: Basic Concept

Partitioning method: Construct a partition of a database *D* of *n* objects into a set of *k* clusters

Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion

– Global optimal: exhaustively enumerate all partitions
– Heuristic methods: *k-means* and *k-medoids* algorithms
– *k-means* : Each cluster is represented by the center of the cluster

– *k-medoids* or PAM (Partition around medoids) : Each cluster is represented by one of the objects in the cluster

The *K-Means* Clustering Method

- Given *k*, the *k-means* algorithm is implemented in 4 steps:
    1. Partition objects into *k* nonempty subsets
    2. Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
    3. Assign each object to the cluster with the nearest seed point.
    4. Go back to Step 2, stop when no more new assignment.

Comments on the *K-Means* Method

Strength

– *Relatively efficient*: *O*(*tkn*), where *n* is # objects, *k* is # clusters, and *t* is # iterations. Normally, *k*, *t* << *n*.

– Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* and *genetic algorithms*

Weakness

– Applicable only when *mean* is defined, then what about categorical data?

– Need to specify *k*, the *number* of clusters, in advance

– Unable to handle noisy data and *outliers*

– Not suitable to discover clusters with *non-convex shapes*

Variations of the *K-Means* Method

A few variants of the *k-means* which differ in

– Selection of the initial *k* means

– Dissimilarity calculations

– Strategies to calculate cluster means

Handling categorical data: *k-modes*

– Replacing means of clusters with modes

– Using new dissimilarity measures to deal with categorical objects

– Using a frequency-based method to update modes of clusters

– A mixture of categorical and numerical data: *k-prototype* method

***K-Medoids* Clustering Method**

- Find *representative* objects, called medoids, in clusters
- *PAM* (Partitioning Around Medoids, 1987)
    - starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
    - *PAM* works effectively for small data sets, but does not scale well for large data sets

PAM (Partitioning Around Medoids)

- PAM (Kaufman and Rousseeuw, 1987), built in Splus
- Use real object to represent the cluster

– Select ***k*** representative objects arbitrarily

– For each pair of non-selected object ***h*** and selected object ***i***, calculate the total swapping cost ***TCih***

- For each pair of ***i*** and ***h***,
    - If *TCih* < 0, ***i*** is replaced by ***h***
    - Then assign each non-selected object to the most similar representative object
- repeat steps 2-3 until there is no change

Hierarchical Clustering

Use distance matrix as clustering criteria. This method does not require the number of clusters *k* as an input, but needs a termination condition

AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Use the Single-Link method and the dissimilarity matrix.
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster

A *Dendrogram* Shows How the Clusters are Merged Hierarchically

- Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.
- A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.

DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Inverse order of AGNES
- Eventually each node forms a cluster on its own

BIRCH

- Birch: Balanced Iterative Reducing and Clustering using Hierarchies
- Incrementally construct a CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering

– Phase 1: scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data)

– Phase 2: use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree

Scales linearly: finds a good clustering with a single scan and improves the quality with a few additional scans

Weakness: handles only numeric data, and sensitive to the order of the data record.

Hierarchical and Non-Hierarchical Clustering

There are two main types of clustering techniques, those that create a hierarchy of clusters and those that do not. The hierarchical clustering techniques create a hierarchy of clusters from small to big. The main reason for this is that, as was already stated, clustering is an unsupervised learning technique, and as such, there is no absolutely correct answer. For this reason and depending on the particular application of the clustering, fewer or greater numbers

of clusters may be desired. With a hierarchy of clusters defined it is possible to choose the number of clusters that are desired. At the extreme it is possible to have as many clusters as there are records in the database. In this case the records within the cluster are optimally similar to each other (since there is only one) and certainly different from the other clusters. But of course such a clustering technique misses the point in the sense that the idea of clustering is to find useful patters in the database that summarize it and make it easier to understand. Any clustering algorithm that ends up with as many clusters as there are records has not helped the user understand the data any better. Thus one of the main points about clustering is that there be many fewer clusters than there are original records. Exactly how many clusters should be formed is a matter of interpretation. The advantage of hierarchical clustering methods is that they allow the end user to choose from either many clusters or only a few.

The hierarchy of clusters is usually viewed as a tree where the smallest clusters merge together to create the next highest level of clusters and those at that level merge together to create the next highest level of clusters. Figure 1.5 below shows how several clusters might form a hierarchy. When a hierarchy of clusters like this is created the user can determine what the right number of clusters is that adequately summarizes the data while still providing useful information (at the other extreme a single cluster containing all the records is a great summarization but does not contain enough specific information to be useful).

This hierarchy of clusters is created through the algorithm that builds the clusters. There are two main types of hierarchical clustering algorithms:

Agglomerative - Agglomerative clustering techniques start with as many clusters as there are records where each cluster contains just one record. The clusters that are nearest each other are merged together to form the next largest cluster. This merging is continued until a hierarchy of clusters is built with just a single cluster containing all the records at the top of the hierarchy.

Divisive - Divisive clustering techniques take the opposite approach from agglomerative techniques. These techniques start with all the records in one cluster and then try to split that cluster into smaller pieces and then in turn to try to split those smaller pieces.

Of the two the agglomerative techniques are the most commonly used for clustering and have more algorithms developed for them. We'll talk about these in more detail in the next section.

The non-hierarchical techniques in general are faster to create from the historical database but require that the user make some decision about the number of clusters desired or the minimum "nearness" required for two records to be within the same cluster. These non-hierarchical techniques often times are run multiple times starting off with some arbitrary or even random clustering and then iteratively improving the clustering by shuffling some records around. Or these techniques sometimes create clusters that are created with only one pass through the database adding records to existing clusters when they exist and creating new clusters when no existing cluster is a good candidate for the given record. Because the definition of which clusters are formed can depend on these initial choices of which starting clusters should be chosen or even how many clusters these techniques can be less repeatable than the hierarchical techniques and can sometimes create either too many or too few clusters

because the number of clusters is predetermined by the user not determined solely by the patterns inherent in the database.

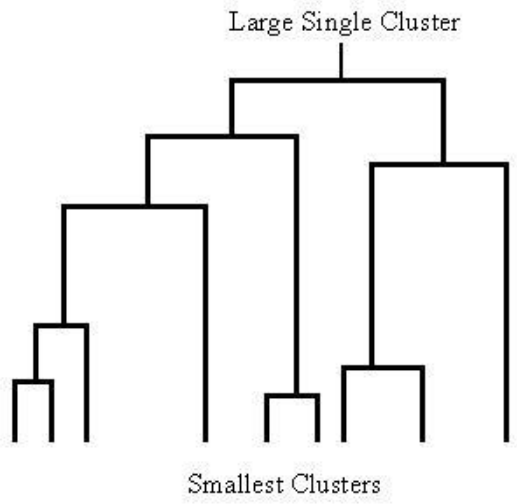


Diagram showing a hierarchy of clusters. Clusters at the lowest level are merged together to form larger clusters at the next level of the hierarchy.

Non-Hierarchical Clustering

There are two main non-hierarchical clustering techniques. Both of them are very fast to compute on the database but have some drawbacks. The first are the single pass methods. They derive their name from the fact that the database must only be passed through once in order to create the clusters (i.e. each record is only read from the database once). The other class of techniques are called reallocation methods. They get their name from the movement or "reallocation" of records from one cluster to another in order to create better clusters. The reallocation techniques do use multiple passes through the database but are relatively fast in comparison to the hierarchical techniques.

Density-Based Clustering Methods

Clustering based on density (local cluster criterion), such as density-connected points

Major features:

– Discover clusters of arbitrary shape
– Handle noise
– One scan
– Need density parameters as termination condition

Several interesting studies:

– DBSCAN:
– OPTICS:
– DENCLUE:
– CLIQUE:

Density-Based Clustering: Background

Two parameters*:*

– *Eps*: Maximum radius of the neighbour hood
– *MinPts*: Minimum number of points in an Eps-neighbour hood of that point

NEps(p):{q belongs to D | dist(p,q) <= Eps}

Directly density-reachable: A point *p* is directly density-reachable from a point *q* wrt. *Eps, MinPts* if

– 1) *p* belongs to *NEps(q)*

– 2) core point condition:

|NEps (q)| >= MinPts

Density-reachable:

– A point *p* is density-reachable from a point *q* wrt. *Eps, MinPts* if there is a chain of points *p1*, …, *pn, p1 = q, pn = p* such that *pi+1* is directly density-reachable from *pi*

Density-connected

– A point *p* is density-connected to a point *q* wrt. *Eps, MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable from *o* wrt. *Eps* and *MinPts*.

DBSCAN: Density Based Spatial Clustering of Applications with Noise

Relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points

Discovers clusters of arbitrary shape in spatial databases with noise

DBSCAN: The Algorithm

– Arbitrary select a point *p*

– Retrieve all points density-reachable from *p* wrt *Eps* and *MinPts*.

– If *p* is a core point, a cluster is formed.

– If *p* is a border point, no points are density-reachable from *p* and DBSCAN visits the next point of the database.

– Continue the process until all of the points have been processed

OPTICS: A Cluster-Ordering Method (1999)

OPTICS: Ordering Points To Identify the Clustering Structure

– Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)

– Produces a special order of the database wrt its density-based clustering structure

– This cluster-ordering contains info equiv to the density-based clusterings corresponding to a broad range of parameter settings

– Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure

– Can be represented graphically or using visualization techniques

OPTICS: Some Extension from DBSCAN

Index-based:

k = number of dimensions

N = 20

p = 75%

M = N(1-p) = 5

– Complexity: O(*kN2*)

Core Distance

Reachability Distance

Max (core-distance (o), d (o, p))

r(p1, o) = 2.8cm. r(p2,o) = 4cm

DENCLUE: using density functions

DENsity-based CLUstEring by Hinneburg & Keim (KDD'98)

Major features

– Solid mathematical foundation

– Good for data sets with large amounts of noise

– Allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets

– Significant faster than existing algorithm (faster than DBSCAN by a factor of up to 45)

– But needs a large number of parameters

Denclue: Technical Essence

- Uses grid cells but only keeps information about grid cells that do actually contain data points and manages these cells in a tree-based access structure.
- Influence function: describes the impact of a data point within its neighborhood.
- Overall density of the data space can be calculated as the sum of the influence function of all data points.
- Clusters can be determined mathematically by identifying density attractors.
- Density attractors are local maximal of the overall density function.

Grid-Based Methods

Using multi-resolution grid data structure

Several interesting methods

– STING

– WaveCluster : A multi-resolution clustering approach using wavelet method

– CLIQUE: Agrawal, et al. (SIGMOD'98)

STING: A Statistical Information Grid Approach

- The spatial area is divided into rectangular cells
- There are several levels of cells corresponding to different levels of resolution

STING: A Statistical Information Grid Approach (2)

- Each cell at a high level is partitioned into a number of smaller cells in the next lower level
- Statistical info of each cell is calculated and stored beforehand and is used to answer queries
- Parameters of higher level cells can be easily calculated from parameters of lower level cell

  count, *mean*, *s*, *min*, *max*

  type of distribution—normal, *uniform*, etc.

– Use a top-down approach to answer spatial data queries

– Start from a pre-selected layer—typically with a small number of cells

– For each cell in the current level compute the confidence interval

STING: A Statistical Information Grid Approach (3)

– Remove the irrelevant cells from further consideration

– When finish examining the current layer, proceed to the next lower level
– Repeat this process until the bottom layer is reached
– **Advantages:**
Query-independent, easy to parallelize, incremental update
O(K), where *K* is the number of grid cells at the lowest level
– **Disadvantages:**
All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected

WaveCluster

A multi-resolution clustering approach which applies wavelet transform to the feature space
– A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub-band.
Both grid-based and density-based
Input parameters:
– # of grid cells for each dimension
– the wavelet, and the # of applications of wavelet transform.

How to apply wavelet transform to find clusters

– Summaries the data by imposing a multidimensional grid structure onto data space
– These multidimensional spatial data objects are represented in a n-dimensional feature space
– Apply wavelet transform on feature space to find the dense regions in the feature space
– Apply wavelet transform multiple times which result in clusters at different scales from fine to coarse

Why is wavelet transformation useful for clustering

– Unsupervised clustering
It uses hat-shape filters to emphasize region where points cluster, but simultaneously to suppress weaker information in their boundary
– Effective removal of outliers
– Multi-resolution
– Cost efficiency
Major features:
– Complexity O(N)
– Detect arbitrary shaped clusters at different scales
– Not sensitive to noise, not sensitive to input order
– Only applicable to low dimensional data

Model-Based Clustering Methods:

1.Attempt to optimize the fit between the data and some mathematical model
Statistical and AI approach Conceptual clustering
A form of clustering in machine learning
Produces a classification scheme for a set of unlabeled objects
Finds characteristic description for each concept (class)

COBWEB (Fisher'87)

A popular a simple method of incremental conceptual learning

Creates a hierarchical clustering in the form of a classification tree

Each node refers to a concept and contains a probabilistic description of that concept

Model-Based Clustering Methods

Attempt to optimize the fit between the data and some mathematical model

Statistical and AI approach

– Conceptual clustering

A form of clustering in machine learning

Produces a classification scheme for a set of unlabeled objects

Finds characteristic description for each concept (class)

– COBWEB (Fisher'87)

A popular a simple method of incremental conceptual learning

Creates a hierarchical clustering in the form of a classification tree

Each node refers to a concept and contains a probabilistic description of that concept

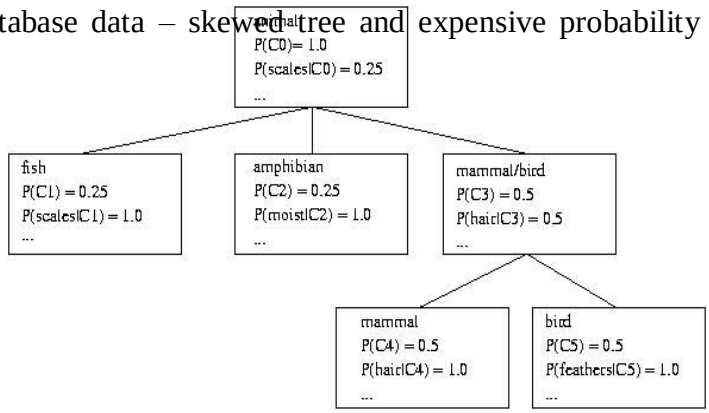
COBWEB Clustering Method

A classification tree More on Statistical-Based Clustering

Limitations of COBWEB

– The assumption that the attributes are independent of each other is often too strong because correlation may exist

– Not suitable for clustering large database data – skewed tree and expensive probability distributions



Outlier Analysis

What are outliers?

– The set of objects are considerably dissimilar from the remainder of the data

– Example: Sports: Michael Jordon, Wayne Gretzky, ...

Problem

– Find top n outlier points

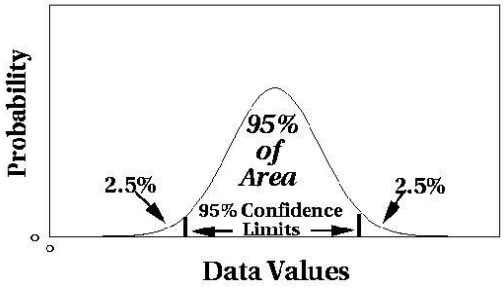
Applications:

– Credit card fraud detection

– Telecom fraud detection

– Customer segmentation

– Medical analysis



Outlier Discovery: Statistical Approaches

Assume a model underlying distribution that generates data set (e.g. normal distribution)

Use discordance tests depending on

– data distribution
– distribution parameter (e.g., mean, variance)
– number of expected outliers

Drawbacks

– most tests are for single attribute
– In many cases, data distribution may not be known

Outlier Discovery: Distance-Based Approach

Introduced to counter the main limitations imposed by statistical methods

– We need multi-dimensional analysis without knowing data distribution.

Distance-based outlier: A DB(p, D)-outlier is an object O in a dataset T such that at least a fraction p of the objects in T lies at a distance greater than D from O

Algorithms for mining distance-based outliers

– Index-based algorithm
– Nested-loop algorithm
– Cell-based algorithm

Outlier Discovery: Deviation-Based Approach

Identifies outliers by examining the main characteristics of objects in a group

Objects that "deviate" from this description are considered outliers

sequential exception technique

– simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects

OLAP data cube technique

– uses data cubes to identify regions of anomalies in large multidimensional data

Spatial Data Mining

A spatial database stores a large amount of space-related data, such as maps, preprocessed remote sensing or medical imaging data, and VLSI chip layout data. Spatial databases have many features distinguishing them from relational databases. They carry topological and/or distance information, usually organized by sophisticated, multidimensional spatial indexing structures that are accessed by spatial data access methods and often require spatial reasoning, geometric computation, and spatial knowledge representation techniques.

Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases. Such mining demands an integration of data mining with spatial database technologies. It can be used for understanding spatial data, discovering spatial relationships and relationships between spatial and non spatial data, constructing spatial knowledge bases, reorganizing spatial databases, and optimizing spatial queries.

Statistical spatial data analysis has been a popular approach to analyzing spatial data and exploring geographic information. The term geostatistics is often associated with continuous geographic space, partial statistics is often associated with discrete space.

In a statistical model that handles nonspatial data, one usually assumes statistical independence among different portions of data. However, different from traditional data sets, there is no such independence among spatially distributed data because in reality, spatial objects are often interrelated, or more exactly spatially co-located, in the sense that the closer the two objects are located, the more likely they share similar properties.

People even consider this as the first law of geography: "Everything is related to everything else, but nearby things are more related than distant things." Such a property of close interdependency across nearby space leads to the notion of spatial autocorrelation.

Spatial Data Cube Construction and Spatial OLAP

"Can we construct a spatial data warehouse?" Yes, as with relational data, we can integrate spatial data to construct a data warehouse that facilitates spatial data mining.

A spatial data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of both spatial and non spatial data in support of spatial data mining and spatial-data related decision-making processes.

There are three types of dimensions in a spatial data cube:

A non spatial dimension contains only non spatial data.

- Non spatial dimensions temperature and precipitation can be constructed for the warehouse in since each contains non spatial data whose generalizations are non spatial (such as "hot" for temperature and "wet" for precipitation).
- A spatial-to-non spatial dimension is a dimension whose primitive-level data are spatial but whose generalization, starting at a certain high level, becomes non spatial. For example, the spatial dimension city relays geographic data for the U.S. map.

Suppose that the dimension's spatial representation of, say, Seattle is generalized to the string "pacific northwest." Although "pacific northwest" is a spatial concept, its representation is not spatial (since, in our example, it is a string). It therefore plays the role of a non spatial dimension.

A spatial-to-spatial dimension is a dimension whose primitive level and all of its high level generalized data are spatial. For example, the dimension equi temperature region contains spatial data, as do all of its generalizations, such as with regions covering 0-5 degrees (Celsius), 5-10 degrees, and so on.

We distinguish two types of measures in a spatial data cube:

1) A numerical measure contains only numerical data. For example, one measure in a spatial data warehouse could be the monthly revenue of a region, so that a roll-up may compute the total revenue by year, by county, and so on. Numerical measures can be further classified into distributive, algebraic, and holistic

2) A spatial measure contains a collection of pointers to spatial objects. For example, in a generalization (or roll-up) in the spatial data cube of Example 10.5, the regions with the same range of temperature and precipitation will be grouped into the same cell, and the measure so formed contains a collection of pointers to those regions.

Mining Spatial Association and Co-location Patterns

For mining spatial associations related to the spatial predicate close to,

we can first collect the candidates that pass the minimum support threshold by Applying certain rough spatial evaluation algorithms, for example, using an MBR structure (which registers only two spatial points rather than a set of complex polygons), and
Evaluating the relaxed spatial predicate, g close to, which is a generalized close to covering a broader context that includes close to, touch, and intersect.

Spatial Clustering Methods

Spatial data clustering identifies clusters, or densely populated regions, according to some distance measurement in a large, multidimensional data set.

Spatial Classification and Spatial Trend Analysis

Spatial classification analyzes spatial objects to derive classification schemes in relevance to certain spatial properties, such as the neighborhood of a district, highway, or river.
Spatial trend analysis deals with another issue: the detection of changes and trends along a spatial dimension. Typically, trend analysis detects changes with time, such as the changes of temporal patterns in time-series data. Spatial trend analysis replaces time with space and studies the trend of non spatial or spatial data changing with space.

Mining Raster Databases

Spatial database systems usually handle vector data that consist of points, lines, polygons (regions), and their compositions, such as networks or partitions. Typical examples of such data include maps, design graphs, and 3-D representations of the arrangement of the chains of protein molecules. However, a huge amount of space-related data are in digital raster (image) forms, such as satellite images, remote sensing data, and computer tomography. It is important to explore data mining in raster or image databases. Methods for mining raster and image data are examined in the following section regarding the mining of multimedia data.

Multimedia Data Mining

A multimedia database system stores and manages a large collection of multimedia data, such as audio, video, image, graphics, speech, text, document, and hypertext data, which contain text, text markups, and linkages. Multimedia database systems are increasingly common owing to the popular use of audio video equipment, digital cameras, CD-ROMs, and the Internet. Typical multimedia database systems include NASA's EOS (Earth Observation System), various kinds of image and audio-video databases, and Internet databases.

Similarity Search in Multimedia Data

For similarity searching in multimedia data, we consider two main families of multimedia indexing and retrieval systems: (1) description-based retrieval systems, which build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation; and (2) content-based retrieval systems, which support retrieval based on the image content, such as color histogram, texture, pattern, image topology, and the shape of objects and their layouts and locations within the image.
Image-sample-based queries find all of the images that are similar to the given image sample. This search compares the feature vector (or signature) extracted from the sample with the

feature vectors of images that have already been extracted and indexed in the image database. Based on this comparison, images that are close to the sample image are returned.

Image feature specification queries specify or sketch image features like color, texture, or shape, which are translated into a feature vector to be matched with the feature vectors of the images in the database. Content-based retrieval has wide applications, including medical diagnosis, weather prediction, TV production,

Web search engines for images, and e-commerce. Some systems, such as QBIC (Query By Image Content), support both sample-based and image feature specification queries. There are also systems that support both content based and description-based retrieval.

Several approaches have been proposed and studied for similarity-based retrieval in image databases, based on image signature:

1) Color histogram–based signature:

In this approach, the signature of an image includes color histograms based on the color composition of an image regardless of its scale orientation. This method does not contain any information about shape, image topology, or texture. Thus, two images with similar color composition but that contain very different shapes or textures may be identified as similar, although they could be completely unrelated semantically.

2) Multi feature composed signature:

In this approach, the signature of an image includes a composition of multiple features: color histogram, shape, image topology, and texture. The extracted image features are stored as metadata, and images are indexed based on such metadata. Often, separate distance functions can be defined for each feature and subsequently combined to derive the overall results. Multidimensional content-based search often uses one or a few probe features to search for images containing such (similar) features. It can therefore be used to search for similar images. This is the most popularly used approach in practice.

3) Wavelet-based signature:

This approach uses the dominant wavelet coefficients of an image as its signature. Wavelets capture shape, texture, and image topology information in a single unified framework.1 This improves efficiency and reduces the need for providing multiple search primitives (unlike the second method above). However, since this method computes a single signature for an entire image, it may fail to identify images containing similar objects where the objects differ in location or size.

4) Wavelet-based signature with region-based granularity:

In this approach, the computation and comparison of signatures are at the granularity of regions, not the entire image. This is based on the observation that similar images may contain similar regions, but a region in one image could be a translation or scaling of a matching region in the other. Therefore, a similarity measure between the query image Q and a target image T can be defined in terms of the fraction of the area of the two images covered by matching pairs of regions from Q and T. Such a region-based similarity search can find images containing similar objects, where these objects may be translated or scaled.

Multidimensional Analysis of Multimedia Data

A multimedia data cube can contain additional dimensions and measures for multimedia information, such as color, texture, and shape.

The example database tested in the Multi Media Miner system is constructed as follows. Each image contains two descriptors: a feature descriptor and a layout descriptor.

The original image is not stored directly in the database; only its descriptors are stored. The description information encompasses fields like image file name, image URL, image type (e.g., gif, tiff, jpeg,mpeg, bmp, avi), a list of all known Web pages referring to the image (i.e., parent URLs), a list of keywords, and a thumbnail used by the user interface for image and video browsing.

The feature descriptor is a set of vectors for each visual characteristic. The layout descriptor contains a color layout vector and an edge layout vector.

Classification and Prediction Analysis of Multimedia Data

Classification and predictive modeling have been used for mining multimedia data, especially in scientific research, such as astronomy, seismology, and geo scientific research.

For Example:

Classification and prediction analysis of astronomy data. Taking sky images that have been carefully classified by astronomers as the training set, we can construct models for the recognition of galaxies, stars, and other stellar objects, based on properties like magnitudes, areas, intensity, image moments, and orientation. A large number of sky images taken by telescopes or space probes can then be tested against the constructed models in order to identify new celestial bodies. Similar studies have successfully been performed to identify volcanoes on Venus.

Mining Associations in Multimedia Data

Association rules involving multimedia objects can be mined in image and video databases. At least three categories can be observed:

Associations between image content and non image content features:

A rule like “If at least 50% of the upper part of the picture is blue, then it is likely to represent sky” belongs to this category since it links the image content to the keyword sky.

Associations among image contents that are not related to spatial relationships:

A rule like “If a picture contains two blue squares, then it is likely to contain one red circle as well” belongs to this category since the associations are all regarding image contents.

Associations among image contents related to spatial relationships:

A rule like “If a red triangle is between two yellow squares, then it is likely a big oval-shaped object is underneath” belongs to this category since it associates objects in the image with spatial relationships.

“What are the differences between mining association rules in multimedia databases versus in transaction databases?”

1) An image may contain multiple objects, each with many features such as color, shape, texture, keyword, and spatial location, Such a multi resolution mining strategy substantially reduces the overall data mining cost without loss of the quality and

completeness of data mining results. This leads to an efficient methodology for mining frequent item sets and associations in large multimedia databases.

2) Because a picture containing multiple recurrent objects is an important feature in image analysis, recurrence of the same objects should not be ignored in association analysis.
3) There often exist important spatial relationships among multimedia objects, such as above, beneath, between, nearby, left-of, and so on.

Audio and Video Data Mining

To facilitate the recording, search, and analysis of audio and video information from multimedia data, industry and standardization committees have made great strides toward developing a set of standards for multimedia information description and compression. For example, MPEG-k (developed by MPEG: Moving Picture Experts Group) and JPEG are typical video compression schemes. The most recently released MPEG-7,formally named "Multimedia Content Description Interface," is a standard for describing the multimedia content data. It supports some degree of interpretation of the information meaning, which can be passed onto, or accessed by, a device or a computer.

The MPEG committee standardizes the following elements in MPEG-7: (1) a set of descriptors, where each descriptor defines the syntax and semantics of a feature, such as color, shape, texture, image topology, motion, or title; (2) a set of descriptor schemes, where each scheme specifies the structure and semantics of the relationships between its components (descriptors or description schemes); (3) a set of coding schemes for the descriptors, and (4) a description definition language (DDL) to specify schemes and descriptors. Such standardization greatly facilitates content-based video retrieval and video data mining.

Text Mining

Information is stored in text databases (or document databases), which consist of large collections of documents from various sources, such as news articles, research papers, books, digital libraries, e-mail messages, and Web pages. Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mail, and the World Wide Web (which can also be viewed as a huge, interconnected, dynamic text database). Nowadays most of the information in government, industry, business, and other institutions are stored electronically, in the form of text databases.

Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured. For example, a document may contain a few structured fields, such as title, authors, publication date, category, and so on, but also contain some largely unstructured text components, such as abstract and contents.

Text Data Analysis and Information Retrieval

Information retrieval (IR) is a field that has been developing in parallel with database systems for many years. Due to the abundance of text information, information retrieval has found many applications. There exist many information retrieval systems, such as on-line library catalog systems, on-line document management systems, and the more recently developed

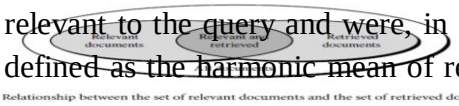
$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}.$$

Web search engines.

Basic Measures for Text Retrieval: Precision and Recall

Precision: This is the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses). It is formally defined as

Recall: This is the percentage of documents that are relevant to the query and were, in fact, retrieved. It is formally defined as F-score, which is defined as the harmonic mean of recall and precision:



Relationship between the set of relevant documents and the set of retrieved documents.

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}.$$

The harmonic mean discourages a system that sacrifices one measure for another too drastically.

Text Retrieval Methods

Retrieval methods fall into two categories: They generally either view the retrieval problem as a document selection problem or as a document ranking problem

In document selection methods, the query is regarded as specifying constraints for selecting relevant documents.

Document ranking methods use the query to rank all documents in the order of relevance. How can we model a document to facilitate information retrieval?" Starting with a set of d documents and a set of t terms, we can model each document as a vector v in the t dimensional space R t , which is why this method is called the vector-space model. Let the term frequency be the number of occurrences of term t in the document d, that is, freq(d; t). The (weighted) term-frequency matrix TF(d; t) measures the association of a term t with respect to the given document d: it is generally defined as 0 if the document does not contain the term, and nonzero otherwise. There are many ways to define the term-weighting for the nonzero entries in such a vector. For example, we can simply set TF(d; t) = 1 if the term t occurs in the document d, or use the term frequency freq(d; t), or the relative term frequency,

that is, the term frequency versus the total number of occurrences of all the terms in the document. There are also other ways to normalize the term frequency. For example, the Cornell SMART system uses the following formula to compute the (normalized) term frequency:

TF(d; t) =0 if freq(d; t) = 0

1+log(1+log(freq(d; t))) otherwise

Inverse document frequency (IDF), that represents the scaling factor, or the importance, of a term t. If a term t occurs in many documents.

A representative metric is the cosine measure, defined as follows. Let v1 and v2 be two document vectors. Their cosine similarity is defined as the tagging approach, where the input is a set of tags, and (3) the information-extraction approach, which inputs semantic information, such as events, facts, or entities uncovered by information extraction.

Various text mining tasks can be performed on the extracted keywords, tags, or semantic information. These include document clustering, classification, information extraction, association analysis, and trend analysis .We examine a few such tasks in the following discussion.

1)Keyword-Based Association Analysis

2)Document Classification Analysis

Mining the World Wide Web

The World Wide Web serves as a huge ,widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services. The Web also contains a rich and dynamic collection of hyperlink information and Web page access and usage information, providing rich sources for data mining. However, based on the following observations, the Web also poses great challenges for effective resource and knowledge discovery.

i)The Web seems to be too huge for effective data warehousing and data mining.

The complexity of Web pages is far greater than that of any traditional text document Collection

The Web is a highly dynamic information source.

The Web serves a broad diversity of user communities

Only a small portion of the information on the Web is truly relevant or useful.

Mining the Web Page Layout Structure

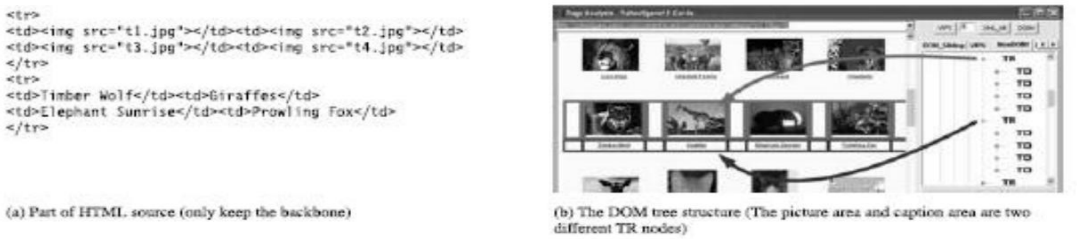
The basic structure of a Web page is its DOM (Document Object Model) structure. The DOM structure of a Web page is a tree structure, where every HTML tag in the page corresponds to a node in the DOM tree.

There are many index-based Web search engines. These search the Web, index Web pages, and build and store huge keyword-based indices that help locate sets of Web pages containing certain keywords. a simple keyword-based search engine suffers from several deficiencies. First, a topic of any breadth can easily contain hundreds of thousands of documents. This can

lead to a huge number of document entries returned by a search engine, many of which are only marginally relevant to the topic or may contain materials of poor quality. Second, many documents that are highly relevant to a topic may not contain keywords defining them. This is referred to as the polysemy problem

Mining the Web's Link Structures to Identify Authoritative Web Pages

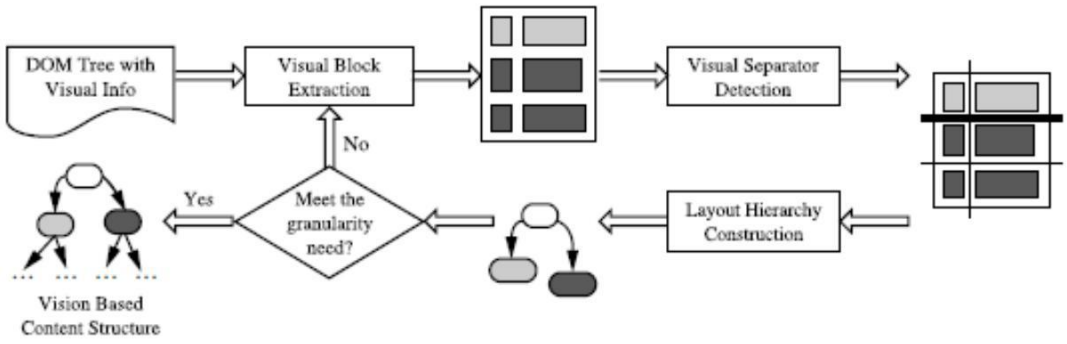
But how can a search engine automatically identify authoritative Web pages for my topic?" Interestingly, the secrecy of authority is hiding in Web page linkages. The Web consists not only of pages, but also of hyperlinks pointing from one page to another. This idea has motivated some interesting studies on mining authoritative pages on the Web These



The HTML source and DOM tree structure of a sample page. It is difficult to extract the correct semantic content structure of the page.

properties of Web link structures have led researchers to consider another important category of Web pages called a hub. A hub is one or a set of Web pages that provides collections of links to authorities. Hub pages may not be prominent, or there may exist few links pointing to them.

An algorithm using hubs, called HITS (Hyperlink-Induced Topic Search), was developed as follows. First, HITS uses the query terms to collect a starting set of, say, 200 pages from an index-based search engine. These pages form the root set. Since many of these pages are presumably relevant to the search topic, some of them should contain links to most of the prominent authorities. Therefore, the root set can be expanded into a base set by including all



The process flow of vision-based page segmentation algorithm.

of the pages that the root-set pages link to and all of the pages that link to a page in the root set, up to a designated size cutoff such as 1,000 to 5,000 pages

We first associate a non-negative authorityweight, ap, and a non-negative hubweight, hp, with each page p in the base set, and initialize all a and h values to a uniform constant

$$a_p = \Sigma_{(q \text{ such that } q \to p)} h_q$$

$$h_p = \Sigma_{(q \text{ such that } q \leftarrow p)} a_q$$

These equations can be written in matrix form as follows. Let us number the pages f1;2; : : : ;ng and define their adjacency matrix A to be an n_n matrix where A(i; j) is 1 if page i links to page j, or 0 otherwise. Similarly, we define the authority weight vector a = (a1;a2; : : : ;an), and the hub weight vector h = (h1;h2; : : : ;hn). Thus, we have

$$\boldsymbol{h} = A \cdot \boldsymbol{a}$$

$$\boldsymbol{a} = A^T \cdot \boldsymbol{h},$$

where AT is the transposition of matrix A. Unfolding these two equations k times, we have
The graph model in block-level link analysis is induced from two kinds of relationships ,that is, block-to-page (link structure) and page-to-block (page layout). The block-to-page relationship is obtained from link analysis. Let Z denote the block-to-page matrix with dimension n_k. Z can be formally defined as follows:

$$\boldsymbol{h} = A \cdot \boldsymbol{a} = AA^T\boldsymbol{h} = (AA^T)\boldsymbol{h} = (AA^T)^2\boldsymbol{h} = \cdots = (AA^T)^k\boldsymbol{h}$$

$$\boldsymbol{a} = A^T \cdot \boldsymbol{h} = A^TA\boldsymbol{a} = (A^TA)\boldsymbol{a} = (A^TA)^2\boldsymbol{a} = \cdots = (A^TA)^k\boldsymbol{a}.$$

where si is the number of pages to which block i links. Zi j can also be viewed as a probability of jumping from block i to page j

The page-to-block relationships are obtained from page layout analysis. Let X denote the page-to-block matrix with dimension k_n.
where f is a function that assigns to every block b in page p an importance value. Specifically, the bigger fp(b) is, the more important the block b is. Function f is empirically defined below,

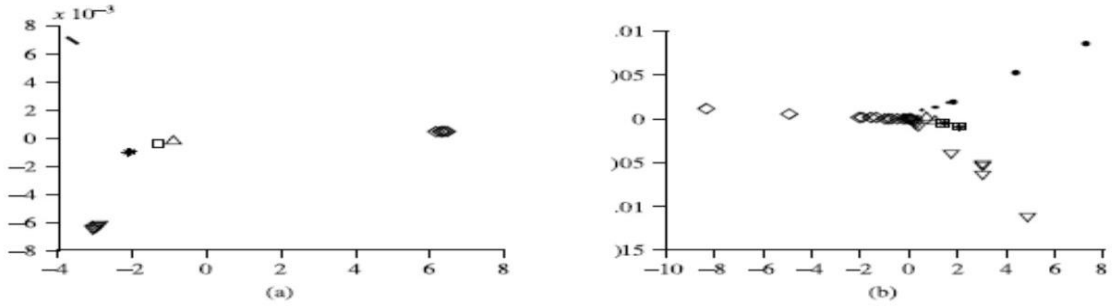
$$Z_{ij} = \begin{cases} 1/s_i, & \text{if there is a link from block } i \text{ to page } j \\ 0 & \text{otherwise.} \end{cases}$$

where a is a normalization factor to make the sum of fp(b) to be 1, that is,

Automatic Classification of Web Documents

In the automatic classification of Web documents, each document is assigned a class label from a set of predefined topic categories, based on a set of examples of pre classified documents. Keyword-based document classification can be used for Web document classification. Such a term-based classification scheme has shown good results in Web page classification.

There have been extensive research activities on the construction and use of the semantic



2-D embedding of the WWW images. (a) The image graph is constructed using block-level link analysis. Each color (shape) represents a semantic category. Clearly, they are well separated. (b) The image graph was constructed based on traditional perspective that the hyperlinks are considered from pages to pages. The image graph was induced from the page-to-page and page-to-image relationships.

Web, a Web information infrastructure that is expected to bring structure to the Web based on the semantic meaning of the contents of Web pages. Web document classification by Web mining will help in the automatic extraction of the semantic meaning of Web pages and build up ontology for the semantic Web.

Web Usage Mining

Besides mining Web contents and Web linkage structures, another important task for Web mining is Web usage mining, which mines Weblog records to discover user access patterns of Web pages.

A Web server usually registers a (Web) log entry, or Weblog entry, for every access of a Web page. It includes the URL requested, the IP address from which the request originated, and a timestamp.

In developing techniques for Web usage mining, we may consider the following.

First, Although it is encouraging and exciting to imagine the various potential applications of Weblog file analysis, it is important to know that the success of such applications depends on what and how much valid and reliable knowledge can be discovered from the large raw log data. Often, raw Weblog data need to be cleaned, condensed, and transformed in order to retrieve and analyze significant and useful information.

Second, with the available URL, time, IP address, and Web page content information, a multidimensional view can be constructed on the Weblog database, and multidimensional OLAP analysis can be performed to find the top N users, top N accessed Web pages, most frequently accessed time periods, and so on, which will help discover potential customers, users, markets, and others.

Third, data mining can be performed on Web log records to find association patterns, sequential patterns, and trends of Web accessing.

POSSIBLE QUESTIONS:

2 Marks:

1) What is clustering?
2) List the various clustering methods.
3) Diffrenciate k-means and k-mediods algorithms.
4) Define Agglomerative clustering techniques
5) Give example for divisive clustering techniques
6) What are issues in spatial data mining?
7) Give the formula for precision and recall in text mining.
8) What are the uses of web mining.
9) What is multimedia mining?

Part- B

1. Describe k-means clustering with an Example.
2. Explain hierarchical methods of clustering.
3. Discuss the different types of clustering methods
4. Explain the various methods for detecting outliers.
5. Explain in detail about Categorization of major clustering methods
6. Explain in detail about Partitioning methods with example
7. Explain the following clustering methods in detail (1) BIRCH (2) DENCLUE
8. Write a short note on the following:
    1) Text Mining
    2) Multimedia Mining

Part-C 10 marks:

1. Explain in detail about the following clustering methods:
    1) Density-based Methods
    2) Grid-based Methods
    3) Model-based Methods
    4) Clustering High-Dimensional Data
2. What is web mining? Discuss the various web mining techniques.

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University Established Under Section 3 of UGC Act 1956)

Pollachi Main Road, Eachanari Post, Coimbatore – 641 021. INDIA

DEPARTMENT OF COMPUTER SCIENCE

DATA MINING AND WAREHOUSING (17CSP103)

ONLINE EXAM QUESTION BANK

UNIT-1

Questions	opt1	opt2	opt3	opt4	Answer
_____ refers to extracting knowledge from large amounts of data.	data extraction	data mining	data warehousing	data integration	data mining
_____ used to organize attributes or attribute values into different levels of abstraction.	data hierarchies	model hierarchies	concept hierarchies	pattern hierarchies	concept hierarchies
_____ process is used to remove noise and inconsistent data	data cleaning	data integration	data selection	data transformation	data cleaning
Which process multiple data sources may be combined	data cleaning	data integration	data selection	data transformation	data integration
_____ process is used to retrieve the relevant data from the database	data cleaning	data integration	data selection	data transformation	data selection
_____ process is used to identify the truly interesting patterns representing knowledge based on some interestingness measures	data cleaning	data integration	data selection	pattern evaluation	pattern evaluation
_____ database consists of a file where each record represents a transaction	Relational	Transactional	Object Oriented	Spatial	Transactional
A set of _____ that describe the objects. These corresponds to attributes in the entity relationship and relational models	messages	methods	variables	Spatial	variables
A set of _____ helps the objects communicate with other objects	messages	methods	variables	functions	Messages
_____ database contains spatial related information	Relational	Transactional	Object Oriented	Spatial	Spatial
_____ database typically stores relational data that include time-related attributes	Temporal	Transactional	Object Oriented	Spatial	Temporal

A _____ database consists of a set of interconnected autonomous component	Temporal	Heterogeneous	Spatial	Object Oriented	Heterogeneous
Many applications involve the generation and analysis of a new kind of data called	Stream	multimedia	Object Oriented	Transactional	Stream
Capturing user access patterns in such distributed information environment is called _____	Web usage Mining	multimedia	Object Oriented	Transactional	Web usage Mining
_____ is the process of finding a model that describes and distinguishes data classes	Prediction	Classification	Analysis	Clustering	Classification
_____ is the statistical methodology that is most often used for numeric prediction	Prediction	Classification	Regression	Clustering	Regression
The data objects that do not comply with the general behavior or model of the data is called	Evolution	Classification	Regression	Outlier	Outlier
In which integration scheme Data mining system will not utilize any function of a Data base or Data warehouse.	No Coupling	Tight coupling	loose coupling	Semi tight coupling	No Coupling
In which integration scheme Data mining system will use some facilities of a Data base or Data warehousehouse	No Coupling	Tight coupling	loose coupling	Semi tight coupling	loose coupling
In which integration scheme Data mining system is smoothly integrated into the DB/DW system	No Coupling	Tight coupling	loose coupling	Semi tight coupling	Tight coupling
_____ is the task of discovering interesting patterns from large amounts of data	database	data warehouse	data mining	data clustering	data mining
_____ is a repository for long term storage of data from multiple sources, organized so as to facilitate management decision makingcilitate management decision making	database	data warehouse	data mining	data clustering	data warehouse
A _____ is a collection of tables each of which is assigned a unique name	Relational	Transactional	Object Oriented	Spatial	Relational

A _____ is a measure that can be computer for a given dtasey by partitioning the data into smaller subsets,computing the measure for each subset and then merging the results in order toarrive at the measures value for the original data set	Predictio n	Distributi ve measures	algebraic measures	None of the above	Distributive measures
An _____ is a measure that can be computed by applying an algebraic function to one or more distributive measuresore distributive measures	Predictio n	Distributi ve measures	algebraic measures	holistic	algebraic measures
A _____ is a measure that must be computed on the entire data set as a whole	holistic	Distributi ve	algebraic	Predictio n	holistic
The _____ of the set is the difference between the largest and smallest values	range	median	mode	sum	range
The _____ of a distribution consists of the minimum value, Q1, median, Q3,maximum	five-number sumary mary	mode	range	average	five-number sumary mary
In Box plots median is marked by _____ with in the box	circle	line	double line	line chart	line
The distance between the first and third quartile is called as _____	Median	mode	InterQuar tileRange	range	InterQuartileRange
The most commonly used percentiles other than median are _____	Quartiles	mode	range	sum	Quartiles
_____ are a popular way of visualizing a distribution	Quartiles	boxplots	IQR	Outlier	boxplots
_____ is a random error or variance in a measured variable	smooth	binning	Regressi on	Noise	Noise
_____ works to remove noise from the data	smoothin g	aggregati on	generaliz ation	normaliz ation	smoothing
_____ operation is used for summarization	smoothin g	aggregati on	generaliz ation	normaliz ation	aggregation
_____ is used to scale attribute data so as to fall with in a small specified range	smoothin g	aggregati on	generaliz ation	normaliz ation	normalization

_____ performs a linear transformation on the original data	smoothing	aggregation	min-max normalization n	binning	min-max normalization n
_____ databases contain word description for objects	Relational	Transactional	Object Oriented	Text	Text
Maps can be represented in _____format in spatial database	lines	circles	vector	scalar	vector
_____ database store image ,audio and video data	Relational	Transactional	multimedia	Object Oriented	multimedia
_____ is a comparison of the general features of target class data objects	characterization	Classification	discrimination	None of the above	discrimination
A _____is a flow chart like tree structure where each node denotes a test on an attribute valuee	neural network	regression	k means	decision tree	decision tree
A _____ when used for classification resembled neural like processing	neural network	regression	k means	decision tree	neural network
_____ represents the percentage of transaction from a transactional database that the given rule satisfies in a association rulen rule satisfies in a associaiton rule	confidence	support	completeness	None of the above	support
_____ routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the datand correct inconsistencies in the data	data cleaning	data integration	data selection	data transformation	data cleaning
_____ techniques such as data cube aggregation attribute subset selection, dimensionality reduction numerosity reduction and discretizationeduction numerosity reduction and discretization	data integration	data selection	data transformation	data reduction	data reduction

KDD stands for	Knowledge discovery in database Database	knowledge database iscovery	knowledge distributee database	None of the above	knowledge distributee database
Histograms partition the values for an attribute into disjoint ranges called _____	Relational	buckets	interval	frequency	buckets
In which reduction data are replaced or estimated by alternative, smaller data representation	numerosity	dimensionality	aggregation	None of the above	numerosity
In which reduction encoding mechanism are used to reduce the data set size	numerosity	dimensionality	aggregation	None of the above	dimensionality
If the original data can be reconstructed from the compressed data with out loss of any information is called	lossy	lossless	discrimination	None of the above	lossless
If the original data can be reconstructed from the compressed data with loss of any information is called	lossy	lossless	discrimination	aggregation	lossy
DWT stands for _____	Discrete Wavelet Transform	Direct Wavelet transform	Dynamic Weather Transform	linear regression	Discrete Wavelet Transform
PCA stands for _____	Principal Components Analysis	Primary Complex Analysis	Principal Complex Analysis	lossless	Principal Components Analysis
In _____ the data are modeled to fit a straight line	linear regression	multiple regression	log-linear regression	Discrete Wavelet Transform	linear regression
_____ is an extension of linear regression which allows a response variable	linear regression	multiple regression	log-linear regression	none	multiple regression

KARPAGAM ACADEMY OF HIGHER EDUCATION

emed to be University Established Under Section 3 of UGC Act 1956)

Pollachi Main Road, Eachanari Post, Coimbatore – 641 021. INDIA

DEPARTMENT OF COMPUTER SCIENCE

DATA MINING AND WAREHOUSING (17CSP103)

ONLINE EXAM QUESTION BANK

UNIT-2

Questions	opt1	opt2	opt3	opt4	Answer
_____ approximates discrete multidimensional probability distributions	linear regression	multiple regression	log-linear regression	aggregation	log-linear regression
In _____ the width of each bucket range is uniform	Equal-width	Equal – frequency	V-optimal	MaxDiff	Equal-width
In _____ histogram the bucket created so that roughly the frequency of each bucket is constant	Equal-width	Equal – frequency	V-optimal	MaxDiff	Equal – frequency
_____ histogram is the one with the least variance	Equal-width	Equal – frequency	V-optimal	MaxDiff	V-optimal
In _____ histogram the difference between each pair of adjacent values	Equal-width	Equal – frequency	V-optimal	MaxDiff	MaxDiff
_____ can be used as a data reduction technique	Equal-width	Equal – frequency	V-optimal	sample	sample
SRSWOR stands for	Simple Random sample without replacement	Simple Random Sample with replacement	Sort Random simple with replacement	Sort Rambus simple with replacement	Simple Random sample without replacement
SRSWR stands for	Simple Random sample without replacement	Simple Random Sample with replacement	Sort Random simple with replacement	Sort Rambus simple with replacement	Simple Random Sample with replacement
In stratified sample D is divided into mutually disjoint parts called _____	Dissimilar matrix	cluster	consequents	Strata	Strata

A _____ is a collection of data objects that are similar to one another	cluster	outlier	predictio n	Data Matrix	cluster
_____ represent object by object structure	Data Matrix	Dissimila r matrix	Singular matrix	outlier	Dissimilar matrix
_____ represent object by variable structure	Data Matrix	Dissimila r matrix	Singular matrix	cluster	Data Matrix
Data Matrix is often called _____	one mode	two mode	three mode	four mode	two mode
Dissimilarity matrix is often called _____	one mode	two mode	three mode	four mode	one mode
Manhattan distance is also called as_____	Euclidea n distance	Minkows ki distance	city block distance	conseque nts	city block distance
_____ is a generalization of both Euclidean distance and Manhattan distance	Euclidea n distance	Minkows ki distance	city block distance	cluster	Minkowski distance
Association rule is also called as _____	Market basket analysis	Euclidea n distance	Minkows ki distance	city block distance	Market basket analysis
In Association rule X-à Y , X represents	Antecede nts	conseque nts	confiden ce	predictab ility	Antecedents
In Association rule X-à Y , Y represents	Antecede nts	conseque nts	confiden ce	predictab ility	consequents
A frequent item set Y is _____	minimal	maximal	predictio n	Antecede nts	maximal
Expansion of DHP is _____	Direct Hashing and Pruning	Dynamic Hash Process	Decision hashing and pruning	Decision Hash Process	Direct Hashing and Pruning
Expansion of DIC is_____	Direct Hashing and Pruning	Dynamic item set Counting	Decision item and counting	Direct item set Counting	Dynamic item set Counting
_____algorithm does not generate candidate itemset	Apriori algorithm	Direct hashing and pruning	FP-Growth	None	FP-Growth
A _____ has only two states 0 or 1	Trinary variable	Binary variable	Unary Variable	none	Binary variable
A binary variable is _____ if the outcomes of the states are not equally important	symmetri c	asymmet ric	different	none	asymmetric

A binary variable is _____ if both of its states are equally valuable and carry the same weight	symmetric	asymmetric	different	none	symmetric
The sim(i,j) is called _____ coefficient	symmetric	asymmetric	Jaccard	differential	Jaccard
A _____ variable is a generalization of the binary variable in that it can take on more than two states	symmetric	asymmetric	Categorical	Ordinal	Categorical
A _____ variable resembles a categorical variable except that the M states of the ordinal value are ordered in a meaningful sequences.	symmetric	asymmetric	Categorical	Ordinal	Ordinal
A _____ variable makes a positive measurement on a nonlinear scale	Ratio Scaled	asymmetric	Categorical	Ordinal	Ratio Scaled
Which algorithm comes under portioning methods	K-mediods	BIRCH	agglomerative	CLARS	K-mediods
Which algorithm comes under portioning methods.	K-means	BIRCH	agglomerative	DBSCAN	agglomerative
Which algorithm comes under Hierarchical methods	K-means	K-mediods	BIRCH	DBSCAN	BIRCH
Which algorithm comes under Density based methods	K-means	K-mediods	agglomerative	DBSCAN	DBSCAN
_____ method quantize the object space into a finite number of cells that form a grid structure	Partitioning	Hierarchical	Density based	Grid-based	Grid-based
_____ method hypothesize a model for each of the clusters ad find the best fit of the data to the given model	Partitioning	Hierarchical	Density based	Model –based	Model –based
_____ is a typical example of grid based method	Partitioning	STING	Density based	Model –based	STING
SOM stands for	Self-organizing feature map	Simple organizing map	Self oriented map	None	Self-organizing feature map
_____ is a clustering approach that performs clustering by incorporation of user specified or application oriented constrained	Partitioning	Hierarchical	Density based	Constrained based clustering	Constrained based clustering

The divisive approach also called _____ approach	top-down	bottom up	partition	none	top-down
Attribute selection measures are also known as _____	partition rule	decision rule	splitting rule	Partitioning	splitting rule
____ uses information gain as its attribute selection measure	partition rule	decision rule	splitting rule	ID3	ID3
A _____ is a heuristic for selecting the splitting criterion that best separates a given data partition	partition rule	decision rule	splitting rule	attribute selection measure	attribute selection measure
The cost complexity pruning algorithm used in _____	CART	ID3	splitting rule	none	CART
DSS stands for _____	Direct Support System	Decision Support System	Dynamic Support system	none	Direct Support System
Data Content refers to _____ operational systems	current values	archived values	derived values	summarized values	current values
_____ stage was an attempt by IT to anticipate somewhat the types of reports that world be requested form time to time.	Decision Support system	information system	Special Extract Programs	Partitioning	Special Extract Programs
In which stage companies began to build more sophisticated systems intended to proved strategic information.	Decision Support system	Information system	Special Extract Programs	none	Decision Support system
Operational systems are _____ systems	Decision Support system	information system	Special Extract Programs	OLAP	OLAP
_____ stage was an attempt by IT to anticipate somewhat the types of reports that world be requested form time to time.	Decision Support system	information system	Special Extract Programs	none	Special Extract Programs
In which stage companies began to build more sophisticated systems intended to proved strategic information.	Decision Support system	information system	Special Extract Programs	none	Decision Support system
Operational systems are _____ systems	Decision Support system	information system	Special Extract Programs	OLAP	OLAP
Database designed for _____ task	Historical	analytical	application	Partitioning	analytical

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University Established Under Section 3 of UGC Act 1956)

Pollachi Main Road, Eachanari Post, Coimbatore – 641 021. INDIA

DEPARTMENT OF COMPUTER SCIENCE

DATA MINING AND WAREHOUSING (17CSP103)

ONLINE EXAM QUESTION BANK

UNIT-3

Questions	opt1	opt2	opt3	opt4	Answer
In Which system the data is stored based on individual applications	Subject oriented data	Integrated data	Time Stamped data	Nonvolatile data	Subject oriented data
In which system the stored data contains the current values	Subject oriented data	Integrated data	Time-Variant data	Nonvolatile data	Time-Variant data
In which system data is pull together from the various applications	Subject oriented data	Integrated data	Time Stamped data	Nonvolatile data	Integrated data
_____ refers to the level of details	Subject oriented data	Integrated data	Time Stamped data	Data granularity	Data granularity
_____ data comes from the various operational systems of the enterprise.	production	Internal	Archived	External	production
_____ data is used to keep the private details of the users in every Organization	production	Internal	Archived	External	Internal
_____ data is mostly dependent on the external resources for a high percentage of information.	production	Internal	Archived	External	External
_____ function is used to extract information from other data sources	Data Extraction	Data Transformation	Data loading	Archived	Data Extraction
In expert systems the knowledge of domain is represented in terms of _____.	If-then rules	Planners	Samples	Charts	If-then rules
_____ is used exclusively for the discovery stage of the KDD process.	OLAP	Cleaning	Enriching	Data mining	Data mining
KDD is a ----------_____.	new technique	statistical technique	multi disciplinary field	database technology	multi disciplinary field

If you know exactly what you are looking for the use _____.	genetic algorithm	SQL	neural network	clustering	SQL
Access type in operational system is _____	read	update,delete	delete, read	read, update,delete	read, update,delete
Access type in operational system is _____	read	update	delete	update, delete	read
Data mart is _____	departmental	corporate	organized	neural network	departmental
Data warehouse is _____	departmental	corporate	organized	neural network	corporate
_____ data comes from various operational system	Internal	production	archived	external	production
_____ data contains the user profile	Internal	production	archived	external	Internal
_____ in a data warehouse is similar to the data dictionary	internal	production	metadata	external	metadata
_____ of the dimension table uniquely identifies each row	primary key	foreign key	unique	internal key	primary key
A dimension table has _____ attributes	one	two	three	many	many
Each dimension table is in a one to many relationship with the _____ table	primary	secondary	fact	none	fact
The _____ schema is used for data design is a relational model consisting of fact and dimension tables.	STAR	Cube	Dimension	none	STAR
Analytical capabilities of OLTP system is _____	low	moderate	very low	small	very low
Analytical capabilities of Data Warehouse system is _____	low	moderate	very low	small	moderate
Data for a single session in OLTP system is _____	Large	moderate	very limited	small	very limited
Data for a single session in Data Warehouse system is _____	Large	moderate	very limited	small to medium size	small to medium size
Response time in OLTP system is	Fast to moderate	Large	Small	Very Fast	Very Fast
Response time in Data warehouse system is	Fast to moderate	Large	Small	Very Fast	Fast to moderate

Data Granularity in OLTP system is	Large	Small	Very Fast	Detail	Detail
Data Granularity in Data Warehouse system is	Large	Small	Detail and summary	Detail	Detail and summary
Access Method in OLTP System is	defined	predefined and adhoc	predefined	current	predefined
Access Method in Data Warehouse System is	defined	predefined and adhoc	predefined	current	predefined and adhoc
Data currency in OLTP system is	current	old	new	current and historical	current
Data currency in Data Warehouse system is	current	old	new	current and historical	current and historical
OLAP means	OnLine Analytical Processing	Online Algebric Processing	Overall Analytical Purpose	OnLine Analysis Processing	OnLine Analytical Processing
---------- provides support for concurrent data access, data integrity, and access security	Transperancy	Multi user	Client / server Architecture	Flexible reporting	Multi user
The representation that accommodates more than three dimension is ----------	Cube	Hyper Cube	Meta Date	ROLAP	Hyper Cube
The technique used to view only the Higher Level of aggregation is ----------	drill down	rolling up	hierarchy	aggregation	rolling up
The technique used to view the Lower Level details ----------	drill down	rolling up	hierarchy	aggregation	drill down
The _____ schema has three business dimensions, namely product , store and time	star	cube	diamond	none	star
The _____ algorithm for portioning where each clusters center is represented by the mean value of the objects in the cluster	k-means	k-mediods	star	none	k-means

PAM refers to	Partitioning Around Mediods	Partial around mediod	Partitioning Against Mediods	None	Partitioning Around Mediods
Expansion of CLARA	Centre LARge Application	Cluster LARge Application	LARge Application	none	Cluster LARge Application
_____ algorithm combines sampling text with PAM	Cluster	K-mediods	K-means	CLARANS	CLARANS
AGNES stands for	AGlomerative NESting	Against Nesting	All Network	None	AGlomerative NESting
A tree structure called a _____ is commonly used to represent the process of hierarchical clustering	Distance	Mean	Dendrogram	None	Dendrogram
To measure the distance between clusters it is sometimes called a _____	single linkage algorithm	minimal spanning algorithm	nearest-neighbor clustering algorithm	farthest neighbor clustering algorithm	nearest-neighbor clustering algorithm
If the clustering process sis terminated when the distance between nearest clusters exceeds an arbitrary threshold it is called a _____	single linkage algorithm	minimal spanning algorithm	nearest-neighbor clustering algorithm	farthest neighbor clustering algorithm	single linkage algorithm
In agglomerative hierarchical clustering algorithm that uses the minimum distance measure is also called a	single linkage algorithm	minimal spanning algorithm	nearest-neighbor clustering algorithm	farthest neighbor clustering algorithm	minimal spanning algorithm
When an algorithm uses the maximum distance to measure the distance between clusters it is sometimes called a	single linkage algorithm	minimal spanning algorithm	nearest-neighbor clustering algorithm	**farthest neighbor clustering algorithm**	farthest neighbor clustering algorithm
If the clustering process is terminated when the maximum distance between nearest clusters exceeds an arbitrary threshold it is called a	single linkage algorithm	minimal spanning algorithm	nearest-neighbor clustering algorithm	**Complete – linkage algorithm**	Complete – linkage algorithm

The _____ between two clusters is defined as the absolute interconnectivity	Relative interconnectivity	Relative Closeness	Edge cut	single linkage algorithm	Relative interconnectivity
A _____ cluster is a set of density connected objects that is maximal with respect to density reach ability	Relative interconnectivity	Relative Closeness	Density-based	None	Density-based
_____ is a popular and simple method of incremental conceptual clustering	COBWEB	COOL WEB	web	minimal spanning algorithm	COBWEB
A _____ clusters objects based on the notion of density	hierarchical method	gird based method	density based method	partitioning method	density based method
OLTP data base size is	100 MB to GB	100 GB to TB	millions	hundreds	100 MB to GB
OLAP data base size is	100 MB to GB	100 GB to TB	millions	hundreds	100 GB to TB
Number of users in OLTP	tens	hundreds	millions	thousands	thousands
Number of users in OLAP	tens	hundreds	millions	thousands	hundreds
A _____ is a subject oriented, integrated , time-variant, and non volatile collection	Spatial data warehouse	Temporal warehouse	Web mining	Text mining	Spatial data warehouse
There are _____ types of spatial data cube	one	two	three	four	three
A _____ contains only numeric data	spatial measure	numeric measure	text measure	none	numeric measure
A _____ contains a collection of pointers to spatial objects	spatial measure	numeric measure	text measure	none	spatial measure

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University Established Under Section 3 of UGC Act 195

Pollachi Main Road, Eachanari Post, Coimbatore – 641 021. INDIA

DEPARTMENT OF COMPUTER SCIENCE

DATA MINING AND WAREHOUSING (17CSP103)

ONLINE EXAM QUESTION BANK

UNIT-4

Questions	opt1	opt2	opt3	opt4
_____ is the percentage of retrieved documents that re in fact relevant to the query	Precision	Recall	Retrieval	information
A _____ database stores and manages a large collection of multimedia data	spatial	temporal	web	multimedia
The object oriented data model inherits the essential concepts of _____ database	Multimedia	Object oriented	Spatial	web
_____ gives the relationship and patterns between data elements	KDD	data	objects	web
_____ is used to describe the whole process of extraction of knowledge from data	data	data mining	KDD	algorithm
_____ need the control of human operation during their execution	noise	data mining	supervised algorithm	spatial
_____ takes all the data at once and tries to create hyposthesis based on the data	visualization	data mining	supervised algorithm	batch learning algorithm
_____ is the random disturbance of a transmitted signal in the context of KDD	noise	data mining	objects	batch learning algorithm
_____ is used at thebegining of datamining process to get a good quality of dataset	gentic	Web mining	visualization	spatial
Important elements in the cleaning operation is the _____	k-nearest	inconsistency	deduplicaiton of records	none
Including records that are closed to each other or living each other's neighbour hood is called as _____	genetic	neural networks	k-nearest neighbour	visualizaiton
_____ was modeled based on human brain	genetic	neural networks	k-nearest neighbour	visualizaiton

knowledge discovery process consists of _____ stages	six	three	four	five
_____ refers to elements of the message that can be derived from other part of the other message	machine language	neural	k-nearest	redundancy
_____ program must be unable to read it by the users	nerual networks	genetic algorihtm	machine language	spatial
_____ is not realistic for most learning situations	data mining	neural networks	background knowledge	spatial
_____ are always defined on binary attributes	association rules	genetic algorithm	neural networks	KDD
KDD stands for _____	Knowledge discovery in database	Knowledge development in data	knowledge data	Knowledge development
_____ is a interaction between technology and nature	genetic algorithm	neural networks	k-nearest	machine language
_____ contains the visual map of data sets	kohonen self organizing map	neural networks	k-nearest	spatial
A _____ network not only has input and outputnodes but also hidden nodes	gentic algorithm	neural network	back propogation	k-nearest
In _____ information can be easily retrived from the database using query tools	Mulit dimeniso nal knowledge	shallow knowledge	hidden knowledge	deep knowledge
In _____ information that can be analyzed using online analytical processing tools	Mulit dimeniso nal knowledge	shallow knowledge	hidden knowledge	deep knowledge
In _____ data can be found relatively easily by using pattern recognition or machine learning algorithm	Mulit dimeniso nal knowledge	shallow knowledge	hidden knowledge	deep knowledge

In _____ information that is stored in the database but can only be located if we have a clue that tells us where to look	Mulit dimenisonal knowledge	shallow knowledge	hidden knowledge	deep knowledge
A _____ consists of a simple three-layered network	responders	photo receptors	perceptrons	none
In perceptron input unit is called	responders	photo receptors	perceptrons	none
In perceptron intermediate units called	responders	photo receptors	perceptrons	associators
In perceptron network output unit is called	Responders	photo receptors	perceptrons	none
OLAP stands for	Online arithmetic processing	Online Analytical Processing	Operation Analytical Processing	none
The _____ is very useful in a data mining context	responders	photo receptors	perceptrons	space-metaphor
In which process pollution is detected	data cleaning	data selection	data enrichment	data coding
In which process data are collected from the operation database	data cleaning	data selection	data enrichment	data coding
In which process additional information are included in the existing database	data cleaning	data selection	data enrichment	data coding
_____ is a creative process	data cleaning	data selection	data enrichment	data coding
Learning tasks can be divided into _____ areas	None	two	three	four
A _____ contains historic data and is subject oriented and static	cleaning	data warehousing	database	operational data
Creative coding is the heart of _____ process	data	data mining	KDD	algorithm

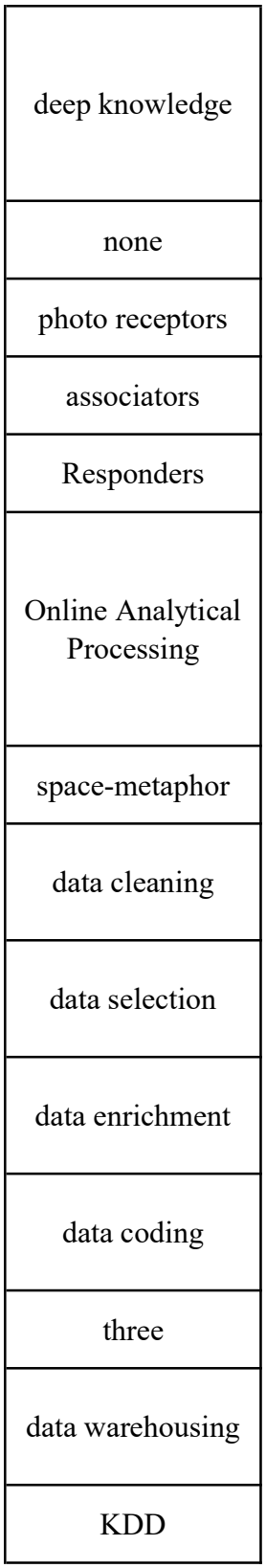
_____ algorithm comes under classification tasks	neural networks	association rules	inductive logic programming	genetic algorithm
_____ algorithm comes under problem solving tasks	neural networks	association rules	inductive logic programming	genetic algorithm
_____ algorithm come under knowledge engineering tasks	neural networks	association rules	inductive logic programming	genetic algorithm
The visual map o given sets and kohenen sel organizing map is a collection of _____	neurons and units	data base	KDD	none
_____ tool is not a single technique it holds variety of different techniques	neurons and units	data base	Data Mining	none
Which is not present in the six stages of KDD	enrichment	coding	selection	replying
A very important element in a cleanning operation is that	de duplicaiton of data	replicaiton	inconsistency	none
At every state of the KDD process the data miner can step back	one phase	two phases	one or more phase	only three
_____ starts with number of observations	analysis	observation	genetic	none
_____ try to find the patterns in the observations	analysis	observation	genetic	none
The theory which give new phenomena that can be verified by new observations is known as	observations	analysis	theory	predection
_____ is a form of adaptation	learning	discovery	finding	none
Geological sample could be represented in terms of _____ rules	do while	for	if-then	none
The _____ revolution gives the individual knowledge worker access to central information systems	client/server	machine learning	SQL	none

6)

Answer

Precision
multimedia
Object oriented
KDD
data mining
supervised algorithm
batch learning algorithm
noise
visualization
deduplicaiton of records
k-nearest neighbour
neural networks

six
redundancy
machine language
background knowledge
association rules
Knowledge discovery in database
genetic algorithm
kohonen self organizing map
back propogation
shallow knowledge
Mulit dimenisonal knowledge
hidden knowledge


deep knowledge
none
photo receptors
associators
Responders
Online Analytical Processing
space-metaphor
data cleaning
data selection
data enrichment
data coding
three
data warehousing
KDD

neural networks
genetic algorithm
inductive logic programming
neurons and units
Data Mining
replying
de duplicaiton of data
one phase
observation
analysis
predection
learning
if-then
client/server

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University Established Under Section 3 of UGC Act 1956)

Pollachi Main Road, Eachanari Post, Coimbatore – 641 021. INDIA

DEPARTMENT OF COMPUTER SCIENCE

DATA MINING AND WAREHOUSING (17CSP103)

ONLINE EXAM QUESTION BANK

UNIT-5

Questions	opt1	opt2	opt3	opt4
A nonspatial dimension contains only_____	spatical data	dimensional data	nonspatial data	numerical data
A spatial measure contains a collection of _____ to spatial objects.	dimensions	objects	pointers	class
If we have found some regularities we formulate _____ explaining the data	observation	theory	prediction	none
_____ data warehouse provide an accurate and consistent view of enterprise information	purchase	enterpirse	analysis	sales
The _____ should have identified the initial user requirements.	technical	warehouse	buisness requirements	process
_____ is used to determine what the overall architecture of the data warehouse should be	requirements	architecture	process	technical blueprint
Data warehouse must be architected to support _____ major driving factors	3	4	2	5
_____ takes data from source systems and makes it available to the data warehouse	data cleaning	data abstraction	data extraction	none
clean and _____ the loaded data into a structure that speeds up queries	transform	extract	partition	aggregate
Make sure data is_____ within itself	inconsistent	replicaiton	consistent	none
The _____ process is the system process that manages the queries and speeds them up by directing queries to the most effective data source	process management	load management	system management	query management
The _____ is the system component that performs all the operations necessary to support the extract and load process	process manager	load manager	system manager	query manager
The bulk of the effort to develop a load manager should be planned within the fir _____ phase	analysis	design	load	production

Meta data describes the _____.	type and format of data	structure of the contents of database	location of data	owner of data
In bottom up approach _____ are used by end-users	data mart	large data warehouse	central database	operational database
_____ are used to load the information from the operational database	Statistical Techniques	Visualization Techniques	Windowing mechanism	Replication Techniques
_____ responses end-users queries in a very short space of time.	Client / Server Technique	Time sharing system	Multiprocessing computer system	Real time system
Expert system contain _____.	spatial data	knowledge of specialists	knowledge of business logic	transaction records
OLAP store their data in _____.	table format	object oriented format	a special multi-dimensional format	points in multi-dimensional space
OLAP tools	do not learn but create new knowledge	do learn but cannot create new knowledge	do learn and also can create new knowledge	do not learn and cannot create new knowledge
Data mining algorithm should not have a complexity that is higher than____.	logn	n^2	nlogn	n
Association rules are defined on _____.	single attribute	n attributes	binary attributes	none of the above
A perceptrons consists of _____.	two layered networks	single layered networks	three layered networks	multiple layered networks

Back propagation method gives _____.	answers and idea as to how they arrived at the answers	answers and no clear idea as to how they arrived at the answers	only ideas as to how to obtain answers	answers and poor ideas as to how they arrived at the answers
A Kohonen's self-organizing map is a collection of _____.	networks	neurons	processors	none of the above
Genetic algorithms can be viewed as a kind of _____ strategy.	self learning	meta learning	knowledge discovering	concept learning
Voroni diagram divide a sample space into _____.	different groups	different divisions	different sub networks	different regions
Neural networks are somewhat better at _____.	problem solving tasks	classification tasks	knowledge engineering	none of the above
SQL retrieves _____.	hidden knowledge	hidden rules	shallow knowledge	deep knowledge
_____ can be found by pattern recognition algorithms	Hidden knowledge	Deep knowledge	Encrypted information	Fine grained segmentation
_____ give a yes or no answer and no explanation of their responses	Genetic algorithm	Neural network	k-nearest neighbor algorithm	Neural network and k-nearest neighbor algorithm

The reporting stage combines _____.	analysis of the results & application of the result to new data	the results & application of the result to new data	analysis of the results & application of the result to existing data	a. analysis of the results & application of the result to new rules
The delivery process is staged in order to _____.	reduce the execution time	to minimize error	minimize risk	measure benefits
Delivery process is designed to deliver _____.	an enterprise data warehouse	a point solution	maximum information	quality solution
Delivery process ensures _____.	to reduce the overall delivery time-slice	benefits are delivered incrementally	to reduce the overall delivery time-slice & benefits are delivered incrementally	to reduce investment
The technical blueprint phase must deliver an overall architecture that satisfies the _____ requirements.	short term	long term	today's	none of the above
----------_____ is part of day-to-day management of the data warehouse.	Creating / deleting summaries	Extracting data	Cleaning data	Populating data
The data extracted from the source systems is loaded into a _____.	data warehouse	data mart	temporary data store	operational database
The purpose of business case is to identify the projected _____.	out put structure	business process	business benefits	business process risks

In order to deliver an early release of part of a data warehouse we should _____.	focus on the business requirements	focus on long term requirements	technical blueprint phases	focus on the business requirements & technical blueprint phases
The technical blueprint must deliver _____ that satisfies the long-term requirements.	overall system architecture	overall benefits	overall expenditure	overall process strategy
In _____ methods, the query is regarded as specifying constraints for selecting relevant documents.	file selection	pointer selection	query selection	document selection
_____ methods use the query to rank all documents in the order of relevance.	model ranking	document ranking	file ranking	description ranking

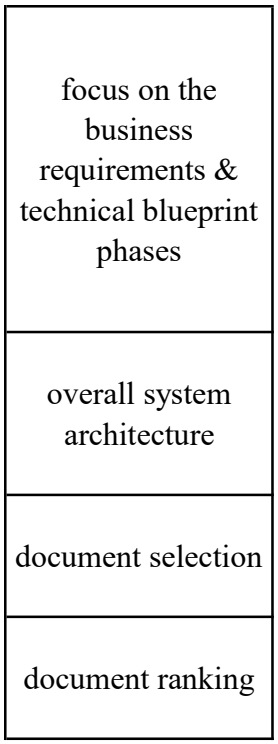
Answer

nonspatial data
pointers
theory
enterpirse
buisness requirements
technical blueprint
3
data extraction
transform
consistent
query management
load manager
production

structure of the contents of database
data mart
Replication Techniques
Multiprocessing computer system
knowledge of specialists
a special multi-dimensional format
do not learn and cannot create new knowledge
nlogn
binary attributes
three layered networks

answers and no clear idea as to how they arrived at the answers
neurons
meta learning
different regions
classification tasks
shallow knowledge
Hidden knowledge
Neural network and k-nearest neighbor algorithm

analysis of the results & application of the result to new data
minimize risk
an enterprise data warehouse
to reduce the overall delivery time-slice & benefits are delivered incrementally
long term
Creating / deleting summaries
temporary data store
business benefits


focus on the business requirements & technical blueprint phases
overall system architecture
document selection
document ranking

Reg. No.....

[09CSP201]

KARPAGAM UNIVERSITY

(Under Section 3 of UGC Act 1956)
COIMBATORE – 641 021
(For the candidates admitted from 2009 onwards)

M.Sc. DEGREE EXAMINATION, APRIL 2010

Second Semester

COMPUTER SCIENCE

DATA MINING AND WAREHOUSING

Time: 3 hours

Maximum : 60 marks

PART – A (20X ½ = 10 Marks) (30 Minutes)
(Question Nos. 1 to 20 Online Examinations)

PART B (5 X 4= 20 Marks) (2 ½ Hours)
Answer ALL the Questions

21.a. Explain in detail about Data warehouse Back-end Tools and Utilities.

(Or)

b. List out the major issues in Data Mining.

22.a. Explain about data integration.

(Or)

b. List out the strategies for data reduction.

23.a. List out the constraints for constraint – based Association mining.

(Or)

b. Explain about the Generating Association Rules from frequent Itemsets.

24.a. List out the criteria for comparing the classification and prediction method.

(Or)

b. Explain Gain ratio and Tree Pruning.

25.a. List out the requirements of clustering in data mining.

(Or)

b. Explain in detail about an Agglomerative and Divisive Hierarchical Clustering.

PART C (3 x 10 = 30 Marks)
Answer any THREE Questions

26. List out the classification of Data Mining Systems and explain each with neat diagram.

27. Write short note on Data Reduction.

28. Write the classification of frequent pattern mining.

29. Write short note on Rule-Based classification.

30. Explain about an Outlier Analysis.

Reg. No.....

[10CSP201]

KARPAGAM UNIVERSITY

(Under Section 3 of UGC Act 1956)
COIMBATORE – 641 021
(For the candidates admitted from 2010 onwards)

M.Sc. DEGREE EXAMINATION, APRIL 2011

Second Semester

COMPUTER SCIENCE

DATA MINING AND WAREHOUSING

Time: 3 hours

Maximum : 60 marks

PART – A (20X ½ = 10 Marks) (30 Minutes)
(Question Nos. 1 to 20 Online Examinations)

PART B (5 X 4= 20 Marks) (2 ½ Hours)
Answer ALL the Questions

21. a. List out the steps associated with the process of knowledge discovery.

Or

b. Explain about spatial Databases.

22. a. Distinguish between OLTP and OLAP.

Or

b. Explain various types of OLAP servers.

23. a. List out the steps involved in Data transformation.

Or

b. List out the possible samples involved in data reduction technique.

24. a. Explain the various kinds of constraints provided by the user in Association mining.

Or

b. What are the preprocessing steps preparing data for classification and prediction.

25. a. Explain how to enhance Basic Decision Tree Induction.

Or

b. Discuss the criteria followed for evaluating and comparing methods in classification and prediction.

34

PART C (3 x 10 = 30 Marks)
Answer any THREE Questions

26. Explain the requirements of clustering in data mining.

27. Explain how to integrate Datawarehousing techniques and decision tree induction.

28. Explain how to mine frequent itemsets without candidate generation.

29. Explain the architecture for on-line analytical mining.

30. Explain the architecture of a typical datamining system.

Reg. No.....

[11CSP201]

KARPAGAM UNIVERSITY

(Under Section 3 of UGC Act 1956)
COIMBATORE – 641 021
(For the candidates admitted from 2011 onwards)

M.Sc. DEGREE EXAMINATION, APRIL 2012

Second Semester

COMPUTER SCIENCE

DATA MINING AND WAREHOUSING

Time: 3 hours Maximum : 100 marks

PART – A (15 x 2 = 30 Marks)
Answer ALL the Questions

1. What is data mining?
2. Give some data mining techniques.
3. Define KDD.
4. What are dimension table?
5. Give any two ways in which summary table differ from fact table.
6. What is the use of unique identifier?
7. Give the purpose of creating data mart?
8. Why aggregations are performed?
9. Give the function of load manager.
10. What is rule based optimizer?
11. Define RAID.
12. Discuss about classification.
13. What are the two forms taken by vertical partitioning?
14. List the types of data in cluster analysis.
15. Application of text mining.

PART B (5 X 14= 70 Marks)
Answer ALL the Questions

16. a. Discuss about the six stages of KDD.

Or

b. Discuss about the major issues in Data mining.

17. a. Define Data preprocessing and explain in detail about the need of data preprocessing.

Or

b. Discuss about concept hierarchy generation in detail.

18. a. Explain in detail about the issues Related to classification and prediction.

Or

b. Discuss in detail about the Decision tree induction classification.

19. a. Write short notes on : i. Mining frequent patterns ii. Correlation

Or

b. Discuss about the various kinds of association rules.

20. a. Write short notes on : i. Web mining ii. Spatial mining.

Or

b. Discuss about content based cluster analysis and outlier analysis.

Reg. No.....

[12CSP201]

KARPAGAM UNIVERSITY
(Under Section 3 of UGC Act 1956)
COIMBATORE – 641 021
(For the candidates admitted from 2012 onwards)

M.Sc. DEGREE EXAMINATION, APRIL 2013
Second Semester

COMPUTER SCIENCE

DATA MINING AND WAREHOUSING

Time: 3 hours Maximum : 100 marks

PART A(15 X 2= 30 Marks)
Answer ALL the Questions

1. What is Data warehousing?
2. What are fact tables and dimension tables?
3. What is an OLAP system?
4. Describe Data cleaning.
5. Explain data Transformation.
6. Write short notes on concept hierarchy generation for categorical data.
7. Explain Association rule in mathematical notations.
8. How are association rules mined from large databases?
9. Give few techniques to improve the efficiency of Apriori algorithm.
10. What is Decision tree?
11. Describe Tree pruning methods.
12. Define the concept of prediction.
13. What are the fields in which clustering techniques are used?
14. What are the requirements of cluster analysis?
15. Write short notes on partitioning method?

PART B (5 x 14 = 70 Marks)
Answer ALL the Questions

16. a. Explain the architecture of data mining system.
Or
b. Explain the various techniques in data mining.

17. a. Discuss about Data Reduction in detail.
Or
b. Explain in detail about Entropy-Based Discretization.

18. a. Explain in details Constraint-Based Association Mining.
Or
b. Write a brief notes on Apriori algorithm.

19. a. Describe Rule-based classification in detail.
Or
b. How to increase an accurancy of Decision Tree Induction? Explain.

20. a. Write a short notes on
i. Text Mining ii. Web Mining iii. Spatial Mining iv. Multimedia Mining
Or
b. What is Hierarchical method? Explain in detail.

Reg. No.....

[15CSP103]

KARPAGAM UNIVERSITY
Karpagam Academy of Higher Education
(Established Under Section 3 of UGC Act 1956)
COIMBATORE – 641 021
(For the candidates admitted from 2015 onwards)

M.Sc., DEGREE EXAMINATION, NOVEMBER 2015
First Semester

COMPUTER SCIENCE

DATA MINING AND WAREHOUSING

Time: 3 hours Maximum : 60 marks

PART – A (20 x 1 = 20 Marks) (30 Minutes)
(Question Nos. 1 to 20 Online Examinations)

(Part - B & C 2 ½ Hours)

PART B (5 x 6 = 30 Marks)
Answer ALL the Questions

21. a. Explain Major Issues in Data Mining.
Or
b. Explain Data Warehouse Implementation.

22. a. Describe Data Cleaning.
Or
b. Explain: i. Data Transformation. ii. Data Reduction.

23. a. Explain Mining Various Kinds of Association Rules.
Or
b. Explain Constraint-Based Association Mining.

24. a. Describe Ensemble Methods—Increasing the Accuracy.
Or
b. Explain the Issues Regarding Classification and Prediction.

25. a. Explain Types of Data in Cluster Analysis.
Or
b. Explain the Categorization of Major Clustering Methods.

PART C (1 x 10 = 10 Marks)
(Compulsory)

26. Write a Case Study on 'Mining Object, Spatial, Multimedia, Text, and Web Data'.

Reg. No.....

[16CSP103]

KARPAGAM UNIVERSITY
Karpagam Academy of Higher Education
(Established Under Section 3 of UGC Act 1956)
COIMBATORE – 641 021
(For the candidates admitted from 2016 onwards)

M.Sc., DEGREE EXAMINATION, NOVEMBER 2016
First Semester

COMPUTER SCIENCE

DATA MINING AND WAREHOUSING

Time: 3 hours Maximum : 60 marks

PART – A (20 x 1 = 20 Marks) (30 Minutes)
(Question Nos. 1 to 20 Online Examinations)

(Part - B & C 2 ½ Hours)

PART B (5 x 6 = 30 Marks)
Answer ALL the Questions

21. a. With neat diagram, explain about three tier data warehouse architecture in detail.

Or

b. How OLAP different from OLTP and how are they similar?

22. a. How are principal components analysis useful in data preprocessing? Explain its procedure.

Or

b. Elaborate the term data cleaning as a process.

23. a. Sketch the importance of market basket analysis.

Or

b. What are constraints based association mining? Explain in detail.

24. a. Discuss the importance of Bayesian belief network.

Or

b. How does back propagation work? Explain in detail.

25. a. How are grid based methods useful in clustering? Briefly describe the different approach behind statistical information grid.

Or

b. What are typical requirements of clustering? Explain in detail.

PART C (1 x 10 = 10 Marks)
CASE STUDY (Compulsory)

26. Suppose that a data warehouse consists of the four dimensions, date, spectator, location and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectator may be Students, Adults or Seniors, with each category having its own charging rate.

i. Draw a star schema for the data warehouse.

ii. Starting with the base Cuboid (date, spectator, location, gate), which specifies OLAP operations should one perform in order to list the total charge paid by student spectators at GM-Place in 2012.
