BAOU
Education
for All

Dr. Babasaheb Ambedkar Open University

(Established by Government of Gujarat)

Data Warehouse and Data Mining

BSCIT-504

Bachelor Of Science (Hons.)-
Information Technology
(BSCIT)

2023

Data Warehousing and Data Mining

Dr. Babasaheb Ambedkar Open University

Data Warehousing and Data Mining

SLM Preparation Team

Mr. Atowar Islam, Royal Global University

Dr. Swapnanil Gogoi, IDOL, Gauahati University

Mr. Biswajit Das, Cotton University

Mr. Dwipen Laskar, Gauhati University

Ms. Daisy Kalita, USTM

Dr. Naba Jyoti Sarmah, Nalbari Commerce College

Ms. Daisy Kalita, USTM

Mr. Biswajit Das, Cotton University

Content Editors

Prof. (Dr.) Nilesh K. Modi	Professor and Director, School of Computer Science, Dr. BabasahebAmbedkar Open University, Ahmedabad
Mr. Nilesh Bokhani	Assistant Professor, School of Computer Science, Dr. Babasaheb Ambedkar Open University, Ahmedabad

Acknowledgement: The content in this book is modifications based on the work created and shared by Krishna Kanta Handiqui State University for the subject Data Warehousing and Data Mining used according to terms described under under a Creative Commons Attribution 4.0 International License.

ISBN -

Printed and published by: Dr. Babasaheb Ambedkar Open University, Ahmedabad While all efforts have been made by editors to check accuracy of the content, the representation of facts, principles, descriptions and methods are that of the respective module writers. Views expressed in the publication are that of the authors, and do not necessarily reflect the views of Dr. Babasaheb Ambedkar Open University. All products and services mentioned are owned by their respective copyrights holders, and mere presentation in the publication does not mean endorsement by Dr. Babasaheb Ambedkar Open University. Every effort has been made to acknowledge and attribute all sources of information used in preparation of this learning material. Readers are requested to kindly notify missing attribution, if any.

Dr. Babasaheb Ambedkar Open University

BSCIT-504

Data Warehouse and Data Mining

BLOCK-1:

BLOCK-2:

BLOCK-3:

BLOCK-4:

UNIT 1: INTRODUCTION TO DATA MINING

UNIT STRUCTURE

1.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define data mining
- describe the different types of data
- describe data mining task primitives
- explain integration of data mining system
- describe the major issues of data mining.

1.2 INTRODUCTION

In this unit, we will learn about data mining. We will also learn about the different types of data like non-dependency oriented data and dependency oriented data. Besides data mining task primitives and how data mining system can be integrated. In addition to this, the major issues of data mining will be discussed in this unit while in the next unit, we will provide an introduction to data warehousing.

1.3 DATA MINING

Data Mining is a non-trivial process of discovering knowledge from huge amount of data, as mining of gold from rocks or sand is called gold mining, similarly data mining is appropriately named as knowledge mining. To extract the knowledge from large data set knowledge discovery from data (KDD) is used.

Data mining is the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data. A wide variation exists in terms of the problem domains, applications, formulations, and data representations that are encountered in real applications. Therefore, “data mining” is a broad umbrella term that is used to describe these different aspects of data processing.

In the modern age, virtually all automated systems generate some form of data either for diagnostic or analysis purposes. Some examples of different kinds of data are as follows:

- **World Wide Web:** The number of documents on the indexed Web is now in the order of billions, and the invisible Web is much larger. User accesses to such documents create Web access logs at servers and customer behavior profiles at commercial sites. Furthermore, the linked structure of the Web is referred to as the Web graph, which itself is a kind of data. These different types of data are useful in various applications. For example, the Web documents and link structure can be mined to determine associations between different topics on the Web. On the other hand, user access logs can be mined to determine frequent patterns of accesses or unusual patterns of possibly unwarranted behavior.
- **Financial Interactions:** Most common transactions of everyday life, such as using an automated teller machine (ATM) card or a credit

card, can create data in an automated way. Such transactions can be mined for many useful insights such as fraud or other unusual activity.

- **User Interactions:** Many forms of user interactions create large volumes of data. For example, the use of a telephone typically creates a record at the telecommunication company with details about the duration and destination of the call. Many phone companies routinely analyze such data to determine the relevant patterns of behavior that can be used to make decisions about network capacity, promotions, pricing, or customer targeting.
- **Sensor Technologies and the Internet of Things:** A recent trend is the development of low-cost wearable sensors, smart phones, and other smart devices that can communicate with one another. By one estimate, the number of such devices exceeded the number of people on the planet in 2008. The implications of such massive data collection are significant for mining algorithms.

The deluge of data is a direct result of advances in technology and the computerization of every aspect of modern life. It is, therefore, natural to examine whether one can extract concise and possibly actionable insights from the available data for application-specific goals. This is where the task of data mining comes in. The raw data may be arbitrary, unstructured, or even in a format that is not immediately suitable for automated processing. For example, manually collected data may be drawn from heterogeneous sources in different formats and yet somehow needs to be processed by an automated computer program to gain insights. To address this issue, data mining analysts use a pipeline of processing, where the raw data are collected, cleaned, and transformed into a standardized format. The data may be stored in a commercial database system and finally processed for insights with the use of analytical methods. In fact, while data mining often conjures up the notion of analytical algorithms, the reality is that the vast majority of work is related to the data preparation portion of the process. This pipeline of processing is conceptually similar to that of an actual mining process from a mineral ore to the refined end product. The term “mining” derives its roots from this analogy. From an analytical perspective, data mining is challenging because of the wide disparity in the problems and data types that are encountered. For example, a commercial product recommendation problem is very different from an intrusion-detection application, even at the level of the input data format or the problem definition. Even within related classes of problems, the differences are quite significant. For example, a product recommendation problem in a multidimensional database is very different from a social recommendation problem due to the differences in the underlying data type. Nevertheless, in spite of these differences, data mining applications are often closely connected to one of four “super problems” in data mining: association pattern mining, clustering, classification, and outlier detection. These problems are so important because they are used as building blocks in a majority of the applications in some indirect form or the other. This is a useful abstraction because it helps us conceptualize and structure the field of data mining more effectively. The data may have different formats or types. The type may be quantitative (e.g., age), categorical (e.g., ethnicity), text, spatial, temporal, or graph-oriented. Although the most common form of data is multidimensional, an increasing proportion belongs to more complex data types. While there is a conceptual portability of algorithms between many data types at a very high level, this is not the case from a practical perspective. The reality is that the precise data type may affect the behavior of a particular algorithm significantly. As a result, one may need to design refined variations of the basic approach for multidimensional data, so that it can be used effectively for a different data type. Therefore, this book will dedicate different chapters to the various data types to provide a better understanding of how the processing methods are affected by the underlying data type.

1.3.1 Various types of Data

One of the interesting aspects of the data mining process is the wide variety of data types that are available for analysis. There are two broad types of data, of varying complexity, for the data mining process: non-dependency-oriented data and dependency-oriented data.

Table 1.1: Example of Multidimensional Data Set

Name	Age	Gender	Race	ZIP Code
Decock S	34	M	Australian	05139
Kohli B	40	M	Indian	10598
Sahni A	35	F	Asian	90201

1.3.1.1 Non-dependency-oriented data

This typically refers to simple data types such as multidimensional data or text data. These data types are the simplest and most commonly encountered. In these cases, the data records do not have any specified dependencies between either the data items or the attributes. An example is a set of demographic records about individuals containing their age, gender, and ZIP code.

Non-dependency-oriented data are the simplest form of data and typically refers to multidimensional data. This data typically contains a set of records. A record is also referred to as a data point, instance, example, transaction, entity, tuple, object, or feature-vector, depending on the application at hand. Each record contains a set of fields, which are also referred to as attributes, dimensions, and features.

The Non-dependency-oriented data are divided into the following categories–

0 **Quantitative Multidimensional Data:** The attributes in Tables 1.1 are of two different types. The age field has values that are numerical in the sense that they have a natural ordering. Such attributes are referred to as continuous, numeric, or quantitative. Data in which all fields are quantitative is also referred to as quantitative data or numeric data. In the data mining literature, this particular subtype of data is considered the most common. This subtype is particularly convenient for analytical processing because it is much easier to work with quantitative data from a statistical perspective.

0 **Categorical and Mixed Attribute Data:** Many data sets in real applications may contain categorical attributes that take on discrete unordered values. For example, in Table 1.1, the attributes such as gender, race, and ZIP code, have discrete values without a natural ordering among them. In the case of mixed attribute data, there is a combination of categorical and numeric attributes. The full data in Table 1.1 are considered mixed-attribute data because they contain both numeric and categorical attributes. The attribute corresponding to gender is special because it is categorical, but with only two possible values. In such cases, it is possible to impose an artificial ordering between these values and use algorithms designed for numeric data for this type. This is referred to as binary data, and it can be considered a special case of either numeric or categorical data.

0 **Binary and Set Data:** Binary data can be considered a special case of either multidimensional categorical data or multidimensional quantitative data. It is a special case of multidimensional categorical data, in which each categorical attribute may take on one of at most two discrete values. It is also a special case of multidimensional quantitative data because an ordering exists between the two values.

0 **Text Data:** Text data can be viewed either as a string, or as multidimensional data, depending on how they are represented. In its raw form, a text document corresponds to a string. This is a dependency-oriented data type. Each string is a sequence of characters(or words) corresponding to the document. However, text documents are rarely represented as strings.

1.3.1.2 Dependency-oriented data

In this type of data, implicit or explicit relationships may exist between data items. For example, a social network data set contains a set of *vertices* (data items) that are connected together by a set of *edges* (relationships). On the other hand, time series contains implicit dependencies. For example, two successive values collected from a sensor are likely to be related to one another. Therefore, the time attribute implicitly specifies a dependency between successive readings.

The knowledge about preexisting dependencies greatly changes the data mining process because data mining is all about finding relationships between data items.

The presence of preexisting dependencies, therefore, changes the expected relationships in the data, and this may be considered interesting from the perspective of these expected relationships. Several types of dependencies may exist that may be either implicit or explicit:

0 **Implicit dependencies:** In this case, the dependencies between data items are not explicitly specified but are known to “typically” exist in that domain.

0 **Explicit dependencies:** This typically refers to graph or network data in which edges are used to specify explicit relationships. Graphs are a very powerful abstraction that is often used as an intermediate representation to solve data mining problems in the context of other data types.

The different dependency-oriented data types are discussed in detail.

0 **Time-Series Data:** Time-series data contain values that are typically generated by continuous measurement over time. For example, an environmental sensor will measure the temperature continuously, whereas an electrocardiogram (ECG) will measure the parameters of a subject’s heart rhythm. Such data typically have implicit dependencies

built into the values received over time. For example, the adjacent values recorded by a temperature sensor will usually vary smoothly over time, and this factor needs to be explicitly used in the data mining process.

0 **Discrete Sequences and Strings:** Discrete sequences can be considered the categorical analog of time-series data. As in the case of time-series data, the contextual attribute is a time stamp or a position index in the ordering. The behavioral attribute is a categorical value. Therefore, discrete sequence data are defined in a similar way to time-series data.

0 **Spatial Data:** In spatial data, many non spatial attributes (e.g., temperature, pressure, image pixel color intensity) are measured at spatial locations. For example, sea-surface temperatures are often collected by meteorologists to forecast the occurrence of hurricanes. In such cases, the spatial coordinates correspond to contextual attributes, whereas attributes such as the temperature correspond to the behavioral attributes. Typically, there are two spatial attributes. As in the case of time-series data, it is also possible to have multiple behavioral attributes. For example, in the sea-surface temperature application, one might also measure other behavioral attributes such as the pressure.

0 **Network and Graph Data:** In network and graph data, the data values may correspond to nodes in the network, whereas the relationships among the data values may correspond to the edges in the network. In some cases, attributes may be associated with nodes in the network. Although it is also possible to associate attributes with edges in the network, it is much less common to do so.

1.4 DATA MINING FUNCTIONALITIES

Data mining functionalities and variety of knowledge are presented as follows:

- **Classification:** Classification analysis is the organization of data in given classes or in short classification is to partition the given data into predefined disjoint groups. For example, a bank loan officer want to analyze which loan applicant are appropriate and which can create risk.
- **Clustering:** Data items are grouped according to logical relationship or customer preferences. For example, data can be mined to identify market segment or customer affinities.
- **Outlier Analysis:** Outlier are data elements that cannot grouped in a given class or cluster. Basically outliers are considered as noise and are always removed from applications.
- **Association:** Association analysis is the discovery of what are commonly called *association rules*. It studies the frequency of items occurring together in transactional databases, and based on a threshold called *support*, identifies the frequent item sets. Another threshold, *confidence*, which is the conditional probability that an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis.
- **Data Characterization:** Data characterization is a summarization of general features of objects in a target class, and it produces what is called characteristic rules. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions. For example, one may want to characterize the OurVideoStore customers who regularly rent more than 30 movies a year. With concept hierarchies on the attributes describing the target class, the attribute-oriented induction method can be used, for example, to carry out data summarization. Note that with a data

cube containing summarization of data, simple OLAP operations fit the purpose of data characterization.

- **Data Discrimination:** Data discrimination produces what are called discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. For example, one may want to compare the general characteristics of the customers who rented more than 30 movies in the last year with those whose rental account is lower than 5. The techniques used for data discrimination are very similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures.
- **Prediction:** Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data.
- **Evolution and Deviation Analysis:** Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

1.5 CLASSIFICATION OF DATA MINING SYSTEM

Data mining systems can be categorized according to various criteria as follows:

- **Classification of data mining systems according to the type of data sources mined:** This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.
- **Classification of data mining systems according to the database involved:** This classification based on the data model involved such as relational database, object oriented database, data warehouse, transactional database, etc.
- **Classification of data mining systems according to the kind of knowledge discovered:** This classification 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 of data mining systems according to mining techniques used:** This classification is 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.

1.6 DATA MINING TASK PRIMITIVES

Each user will have a data mining task in mind, that is, some form of data analysis that he or she would like to have performed. A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives. These primitives allow the user to inter- actively communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths.

- **The set of task-relevant data to be mined:** This specifies the portions of the database or the set of data in which the user is interested.

This includes the database attributes or data warehouse dimensions of interest (referred to as the relevant attributes or dimensions).

- **The kind of knowledge to be mined:** This specifies the data mining functions to be per- formed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.
- **The background knowledge to be used in the discovery process:** This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found. Concept hierarchies are a popular form of back- ground knowledge, which allow data to be mined at multiple levels of abstraction. User beliefs regarding relationships in the data are another form of back- ground knowledge.
- **The interestingness measures and thresholds for pattern evaluation:** They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures. For example, interestingness measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.
- **The expected representation for visualizing the discovered patterns:** This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes. A data mining query language can be designed to incorporate these primitives, allowing users to flexibly interact with data mining systems. Having a data mining query language provides a foundation on which user-friendly graphical interfaces can be built.

1.7 INTEGRATION OF DATA MINING SYSTEM

If a data mining system is not integrated with a database or a data warehouse system, then there will be no system to communicate with. This scheme is known as the non-coupling scheme. In this scheme, the main focus is on data mining design and on developing efficient and effective algorithms for mining the available data sets.

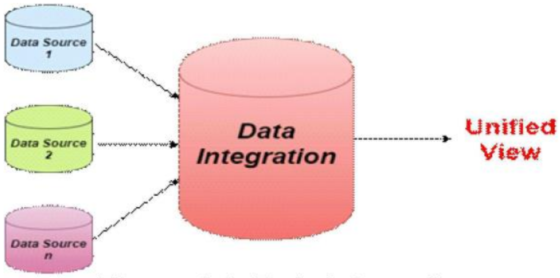


Figure 1.1: Data Integration

The list of integration schemes is as follows:

- **No Coupling:** In this scheme, the data mining system does not utilize any of the database or data warehouse functions. It fetches the data from a particular source and processes that data using some data mining algorithms. The data mining result is stored in another file.
- **Loose Coupling:** In this scheme, the data mining system may use some of the functions of database and data warehouse system. It fetches the data from the data respiratory managed by these systems and performs data mining on that data. It then stores the mining result either in a file or in a designated place in a database or in a data warehouse.
- **Semi-tight Coupling:** In this scheme, the data mining system is linked with a database or a data warehouse system. In addition to that, efficient implementations of a few data mining primitives can be provided in the database.
- **Tight coupling:** In this coupling scheme, the data mining system is smoothly integrated into the database or data warehouse system. The data mining subsystem is treated as one functional component of an information system.

1.8 MAJOR ISSUES OF DATA MINING

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. In this section, we will discuss the major issues regarding–

- Mining methodology and user interaction
- Performance issues
- Diverse data types issues

The following diagram describes the major issues.

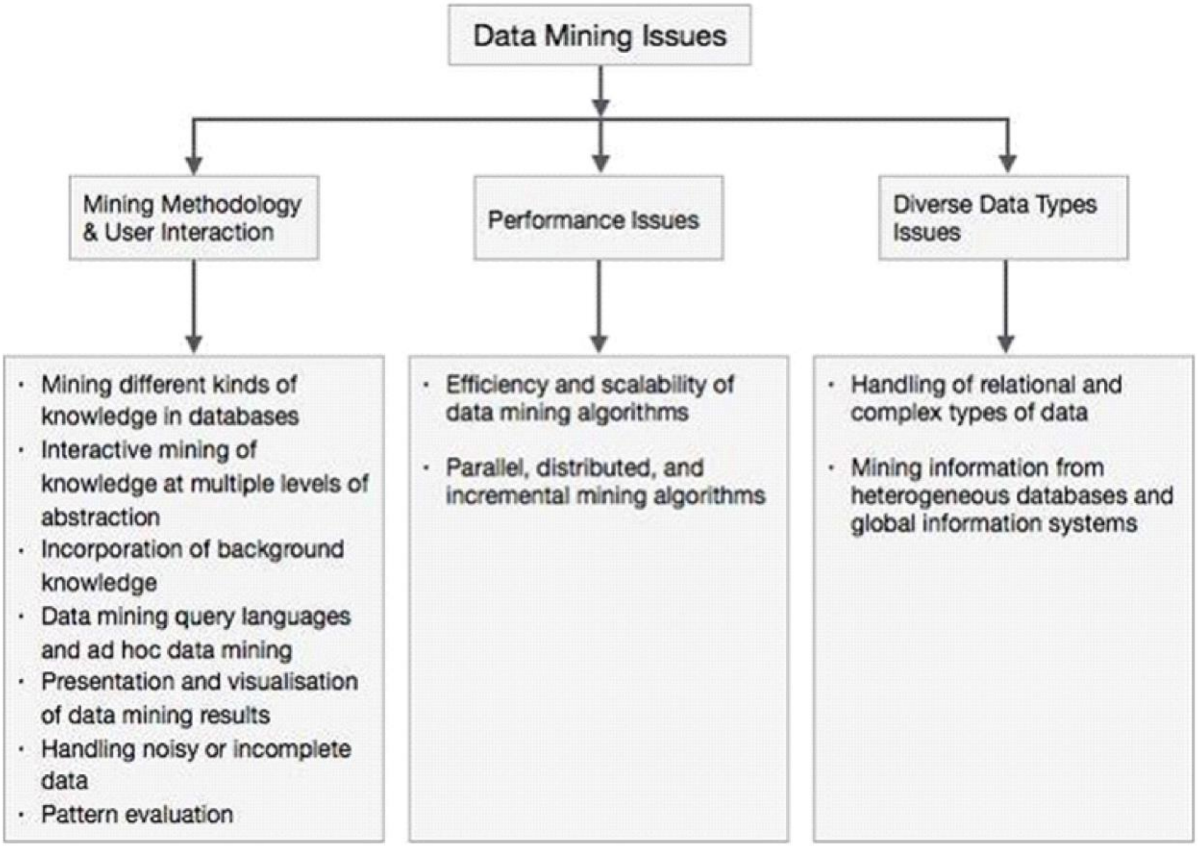

Data Mining Issues
Mining Methodology & User Interaction
Performance Issues
Diverse Data Types Issues
· Mining different kinds of knowledge in databases
· Interactive mining of knowledge at multiple levels of abstraction
· Incorporation of background knowledge
· Data mining query languages and ad hoc data mining
· Presentation and visualisation of data mining results
· Handling noisy or incomplete data
· Pattern evaluation
· Efficiency and scalability of data mining algorithms
· Parallel, distributed, and incremental mining algorithms
· Handling of relational and complex types of data
· Mining information from heterogeneous databases and global information systems

Figure 1.2: Data Mining Issues

1.8.1 Mining Methodology and User Interaction Issues

It refers to the following kinds of issues–

- **Mining different kinds of knowledge in databases:** Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction:** The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.

- **Incorporation of background knowledge:** To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining:** Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results:** Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data:** The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation:** The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

1.8.2 Performance Issues

There can be performance-related issues such as follows:

- **Efficiency and scalability of data mining algorithms:** In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms:** The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions are merged. The incremental algorithms, update databases without mining the data again from scratch.

1.8.3 Diverse Data Types Issues

- **Handling of relational and complex types of data:** The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems:** The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

CHECK YOUR PROGRESS

Q.1: Answer the following multiple choice questions:

. ...is an essential process where intelligent methods are applied to extract data patterns.

i) Data warehousing ii) Data mining

iii) Text mining iv) Data selection

Q.2: Which of the following is not a data mining functionality?

i) Characterization and Discrimination

ii) Classification and regression

iii) Selection and interpretation

iv) Clustering and Analysis

Q.3:........................ is a summarization of the general characteristics or features of a target class of data.

i) Data characterization ii) Data classification

iii) Data discrimination iv) Data selection

Q.4:..........................is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.

i) Data characterization ii) Data classification

iii) Data discrimination iv) Data selection

Q.5: .... is the process of finding a model that describes and distinguishes data classes or concepts.

i) Data characterization ii) Data classification

iii) Data discrimination iv) Data selection

1.9 LET US SUM UP

- Data Mining is a non-trivial process of discovering knowledge from huge amount of data, as mining of gold from rocks or sand is called gold mining, similarly data mining is appropriately named as knowledge mining.
- To extract the knowledge from large data set Knowledge Discovery from Data (KDD) is used.
- Data mining is the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data.
- Non-dependency-oriented data are the simplest form of data and typically refers to multidimensional data. This data typically contains a set of records.
- A record is also referred to as a data point, instance, example, transaction, entity, tuple, object, or feature-vector, depending on the application at hand.
- In case of dependency-oriented data, implicit or explicit relationships may exist between data items.
- Classification analysis is the organization of data in given classes or in short classification is to partition the given data into predefined disjoint groups.
- Association analysis is the discovery of what are commonly called *association rules*.

1.10 FURTHER READING

1) Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.

2) Pujari, A. K. (2001). *Data Mining Techniques.* Universities Press.

3) Saxena, A., Saxena, K. Saxena, S. (2015). *Data Mining and Warehousing.* BPB Publications.

4) https://www.tutorialspoint.com/data_mining/dm_quick_guide.htm

1.11 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: ii) Data mining

Ans. to Q. No. 2: iii) Selection and interpretation

Ans. to Q. No. 3: i) Data characterization

Ans. to Q. No. 4: iii) Data discrimination

Ans. to Q. No. 5: ii) Data classification

1.12 MODEL QUESTIONS

Q.1: What is data mining?

Q.2: What is KDD?

Q.3: Explain the various types of basic data.

Q.4: Explain data mining functionality.

Q.5: Explain the classification of data mining.

Q.6: Explain the major issues of data mining.

*** ***** ***

UNIT 2: INTRODUCTION TO DATA WAREHOUSING

UNIT STRUCTURE

2.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define data warehouse and data base management system
- differentiate between data warehouse and data base management system
- explain the reasons behind the requirement of data warehouse
- describe operational and informational data stores

- list data warehouse characteristics
- describe how a data warehouse is built and what is its structure
- calculate the cost of data warehousing.

2.2 INTRODUCTION

In the later part of 1980, Barry Devlin and Paul Murphy had developed a data warehouse where the flow of information from operational databases to decision support system has been proposed. Requirement of different information to make strategic decisions has become rapidly increased in 1990 due to the increase of businesses and global growth of different corporations. Due to the increase of competition and complexity in businesses, different organizations require more information that will help to make strategic decisions. The operational databases provide information only for executing daily normal operations but it could not be able to provide strategic information.

In 1990, a new concept termed as data warehouse has been evolved which can be able to provide strategic information. Different organizations had started to build data warehouse to make capable enough their decision support mechanism so that it will help to simplify and grow their businesses. In the previous unit we have got an introduction to data mining. In this unit, we will give an introduction to data ware housing. We will also learn about the data base management system and different aspects of data warehouse. In the next unit, we will explore the concept of OLAP in detail.

2.3 DATA BASE MANAGEMENT SYSTEM AND DATA WAREHOUSE

In this section, the definitions of data warehouse and data base management system (DBMS) are presented. The differences between data warehouse and DBMS are also discussed in the later part of this section.

2.3.1 Data Base Management System

A database management system (DBMS) can be defined as a software system that is responsible for data storing,data manipulation and data validation in a database. It supports data access from a database with an efficient and secured manner. Management of data and data format in a database are also supported by DBMS.

2.3.2 Data Warehouse

A data warehouse is a collection of integrated data from various types of sources that help to prepare analytical reports and in the efficient performance of decision making systems. It also supports structured and ad hoc queries. Data in a data warehouse are structured according to different subjects. The format of data may be different in different sources as they may be received from different applications. But in data warehouse, all data must be stored with a common format and to make it possible, data cleaning and data integration techniques are applied to convert inconsistent data to consistent data. The data warehouse stores data according to a particular time unit. For example, examination results of a college from 2014 to 2018. So historical information are only stored in data warehouse. Physical data storage for data warehouse is always made separate to the operational data bases and so it does not require any transaction processing, recovery and concurrency control mechanism. In most of the times, the data in a data warehouse cannot be modified. Only new data are continuously loaded to it.

2.3.3 Difference between Data Warehouse and DBMS

There are some differences between data warehouse and DBMS. These differences are discussed as below:

- DBMS contains transactional data and it is termed as OnLine Transaction Processing (OLTP) system. On the other hand, data warehouse contains analytical data and it is termed as OnLine Analytical Processing (OLAP) system. So DBMS are constructed to record data and data warehouses are constructed to analyze data.

- DBMS stores application oriented data and at most of the times it is based on single application. But data warehouse stores subject oriented historical data received from multiple heterogeneous sources.
- In DBMS, data are changed regularly as different transactions are performed frequently. But in case of data warehouse, data are not allowed to be changed or modified at most of the times.
- Recovery and concurrency control mechanism are very essential in DBMS as data are changed regularly due to different transaction processing operations. But data warehouse does not require any recovery and concurrency control mechanism.
- Data warehouse deals with long duration historical data but DBMS do not operate on long duration data as most of the times it deals with the current data with short time duration.
- DBMS can provide very less support to the decision making system but decision support mechanism is significantly supported by the data warehouse.

2.4 THE NEED FOR DATA WAREHOUSING

Building and utilization of a data warehouse is termed as data warehousing. Since 1990, data warehousing has become an essential part of every organization due to the increase in requirement of strategic information to make their decision making system efficient. Strategic information always helps the executives and managers to provide better business policies for their companies so that better results in businesses have been achieved by them than their competitors. Because of data warehousing, the time required for data analysis is reduced significantly. Historical and analytical information from data warehouse are utilized by the executive and the managers to achieve the following objectives:

- To learn about different operations related to the organization.
- To learn and identify different primary factors that can affect the business and monitor these factors in a regular period of time so that the business performance of the organizations can be compared to that of its competitors.

- To regularly monitor the requirements of customers and their preferences.
- To continuously monitor the results of sales and marketing.
- To regularly monitor product quality and services.
- To learn about new technologies that can be utilized to make businesses grow and to simplify business related complexities.

Data warehousing provides a common data format for all the stored integrated data that are received from various sources and so for this reason the data analysis and sharing of analyzed data become easier. Because of the common data format, data warehousing can minimize the possibility of error in data interpretation and improve data consistency.

Data warehousing provides easier data accessibility as data are stored separately in one place. It can store multidimensional data and provide support to the users to perform query analysis and analysis of stored data.

Data warehouse also maintains data security by providing secure access to the stored data by authenticated users. In this environment, the users are allowed to access only those data which are specific to them. It can also provide accessibility of corporate data to the authenticated customers and vendors for the development of new business strategies.

In recent times, data warehousing has become an integral part of every large organization. Data warehousing is commonly applied in the following domains:

- Banking and financial services
- Analysis of biological data
- Quality product manufacturing
- Consumer data analysis
- Retail sectors
- E-commerce
- Telecom sectors
- Logistics and Inventory management
- Insurance sectors
- Educational institutions for educational data analysis

2.5 OPERATIONAL DATA STORES

Operational data of different organizations are constructed by various on-line transaction processing operations of operational applications. These data are recent, complete, non redundant and modifiable data. Operational data store (ODS) is a database that stores operational data from various sources and process these data. After processing,ODS transfers these data to operational systems and the data warehouses. So, ODS stores recent integrated operational data about various products, customers etc. These data are not application specific and accessible from all parts of an organization.

ODS has similarities as well as differences with data warehouse. Similarities of ODS and data warehouse are as follows:

- Both ODS and data warehouse contains subject oriented data. In both cases data are not application specific.
- Data in ODS are entirely integrated thus showing a similarity with data warehouse.

Differences of ODS and data warehouse are summarized below:

- First of all, the architecture of ODS and data warehouse is fairly different.
- ODS stores short term current data but on the other hand data warehouse contains long duration historical data.
- In case of ODS, data are regularly changeable or updatable. When new data are transmitted to the ODS, then these recent data will overwrite the older version data of the related fields. So, no historical data are available in ODS. But in case of data warehouse, data are not changeable or updateable. So all historical data are available in data ware house.

2.6 INFORMATIONAL DATA STORES

Informational data store is a collection of summarized and redundant data about different subjects like product, customer, vendor etc. Informational data are utilized to provide appropriate response to the different queries placed by corporate executives and managers for decision making purpose. Informational data for a corporation can be collected from different applications, databases, computer systems and operational data stores which are available in the corporation. These data are not changeable or updatable.

Differences between informational data store and operational data store are discussed below:

- Data model in case of an operational data store is normalized to maintain ACID properties i.e., atomicity, consistency, isolation, and durability. But it is not required in case of informational data store.
- Informational data store contains current, redundant, summarized and historical data. But operational data store contains only short duration current data.
- Data accessing in case of informational data store is performed primarily on ad hoc basis. On the other hand, only predefined and structured access of data is possible in caseof operational data store.
- In case of informational data store, data are not changed or updated frequently. At most of the times, periodic and planned batch wise data updates are observed in informational data store. But in operational data store, data are continuously changed or updated fewer current data.
- The number of concurrent users of informational data store is always fewer than that of the operational data store.

2.7 DATA WAREHOUSE CHARACTERISTICS

The main characteristics of a data warehouse are described as follows:

- Data warehouse is a type of database that stores subject-oriented data achieved from various sources or applications to provide support to the decision making system of a corporation by performing data analysis. Data are stored in data warehouse according to some time period so that data analysis becomes accurate.
- Data warehouse is a collection of integrated and standardized data. All applicable data from various sources and applications are

combined together and stored in data warehouse for efficient decision making purpose. These integrated data may be in different data formats as they are transmitted from different types of operational systems and applications. So, these data must be standardized by removing the inconsistencies from them so that all data are available with a common data format and can be utilized efficiently in the decision making.

- Data warehouse is non-volatile. Data warehouse contains current and long duration historical data. In most of the times, data are not changed or updated in data warehouse. Only new data are included to it. But if it supports change of data then data can be changed or updated periodically.
- Small number of user is supported by data warehouse.
- A data warehouse separates operational data stores and informational data store.
- It has been observed that external data are also very useful for efficient decision making by a corporation. Data warehouse integrates the useful external data from various sources or applications available outside of a company and maps it to the related applications of the company for better decision making.

2.8 DATA WAREHOUSE STRUCTURE

Two types of data warehouse architecture are frequently used by different corporations. These are two-tier architecture and three-tier architecture.

Two-Tier Architecture: The data warehouse two-tier architecture is based on **client–server architecture.** In this architecture, direct communication between the client and the data server is available. The client layer of this architecture is responsible for providing user interface, data access, data aggregation, data analysis, query specification and report formatting. The data warehouse server is the server layer of this architecture. Data logic and data services are executed by this layer. It also stores and maintains metadata. The two-tiered architecture lacks of scalability and flexibility. Large number of end-users also cannot be supported by this type of architecture. But it is easy to maintain and data modification is also easy in this case.

Three-Tier Architecture: Three-tier architecture is the most popular data warehouse architecture. This architecture consists of three layers that are **bottom tier**, **middle tier** and **top tier**.

- **Bottom-Tier:** The data warehouse database server is placed at the bottom tier in three-tier data warehouse architecture. This database server is a relational database system.
- **Middle-Tier:** The OLAP Server is implemented in the middle-tier.
- **Top-Tier:** The client layer is placed at the top-tier. It consists of different tools for data analysis, query, data mining and reporting.

There are seven basic components of data warehousing architecture as shown in figure 2.1 and discussed in the following parts of this section.

Seven basic components:

i) **Data warehouse Database:** In most of the cases, the data warehouse database is a relational database management system (RDBMS). It is the central database of the data warehousing system. But it has been observed that the traditional RDBMS system cannot be optimized for data warehousing due to some of the factors that can affect the performance of the data warehouse. Some examples of these factors are very large database size, ad-hoc query, aggregates, multi-table joins etc. Some technological approaches that can be used as a solution to this issue are as follows:

   - Parallel relational databases can be used in a data warehouse for better scalability.
   - Multidimensional database can also be used to remove limitations that are available due to the relational data model.
   - Efficient latest index structures can be used to avoid relational table scan which can improve speed.

ii) **Tools for data sourcing, data cleanup, data transformation and data migration:** The tools for data sourcing, data cleanup, transformation, and migration provide the conversion operations, structural modifications, summarizations and key changes. This component of data warehouse structure is required to convert dissimilar data into information so that it can be utilized by the decision support system. It is required to transmit data from various operational systems to the data warehouse. The metadata is also maintained by this component.

iii) **Metadata:** Metadata is defined as the data about data that describes the data warehouse. It is a very essential component of data warehouse architecture. It is utilized to build and manage the data warehouse. It identifies the data source, data usage, data values and data features of data warehouse. It also describes how data warehouse data can be processed and changed or updated. Metadata repository is responsible for metadata management. There are two classes of metadata available in a data warehouse *technical metadata* and *business metadata*.

- **Technical Metadata:** Technical metadata is the information about the data which are utilized by data warehouse administrators and designers for data warehouse development and management purpose. Some of the information stored in technical metadata are mentioned as follows:
    - Information about data sources
    - Information about the data transformation methods from operational databases to the data warehouse
    - Information about the procedures that are used for data cleanup and data enhancement
    - Information about data access and access permission
    - Information about data warehouse data structures
- **Business Metadata:** Business metadata stores information which helps the end-users to understand easily the various data contained in a data warehouse. Business metadata stores information about different subject areas of data warehouse. It stores the information that supports all data warehousing components. It also contains information regarding data usage, data history, data ownership etc.

iv) **Data Marts:** A data mart is a structure that stores data that are related to specific subject area and used by the specific group of users and departments of an organization. So data marts are smaller in size and more flexible than data warehouse. It is a subset of the data warehouse. Most of the times, data marts can be used as an alternative to large data warehouses because it is less expensive and less time is required to build it. Data marts can also reduce the response time for end-users by permitting users to be able to access only those specific data that are required by the users. Both data marts and data warehouse can be formed in the same database. Data mart can also be created in a physically separate database.

v) **Access Tools:** Access tools in data warehousing are utilized to provide information to business users for taking efficient strategic decisions. Access tools provide interactions between users and data warehouse system. There are four types of access tools available in data warehousing. These are *query and reporting tools*, *application development tools*, *data mining tools*, *OLAP tools and executive information system tools*.

vi) **Data warehouse administration and management:** Data warehouse administration and management component is responsible for security management, priority management, metadata management, monitoring data quality, data cleaning, data replicating, data distribution, monitoring data updates from various sources, assessment of data warehouse usage, data warehouse usage reporting etc. It is also responsible for managing data warehouse storage and maintaining backup and recovery process.

vii) **Information delivery system:** The information delivery component is responsible for the distribution of data warehouse data and different information items to end-user products and other data warehouses.

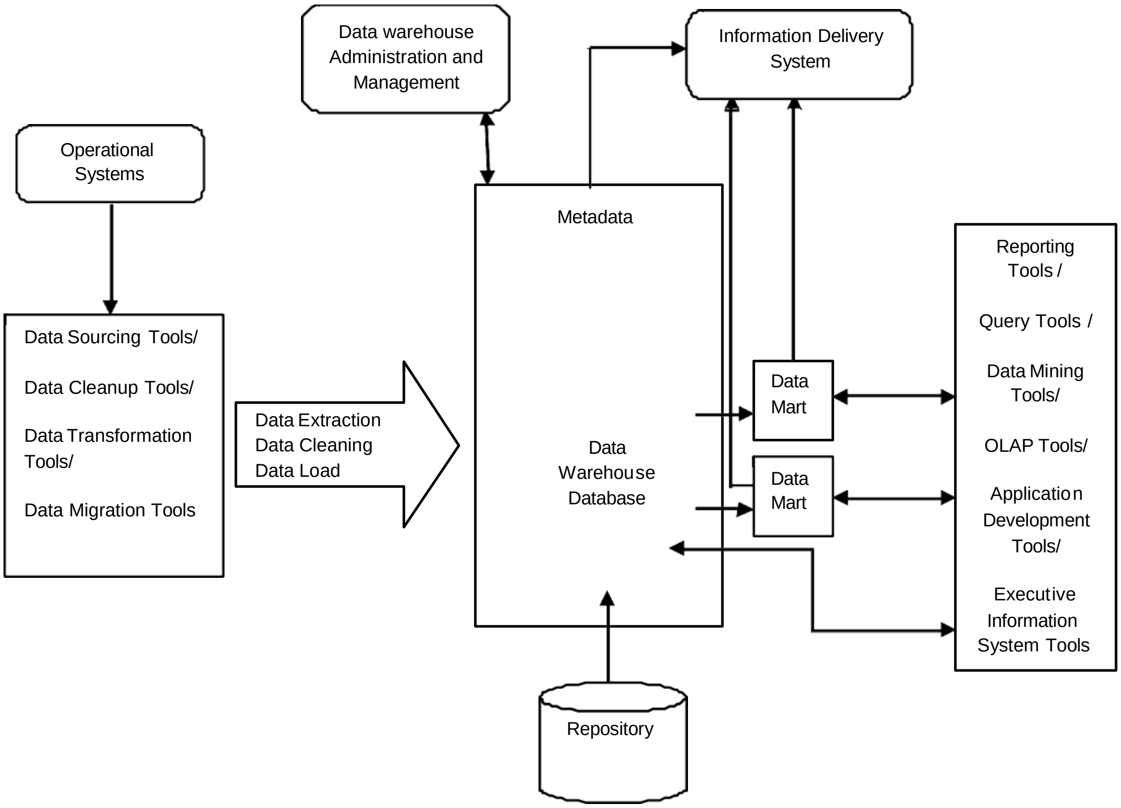


Figure 2.1: Data Warehousing Architecture

2.9 BUILDING A DATA WAREHOUSE

There are two approaches to build a data warehouse. These are top down approach and bottom up approach.

- ***In top down approach***, at first an enterprise data model is developed and then enterprise wise business requirements are collected. After these two processes, the process to build a data warehouse with data marts is started. In this approach, new data mart can be constructed very easily from the data warehouse. But this approach is not flexible enough for the development of a data warehouse where different departmental requirements may be changed. This approach is expensive and it requires more time for initial set up.
- ***In bottom up approach***, at first, individual data marts specific to different subject area and business requirements are constructed. Then these data marts are integrated to build the data warehouse. In this approach, extension of the data warehouse to include new

business area is very easy as it requires only the creation of new data mart and integration of it to the data warehouse. It require less time for initial set up. But in this approach, the integration of data mart to the data warehouse is a difficult process.

For building a data warehouse: design considerations; technical considerations and implementation considerations are discussed in the following subsections.

2.9.1 Design Consideration

All the components of data warehouse, all kind of data sources related to the data warehouse and all kind of information requirements are considered for designing an efficient data warehouse. Data integration from various sources of different types is a very important design consideration of data warehouse. Regular interactions with the end users are also a basic requirement for a successful data warehouse designing. Some other important points for successful data warehouse design are presented below:

- Data model of a data warehouse should be closely related to the data content and structure of the data warehouse. Data warehouse and data marts may have different data models.
- A data warehouse must have some method and technology to maintain metadata repository and to include new information to the metadata regularly.
- Data placement and data distribution strategies should be properly configured in the process of data warehouse designing.
- In recent times, there are various types of tools available that can be used to implement data warehouse. The most appropriate tools related to a particular data warehouse environment should be selected by the data warehouse designers to implement a data warehouse otherwise, it may affect the performance of the data warehouse.
- Data warehouse designer must clearly recognize the end users' requirements to access different data.

2.9.2 Technical Consideration

The technical considerations to build a data warehouse are discussed below:

- The hardware platform for a data warehouse server should be capable enough to store and maintain the required amount of data for decision support systems of an organization. The data warehouse server must be specialized so that it can perform all the related jobs of the data warehouse.
- Metadata repository must be supported by the hardware platform and the associated software.
- A balance between all kinds of computing components should be maintained to design and implement a data warehouse.
- The data warehouse DBMS must be compatible with the data warehouse so that it can easily handle very large size databases and it can also process the complex ad hoc queries efficiently.
- The communications networks should be capable enough to transmit large amount of data and for this purpose the latest hardware and software can be utilized.

2.9.3 Implementation Consideration

Implementation of a data warehouse can be performed by combining all the associated products and objects of a data warehouse. Now to implement a data warehouse, the basic steps to build a data warehouse must be performed. These are stated below:

- At first, all types of information regarding the requirement of a data warehouse by an organization must be collected and analyzed. Then a data model for the data warehouse has to be configured.
- In the next step, various data sources for the data warehouse must be identified. Then an appropriate DBMS and hardware platform has to be selected for the data warehouse server.

- At first, data are collected from the operational databases and then data transformation and data cleaning operation are performed so that these data can be stored in the data warehouse database.
- Different tools and software are selected next for the proper functioning of the data warehouse. For example: database access tools, reporting tools, database connectivity software, data analysis software, presentation software.
- Finally, the data warehouse must be updated when it is required.

Some other implementation considerations to build a data warehouse are as follows:

- Selection of appropriate access tools is one of the important considerations to implement a data warehouse. At the moment, we do not have any tool that can be used for all types of data warehouse access. So the most appropriate set of tools is used for this purpose.
- Data collection, data transformation, data cleanup and data migration operations must be properly performed for successful implementation of a data warehouse.
- There are two data storage approaches available for data warehouse data. In the first approach, a separate storage media is used to store some amount of older warehouse data that are detailed and less important. The other warehouse data are stored in the bulk storage media and it is maintained by the data warehouse server.

In the second approach, data warehouse data are divided depending upon different requirements and data types. Then data are distributed to the related multiple servers. In this approach, metadata must be stored in one source and it must be maintained by a single server.

- Metadata is a very important part of any data warehouse. So metadata must be properly collected and maintained in the process of data warehouse building. The metadata must be accessible by all the data warehouse users so that they can use the data warehouse efficiently.

- A classification of data warehouse users has to be made for the proper use of a data warehouse. In general, the users are classified into the following categories depending upon their skills and access level of the warehouse.
  - 0 **Casual users:** Casual users can access formatted information from a data warehouse. These users can execute only those queries and reports which are already there in a data warehouse.
  - 0 **Power user:** Power user can create and execute simple ad hoc queries. These users can also examine the results of simple queries and reports. This kind of users requires access tools to perform their tasks.
  - 0 **Experts:** Expert users are capable of developing complex queries. These users can perform complex analysis on the data warehouse information. These users also require different tools to perform their jobs efficiently. They possess a good understanding about the data warehouse data and tools.

2.10 THE COST OF DATA WAREHOUSING

The cost to build a data warehouse is not found to be similar for all types of businesses and corporations. It varies depending upon different business environments and requirements of information. It is not possible to estimate the exact cost to build a data warehouse. The cost of data warehousing is affected by different factors. Most significant factors are discussed below:

- **Hardware Costs:** A compatible hardware platform is required to build an efficient data warehouse. Different types of hardware like storage media, CPUs, data communication infrastructures, workstations etc are required for building a data warehouse. So a significant amount of hardware cost can be estimated for building a data warehouse. On the other hand, for any data warehouse, amount of data and data usage may be increased day by day and it will also increase the hardware costs. It is also observed that the number of data warehouse users may affect the hardware cost of a data warehouse. If the number of users is increased then the hardware cost is also increased as the hardware requirement becomes more. Sometimes for better performance, additional or the latest technology hardware is required in data warehousing which increases its hardware cost.

- **Software cost:** The hardware platform of data warehousing require different softwares or software tools to perform their jobs efficiently. In recent times, there are various open source softwares available but in case of data warehousing, all possible warehousing tasks cannot be performed by using only open source software. So, a cost related to the required software or software tools is also associated with the process of building a data warehouse. On the other hand, software cost of data warehousing may increase due to the requirement of software maintenance. Software prices also may be increased in future. DBMS is one of the important softwares that is required in data warehousing.
- **Human resource cost:** Data warehousing always requires different types of users and skilled professionals to use the data warehouse for different business purposes and for regular maintenance of it. For example, at least one dedicated system manager, a software engineer and backend developers are required to support the warehouse database. We know that data warehouse must be updated regularly and some skilled professionals are also required to monitor and execute this job. So a cost has to be estimated for building data warehouse related to the employment of these human resources. Different types of trainings to improve efficiency of the human resources are also provided in every corporation. So a cost related to the user trainings is also associated to build a better data warehouse.

CHECK YOUR PROGRESS

Q.1: i) Which of the following is not true for data warehouse?

A) Data warehouse contains subject oriented data

B) Data warehouse contains only short duration recent data.

C) Data warehouse provides analytical information for decision making process.

D) Data warehouse is different from traditional database systems.

ii) Data model of ....is not required to normalize for maintaining ACID properties.

A) Informational data store B) Operational data store

C) DBMS D) Data warehouse

iii) OLAP server of a data warehouse is implemented in the .... tier of the three-tier structure.

A) Bottom tier B) Second tier

C) Middle tier D) Top tier

iv) Which of the following is not a basic component of data warehouse architecture?

A) Data warehouse database

B) Metadata

C) Data marts

D) Data warehouse users

v) Which of the following is not true for casual users of data warehouse?

A) Casual user cannot create queries.

B) Casual user can access formatted information from a data warehouse.

C) Casual users can perform complex analysis on the data warehouse information.

D) Both (A) and (B)

vi) Which of the following is a basic factor to estimate the cost of a data warehouse?

A) Hardware B) Software

C) Human resources D) All of the above

Q.2: State whether the following statements are true or false:

i) All data in a data warehouse must be stored with a common format.

ii) Data warehouse contains transactional data and it is termed as OnLine Transaction Processing (OLTP) system.

iii) Both operational data store and data warehouse contains subject oriented data.

iv) The number of concurrent users of informational data store is always very less than of the operational data store.

v) A data warehouse combines operational data stores and informational data store.

vi) Data marts are larger in size and less flexible than data warehouse.

2.11 LET US SUM UP

- A data warehouse is a collection of integrated data from different types of sources that help to prepare analytical reports and provide required analytical information to the decision making systems of a corporation.
- All data in a data warehouse must be stored with a common format. Due to the common data format, data warehousing can minimize the possibility of error in data interpretation and improve data consistency.
- The data warehouse stores data according to a particular time unit.
- DBMS contains transactional data and data warehouse contains analytical data.
- DBMS stores application oriented data and data warehouse stores subject oriented historical data received from multiple heterogeneous sources.

- In case of data warehouse, data are not allowed to be changed or modified in most of the times.
- Data warehouse does not require any recovery and concurrency control mechanism.
- Data warehousing provides easier data accessibility as data are stored separately in one place. It can store multidimensional data and provide support users to perform query analysis and analysis of stored data.
- Data warehouse maintains data security by providing secure access to stored data by authenticated users.
- In recent times, data warehousing is applied in different domains like banking and financial services, analysis of biological data, quality product manufacturing, retail sectors, E-commerce, telecom sectors etc.
- Operational data of different organizations are constructed by various on-line transaction processing operations of operational applications. These data are recent, complete, non redundant and modifiable data.
- ODS contains subject oriented data like data warehouse. Data in ODS is entirely integrated.
- The architectures of ODS and data warehouse are different.
- Informational data store is a collection of recent, summarized and redundant data about different subjects. Informational data are used to provide appropriate response to the different queries placed by corporate executives and managers for decision making purpose.
- Data model in case of an informational data store is not required to be normalized for maintaining ACID properties.
- Data accessing In case of informational data store is performed primarily on ad hoc basis.
- In case of informational data store, data are not updated regularly. In most of the times periodic and planned batch wise data updates are observed in informational data store.
- Two-tier architecture and three-tier architecture are the two popular data warehouse architecture.
- The data warehouse two-tier architecture is based on **client–server architecture.**

- Three-tier architecture consists of three layers that are bottom tier, middle tier and top tier.
- A data warehouse architecture contains seven basic components which are data warehouse database, tools for data sourcing, data cleanup, data transformation and data migration, metadata ,data marts, access Tools, data warehouse administration and management and information delivery system
- Metadata is the data about data that describes the data warehouse. Two classes of metadata are available in data warehousing that are technical metadata and business metadata.
- A data mart is a structure that stores data related to specific subject area and used by the specific group of users and departments of a corporation. Data marts are smaller in size and more flexible than data warehouse.
- Four types of access tools available in data warehousing which are query and reporting tools, application development tools, data mining tools, OLAP tools and executive information system tools.
- Two approaches are available to build data warehouse that are top down approach and bottom up approach.
- Data warehouse users are classified into three classes that are casual users, power user and experts.

2.12 FURTHER READING

1) Berson, A., & Smith, S. J. (1997). *Data Warehousing, Data Mining and OLAP*. McGraw-Hill, Inc.
2) Inmon, W. H. (2005). Building the Data Warehouse. John Wiley & Sons.
3) Ponniah, P. (2004). *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals*. John Wiley & Sons.

2.13 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) B; ii) A; iii) C; iv) D; v) C; vi) D

Ans. to Q. No. 2: i) True; ii) False; iii) True; iv) True; v) False; vi) False

2.14 MODEL QUESTIONS

Q.1: Write down the differences between data warehouse and DBMS.

Q.2: Why is data warehouse required by a corporation?

Q.3: Write down the similarities and differences between operational data store and data warehouse.

Q.4: Write down the characteristics of data warehouse.

Q.5: What is metadata? Why is it required?

Q.6: What is data mart?

Q.7: Write down the different design considerations for building a data warehouse.

Q.8: Write down different technical considerations for building a data warehouse.

Q.9: Write down different implement considerations for building a data warehouse.

*** ***** ***

UNIT 3: INTRODUCTION TO OLAP

UNIT STRUCTURE

3.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- describe the different uses of OLTP and OLAP
- define multi-dimensional data cube
- Describe the different operations of OLAP like Roll-Up, Slice, etc.
- Differentiate between OLTP and OLAP.

3.2 INTRODUCTION

In the previous unit, we have learned about the data base management system and different aspects of data warehouse. In this unit we will learn about OLAP and OLTP. In this unit, we will also learn how different OLAP operations like Roll-Up, Roll-down, Slice, Dice and Pivot can be performed on a multidimensional data base i.e data cube to analysis the data. In addition to this, we will also learn to differentiate OLTP and OLAP. In the next unit, we will explore the concept of data preprocessing in detail.

3.3 OLTP AND OLAP

OLTP (Online Transactional Processing) is a category of data processing that is focused on transaction-oriented tasks. OLTP is used for business task. It provides a multi-dimensional view of different business tasks. In OLTP, backup and recovery process is maintained religiously. Following are some examples of OLTP.

- Online Banking Transaction
- Online Shopping
- Booking Online ticket
- Order entry etc

OLTP applications typically possess the following characteristics:

- Transactions that involve small amounts of data
- Indexed access to data
- A large number of users
- Frequent queries and updates
- Fast response times

Online Analytical Processing (OLAP) allows the user to view and analyse data in multiple views. This analytical processing enables the user to select and view data from different points of view. OLAP is used to access live data online and to analyze it. It also allows to process data in multi dimensions.

3.4 OLAP OPERATION

There are different types of OLAP operations. They are listed below.

- Roll-Up
- Roll-down
- Slice
- Dice
- Pivot

The above mentioned OLAP operations can be explained with the help of a data cube **C** which is used to store the sales data of an electronic enterprise. Thus the three dimensions are **Year, Items and City.** These dimensions allow to keep the record of monthly sales of different products or sales of products from different cities in different years etc. as given in figure 3.1.

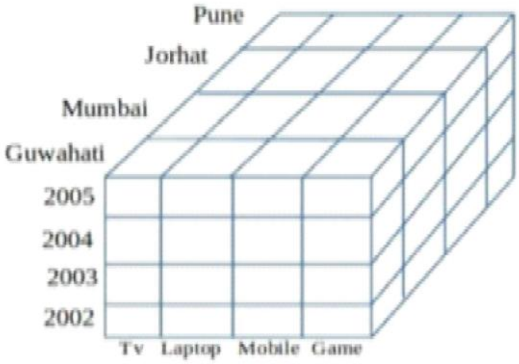


Figure 3.1: Data Cube C[Year, Items, City]

3.4.1 Roll-Up or Drill Up

The roll-up operation performs aggregation on a data cube either by climbing up the hierarchy or by dimension reduction. The figure 3.2 shows the result of **Roll up** operation by climbing up the hierarchy of locations that is **city** to state. In other words, we can say that resulting cuboid group the data by **State** rather than city.

Roll-up C[Year, Items, City] = C[Year, Items, State]

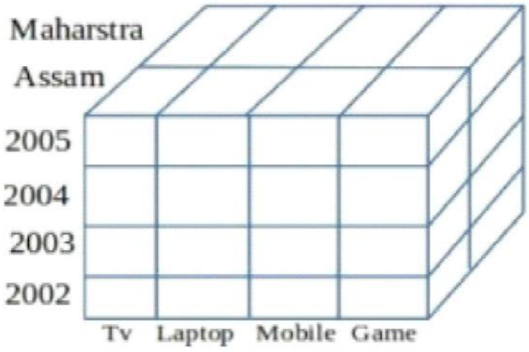


Figure 3.2: Roll-up Operation

Here, each data cell of the cuboid is the aggregation of those data cell that are merged due to roll-up operation. In other word, the result stored in the data cell C[2005, Tv, Assam] is the sum of the data stored in the data cell (figure 3.1) C[2005, Tv, Guwahati] and c[2005, Tv, Jorhat].

When roll-up is performed by dimension reduction, then one or more dimensions from the data cube are removed.

3.4.2 Roll-down or Drill-down

This operation is opposite to roll-up operation. This operation can be performed by stepping down the hierarchy or by adding new dimension. The drill-down operation is concerned with switching from aggregation to more details. The following figure 3.3 shows the result of **Roll-down** operation by stepping down the hierarchy of time that is **Year** to **Month**. In other word we can say that resulting cuboid group the data by **Month** rather than **Year**.

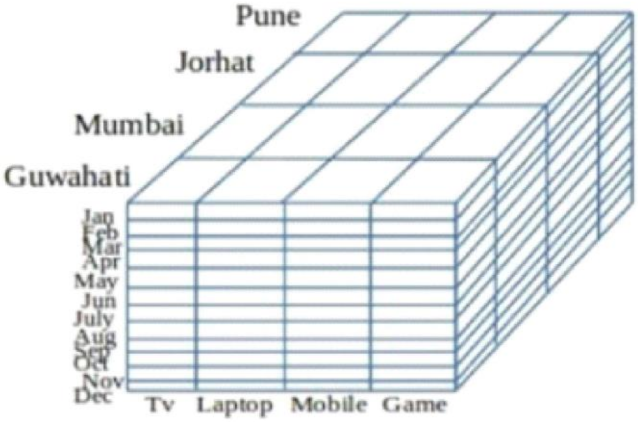


Figure 3.3: Drill-down operation

3.4.3 Slicing

The slice operation selects one particular dimension from the given cube and produces a new sub cube.

The following figure 3.4 shows the slicing operation. Here, the slicing operation is performed for the dimension **Year** using the criterion **Year=‘‘2004"**

slice$_{\text{Year=2004}}$ C[Year, Items, City]= C[Items, City]

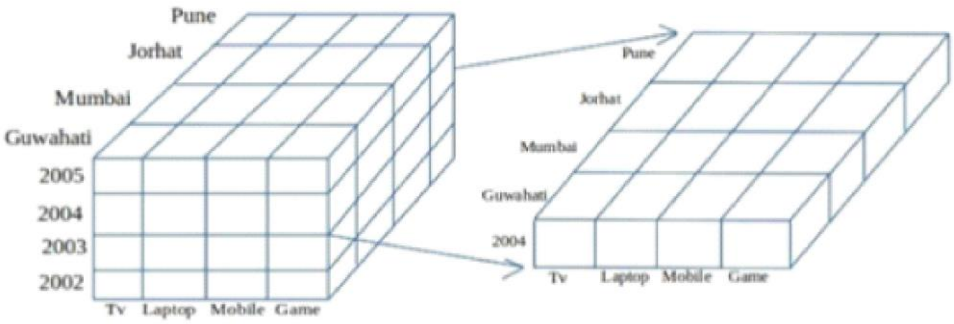


Figure 3.4: Slicing Operation

Each data cell of the resultant sub cube will contain city wise details of all items for the year 2004.

3.4.4 Dicing

The dicing operation is for selecting a smaller cube and it analyzes the cube from different perspectives. The dicing selects two or more dimensions from the original cube and produces a sub cube. The following figure shows the dicing operation. Here, the dicing operation is performed in the following criterion, which involves two dimensions **City** and **Year**.

Dice $_{\text{year="2003" or "2004 and city= "Guwahati" or "Jorhat"}}$ **C[Year, Item, City] = C[Year', Item, City']**

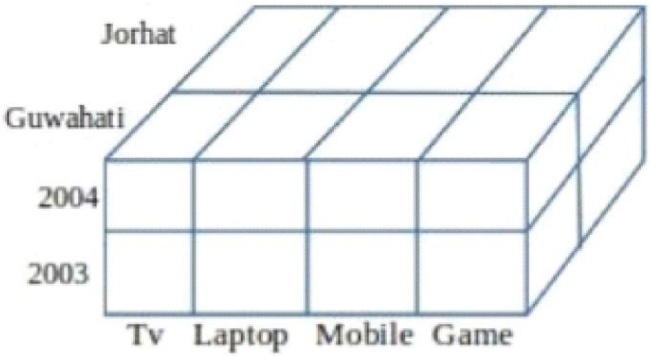


Figure 3.5: Dicing Operation

3.4.5 Pivot

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data.

CHECK YOUR PROGRESS

Q.1: State whether the following statements are true (T) or false (F)

i) Roll up operation performs aggregation on data cube.

ii) Roll-down operation can be performed by stepping up the hierarchy or by deleting an existing dimension.

iii) Slice operation selects a particular cell from a data cube and produce a new data cube.

iv) Dicing selects two or more dimensions from the original cube and produces a sub cube.

Q.2: Fill in the blanks:

i) The drill-down operation is concerned with switching from .... to more details.

ii) Pivot operation.............................. the data axes in view in order to provide an alternative presentation of data.

3.5 OLAP VERSUS OLTP

The differences between OLAP and OLTP are listed below.

Sl. No.	Data Warehouse (OLAP)	Operational Database (OLTP)
1	Involves historical processing of information.	Involves day-to-day processing.
2	OLAP systems are used by knowledge workers such as executives, managers and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
3	Useful in analyzing the business.	Useful in running the business.
4	It focuses on Information out.	It focuses on Data in.
5	Based on Star Schema, Snowflake, Schema and Fact Constellation Schema.	Based on Entity Relationship Model.
6	Contains historical data.	Contains current data.
7	Provides summarized and consolidated data.	Provides primitive and highly detailed data.
8	Provides summarized and multidimensional view of data.	Provides detailed and flat relational view of data.
9	Number or users is in hundreds.	Number of users is in thousands.
10	Number of records accessed is in millions.	Number of records accessed is in tens.

11	Database size is from 100 GB to 1 TB	Database size is from 100 MB to 1 GB.
12	Highly flexible.	Provides high performance.

3.6 LET US SUM UP

- OLTP is an application oriented that provides a multi-dimensional view of different business tasks.
- OLAP is an analytical processing that enables user to select and view data from different point of view.
- Different types of OLAP operations are Roll-Up, Roll-down, Slice, Dice and Pivot.
- Aggregation on a data cube is performed by the roll-up operation.
- The drill-down operation is concerned with switching from aggregation to more details.
- To produce a new sub cube by selecting one particular dimension we can perform Slice operation.
- The dicing selects two or more dimensions from the original cube and produces a sub cube.
- Pivot operation rotates the data axes in view in order to provide an alternative presentation of data.

3.7 FURTHER READING

1) Pujari, A. K. (2001). *Data Mining Techniques*. Universities Press.

2) Tan, P. N., Steinbach, M. & Kumar, V. (2013). *Data Mining Cluster Analysis: Basic Concepts and Algorithms. Introduction to Data Mining*.

3.8 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) True; ii) False; iii) False; iv) True

Ans. to Q. No. 2: i) aggregation; ii) rotate

3.9 MODEL QUESTIONS

Q.1: What is OLTP?

Q.2: What are the different characteristics of OLTP?

Q.3: What is OLAP?

Q.4: What are the different OLAP operations?

Q.5: Explain the different OLAP operations with a suitable example.

Q.6: How is the **Roll-Up** operation different from **Roll-Down** operation?

Q.7: Differentiate between **Slicing** and **Dicing** operation.

Q.8: Differentiate between OLAP and OLTP.

*** ***** ***

UNIT 4: DATA PREPROCESSING

UNIT STRUCTURE

4.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- describe the different stages of data processing
- define data summarization and data cleaning
- explain data transformation
- describe data reduction.

4.2 INTRODUCTION

In the previous unit, we have learned about OLAP and OLTP. We have also learned how different OLAP operations like Roll-Up, Roll-down, Slice, Dice and Pivot can be performed on a multidimensional data base. In this unit, we will learn about data preprocessing as well as the different stages involved in data preprocessing. Here, we will discuss different data processing techniques involved in data mining such as data cleaning, data transformation, data reduction. We will also discuss about concept hierarchies in this unit. In the next unit, we will explore the concept of multidimensional data in the form of data cube in detail along with different data warehouse schema's.

4.3 DATA PREPROCESSING

In data mining, data preprocessing is a technique which involves transformation of raw data into an understandable format. Data preprocessing is a proven method of resolving real-world data's issues such as inconsistent, incomplete, and/or lacking in certain behaviors or trends.

There are six stages involved in data processing:

- **Data Collection:** Data collection is the first step that is involved in data processing technique. Data is collected from available trustworthy and well-built available resources that include data lakes and data warehouse.
- **Data Preparation:** Data preparation stage is the second stage involved in data processing. Data preparation is also referred to as "pre-processing". At this stage, the raw data that is collected in first stage is cleaned up and organized for the following stage of data processing. During preparation, raw data is checked for any kind of errors and this stage also eliminates bad data (redundant, incomplete, or incorrect data) and begins to create high-quality data for the best business intelligence.
- **Data Input:** The clean data is then entered into its destination and translated into a language that it can understand. Data input is the first stage in which raw data begins to take the form of usable information.
- **Processing:** During this stage, the data input to the computer in the previous stage is actually processed for interpretation. Processing is done using machine learning algorithms, although the process itself may vary slightly depending on the source of data being processed (data lakes, social networks, connected devices etc.) and its intended use (examining advertising patterns, medical diagnosis from connected devices, determining customer needs, etc.).
- **Data Output/Interpretation:** The output/interpretation stage is the stage at which data becomes finally usable to non-data scientists. It is translated, made readable, and is often in the form of graphs, videos,

images, plain text, etc. Members of the company or institution can now begin to self-serve the data for their own data analytics projects.

- **Data Storage:** The final stage of data processing is storage. After all the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on. Besides, properly stored data is a necessity for the sake of compliance with data protection legislation like GDPR. When data is properly stored, it can be quickly and easily accessed by members of an organization when needed.

4.4 DATA SUMMARIZATION

Summarization is a key data mining concept which involves techniques for finding a compact description of a dataset. Simple summarization methods such as tabulating the mean and standard deviations are often applied for exploratory data analysis, data visualization and automated report generation.

CHECK YOUR PROGRESS

Q.1: Fill in the blanks:

i) Data preparation is also known as ....

ii) .... is the first stage of data processing in which raw data begins to take the form of usable information.

iii) Summarization involves techniques for finding a compact description of a ....

4.5 DATA CLEANING

For decision making we need data warehouse and it is very essential that the data in data warehouse be correct. Since large volume of data are collected from heterogeneous sources, so there is a high chance of having error in data. Therefore, to construct an error free and high-quality data warehouse data cleaning is essential. Data cleaning technique includes:

- using transformation rules, e.g., translating attribute name like 'age' to 'DOB'
- using domain-specific knowledge.
- performing parsing and fuzzy matching, e.g., for multiple data sources, one can designate a preferred source as a matching standard, and
- auditing, i.e., discovering facts that flag unusual patterns.

4.6 DATA TRANSFORMATION

Data transformation is a process of transforming heterogeneous data that are collected from different data sources to an uniform structure so that data can be combined and integrated.

Data transformation operations would contribute toward the success of the mining process.

- **Smoothing:** It helps to remove noise from the data.
- **Aggregation:** Summary or aggregation operations are applied to the data. i.e., the weekly sales data is aggregated to calculate the monthly and yearly total.
- **Generalization:** In this step, Low-level data is replaced by higher-level concepts with the help of concept hierarchies. For example, the city is replaced by the county.
- **Normalization:** Normalization is performed when the attribute data are scaled up or scaled down. Example: Data should fall in the range -2.0 to 2.0 post-normalization.
- **Attribute Construction:** these attributes are constructed and included in the given set of attributes which are helpful for data mining.

4.7 DATA REDUCTION

Data reduction is a technique that is applied to a data warehouse to obtain a reduced representation of the data set that is much smaller in volume, yet it closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient and yet it must produce the same (or almost the same) analytical results.

There are several data reduction strategies. Those are shown below:

- **Data Cube Aggregation:** Aggregation operations are applied to the data in the construction of a data cube.
- **Dimensionality Reduction:** In dimensionality reduction redundant attributes are detected and removed. This reduces the data set size.
- **Data Compression:** Encoding mechanisms are used to reduce the data set size.
- **Numerosity Reduction:** In numerosity reduction, the data are replaced or estimated by alternative.
- **Concept Hierarchy Generation:** In concept hierarchy, the raw data values for attributes are replaced by ranges or higher conceptual levels.

4.8 CONCEPT HIERARCHIES

Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts with higher-level concepts.

In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides the users with the flexibility to view the data from different perspectives.

Data mining on a reduced data set means fewer input/output operations and it is more efficient than mining on a larger data set.

Because of these benefits, discretization techniques and concept hierarchies are typically applied before data mining rather than during mining.

CHECK YOUR PROGRESS

Q.2: Fill in the blanks:

i) ....is a process of transforming heterogeneous data.

ii)............................. helps to remove noise from the data.

iii)............................. technique is applied to a data warehouse to obtain a reduced representation of the data set.

iv) .... can be used to reduce the data by collecting and replacing low-level concepts with higher-level concepts.

4.9 LET US SUM UP

- Data preprocessing is a technique which involves transformation of raw data into an understandable format.
- Six stages involved in data processing are Data Collection, Data Preparation, Data input, Processing, Data output, Storage.
- Summarization is a technique for finding a compact description of a dataset.
- Data cleaning is essential to construct an error free and high-quality data warehouse.
- Data transformation is a process of transforming heterogeneous data to an uniform structure.
- Normalization is performed when the attribute data are scaled up or scaled down.
- **Data reduction is** a technique that is applied to a data warehouse to obtain a reduced representation of the data set.
- In dimensionality reduction redundant attributes are detected and removed. This reduces the data set size.
- Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts with higher-level concepts.

4.10 FURTHER READING

1) Pujari, A. K. (2001). *Data Mining Techniques*. Universities Press.
2) Tan, P. N., Steinbach, M. & Kumar, V. (2013). *Data Mining Cluster Analysis: Basic Concepts and Algorithms. Introduction to Data Mining*.

4.11 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) Preprocessing; ii) Data input; iii) Data set

Ans. to Q. No. 2: i) Data transformation; ii) Smoothing; iii) Data reduction; iv) Concept hierarchies.

4.12 MODEL QUESTIONS

Q.1: What is data preprocessing? What are the different stages involve in data preprocessing?

Q.2: What is data summarization?

Q.3: Why is data cleaning important? What are the different data cleaning techniques?

Q.4: How do data transformation operations contribute toward the success of the mining process?

Q.5: What is data reduction?

Q.6: What are the different data reduction strategies?

Q.7: Explain the concept hierarchies.

*** ***** ***

UNIT 5: MULTIDIMENSIONAL DATA

UNIT STRUCTURE

5.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- describe the concept of data model and multidimensional representation of data
- define data cube of multidimensional data representation
- explain the concepts of dimension modelling
- describe the components of multidimensional data model
- describe the different data warehouse schema such as star, snowflake and fact constellation.

5.2 INTRODUCTION

We are already familiar with the concepts of data warehousing and OLAP (Online Analytical Processing). We know that the core of the design of the data warehouse lies in a multidimensional view of the data model. A data model is a description of the organization or the structure of data in an information system. Data warehouse users explore data to find useful patterns by studying how certain attributes of data elements (i.e., *measures*) are related to other attributes (i.e., *dimensions*). Initially, the user has to specify which attributes of the original data is to be treated as measures and which to treated as dimensions. The data is conceptually organized as multi-dimensional array, where each dimension corresponds to a dimension of the warehouse, and the values stored in each cell of the array corresponds to the measure of the warehouse.

In this unit, we will learn about multidimensional view of data, data cubes and different data warehouse schema such as *star schema*, *snowflake schema* and *fact constellation*. In the next unit we will explore data warehouse architecture, data warehouse design, OLAP three-tier architecture, indexing and querying in OLAP, OLAM etc.

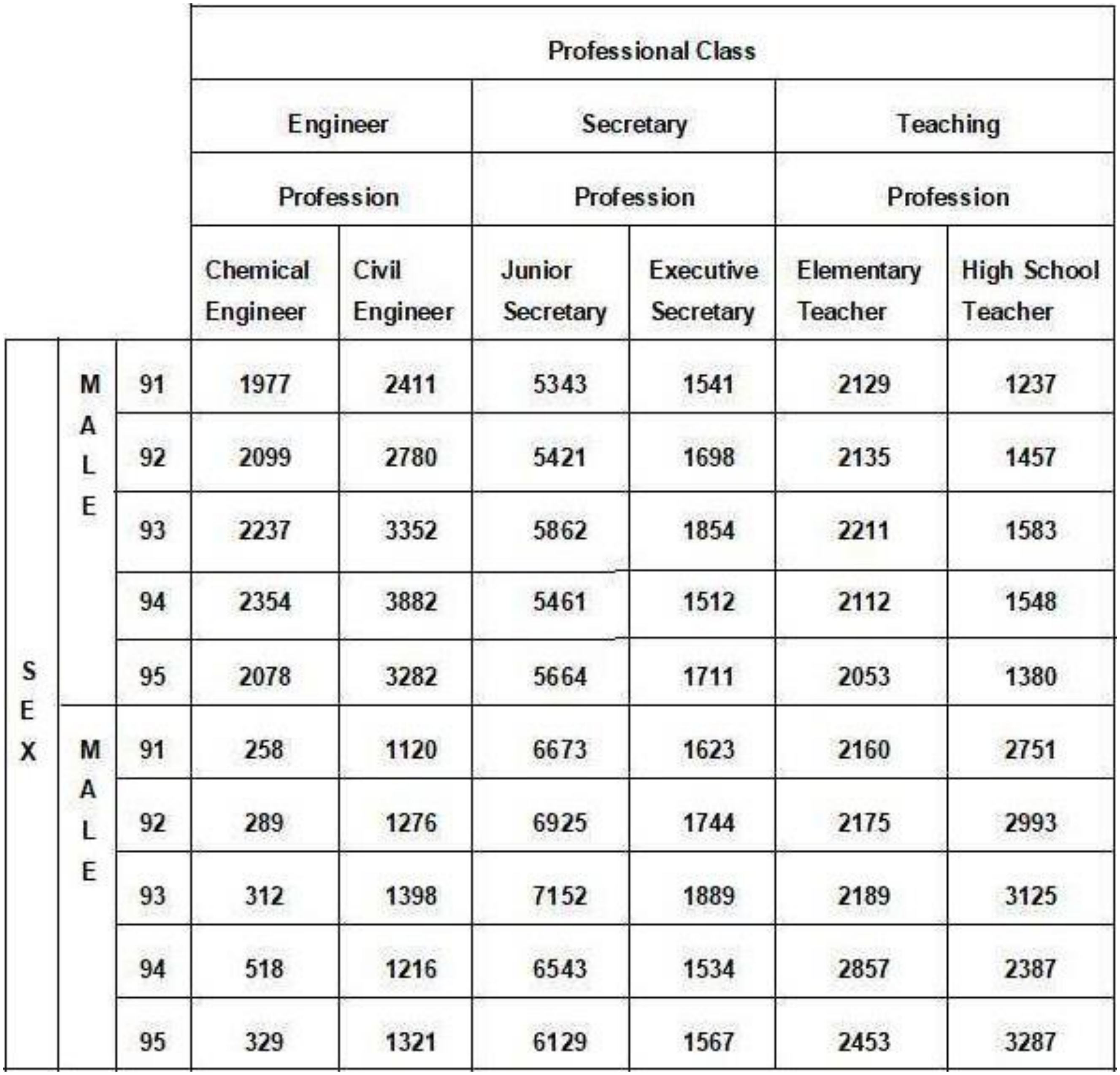
5.3 DATA CUBE

Multidimensional data model stores data in the form of data cube. To understand the concept of multidimensional view of data, let us take the data set represented in 2-D in the following table 5.1 (*example is taken from S Choudhury, 2009*) for an employment data warehouse *"employment in California"* in order to keep records of the employment details with respect to the dimensions *sex, year* and *profession.*

A data cube allows data to be modelled and viewed in multiple dimensions. An OLAP data cube is also known as *hypercube*. It is defined by *dimensions* and *facts*. *Dimensions* are the perspectives or entities with respect to which an organization wants to keep records. For example, in *Employment in California*, there are three dimensions namely *sex, year* and *profession*. Each dimension may have a table associated with it, called a *dimension table*. For example, a dimension table for *sex* may contain the attributes item *male* and *female*

A multidimensional data model is typically organized around a central theme, like *employment*, for instance. This theme is represented by a fact table. Facts are numerical measures by which it analyzes relationships between dimensions. Examples of facts for above example data warehouse may include *total male employees*, *total civil engineers and total employees* *in a year etc.* The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables.

Table 5.1: Statistical Table: Two-dimensional Representation



			Professional Class					
			Engineer		Secretary		Teaching	
			Profession		Profession		Profession	
			Chemical Engineer	Civil Engineer	Junior Secretary	Executive Secretary	Elementary Teacher	High School Teacher
SEX	MALE	91	1977	2411	5343	1541	2129	1237
		92	2099	2780	5421	1698	2135	1457
		93	2237	3352	5862	1854	2211	1583
		94	2354	3882	5461	1512	2112	1548
		95	2078	3282	5664	1711	2053	1380
	MALE	91	258	1120	6673	1623	2160	2751
		92	289	1276	6925	1744	2175	2993
		93	312	1398	7152	1889	2189	3125
		94	518	1216	6543	1534	2857	2387
		95	329	1321	6129	1567	2453	3287

The 3-D data cube representation of the above data set information in table 5.1.; can be represented as shown below in the figure: 5.1. (Example is *taken from [A K Pujari, 2009]).*

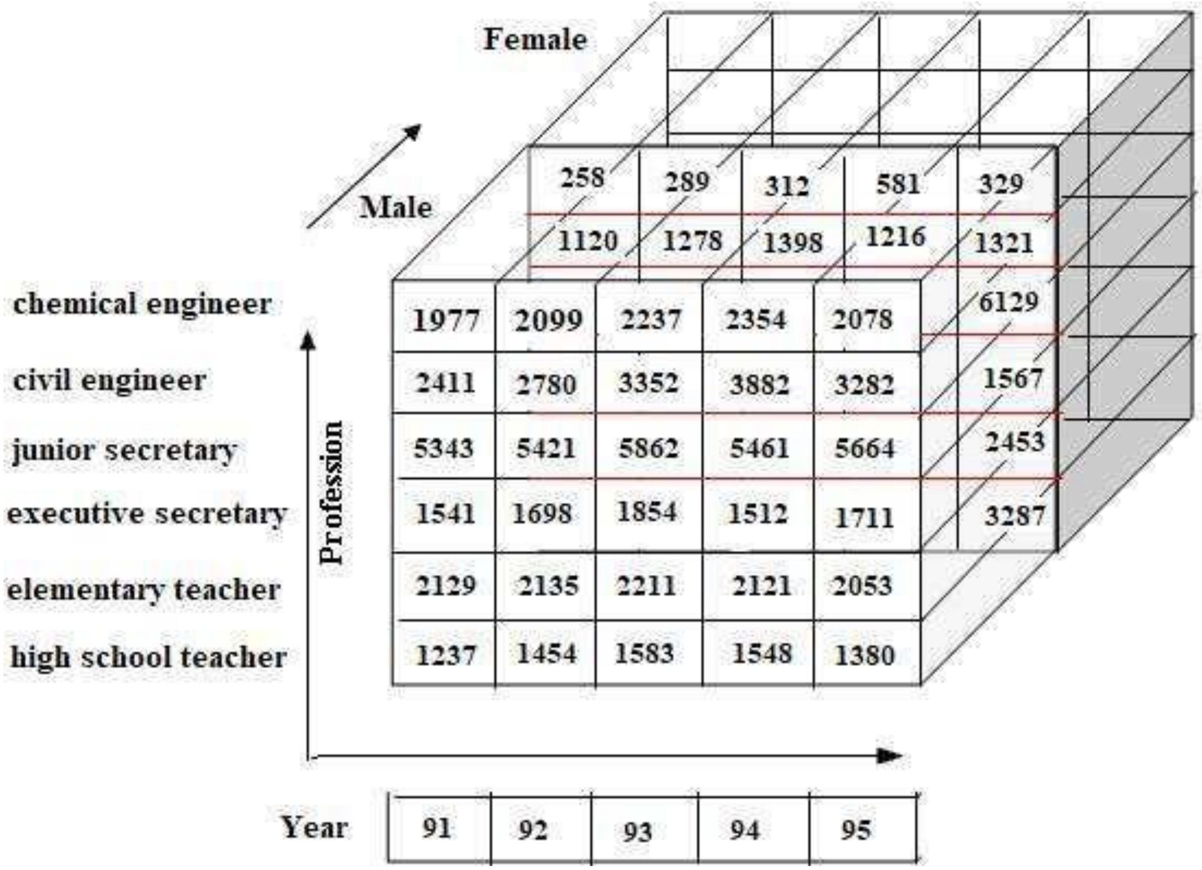


Figure: 5.1. Multidimensional representation of data (data cube)

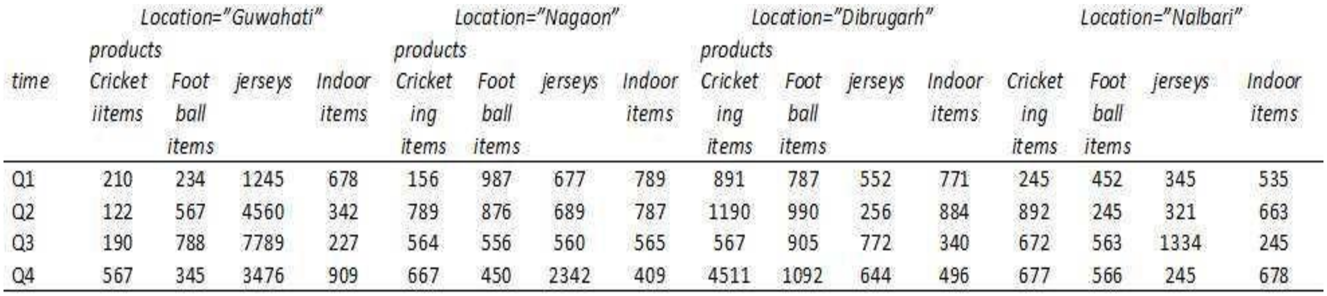
5.4 LATTICE OF CUBOIDS

The data cube is a metaphor for multidimensional data storage. It helps to represent the dimension hierarchy in the multidimensional view of data i.e. multidimensional data can be represented as a *lattice of cuboids*. The important thing to remember is that data cubes are n-dimensional and do not confine data to 3-D. The cuboid that holds the lowest level of summarization is called the *base cuboid (n-D cuboid)* and it consists of all the data cells. Any n-D data can be displayed as a series of (n”1)-D cubes which are obtained by grouping the cells and computing the numeric measures (facts) of all n-dimensions.

The cuboid consisting of one cell with numeric measures of all *n* dimensions holds the highest level of summarization. It is called the *apex cuboid* (*0-D cuboid)*. All the other cuboids lie between the base cuboid and apex cuboid in the lattice of cuboids.

Let us take another example of a data warehouse for *All Sports* to keep records of the store's sales with respect to the dimensions *time*, *product*, *branch*, and *location*. These dimensions allow the store to keep track of things like monthly sales of products and the branches and locations at which the products were sold.

Table 5.2: A 3-D view of sales data for *All Sports*, according to the dimensions *time*, *product* and *location*. The measure displayed is rupees (in thousands)



	Location="Guwahati"				Location="Nagaon"				Location="Dibrugarh"				Location="Nalbari"			
	products				products				products							
time	Cricket iitems	Foot ball items	jerseys	Indoor items	Cricket ing items	Foot ball items	jerseys	Indoor items	Cricket ing items	Foot ball items	jerseys	Indoor items	Cricket ing items	Foot ball items	jerseys	Indoor items
Q1	210	234	1245	678	156	987	677	789	891	787	552	771	245	452	345	535
Q2	122	567	4560	342	789	876	689	787	1190	990	256	884	892	245	321	663
Q3	190	788	7789	227	564	556	560	565	567	905	772	340	672	563	1334	245
Q4	567	345	3476	909	667	450	2342	409	4511	1092	644	496	677	566	245	678

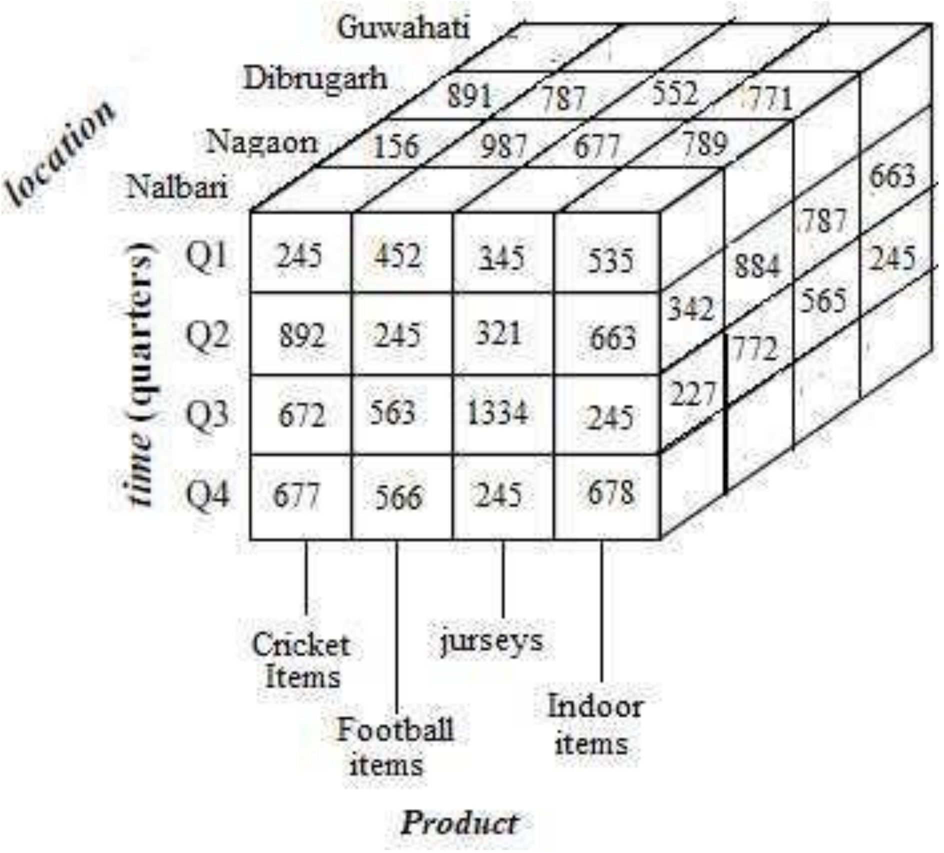


Figure 5.2: A 3-D data cube representation of the data in table 5.2, according to the dimensions *time*, *product* and *location*. The measure displayed is rupees (in thousands)

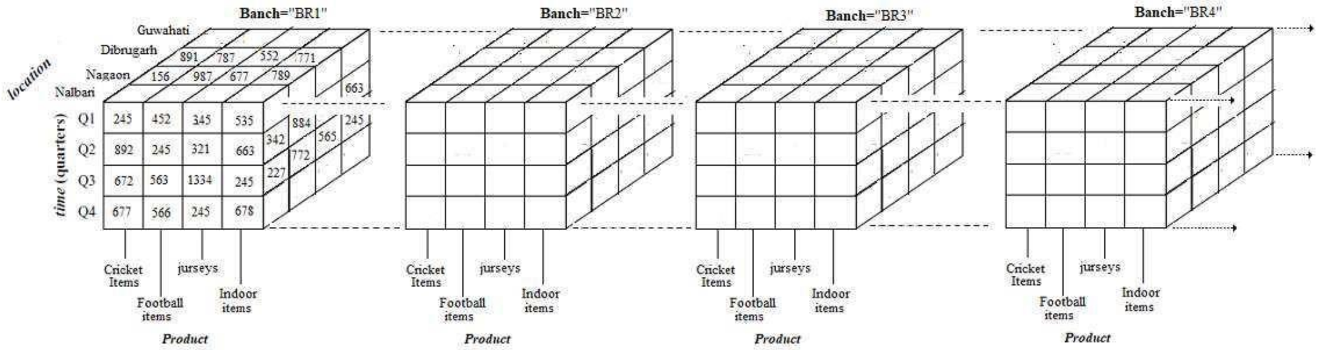


Figure 5.3: A 4-D data cube representation of sales data, according to the dimensions *time, product, location, and Branch.* The measure displayed is Rupees sold (in thousands). For improved readability, only some of the cube values are shown

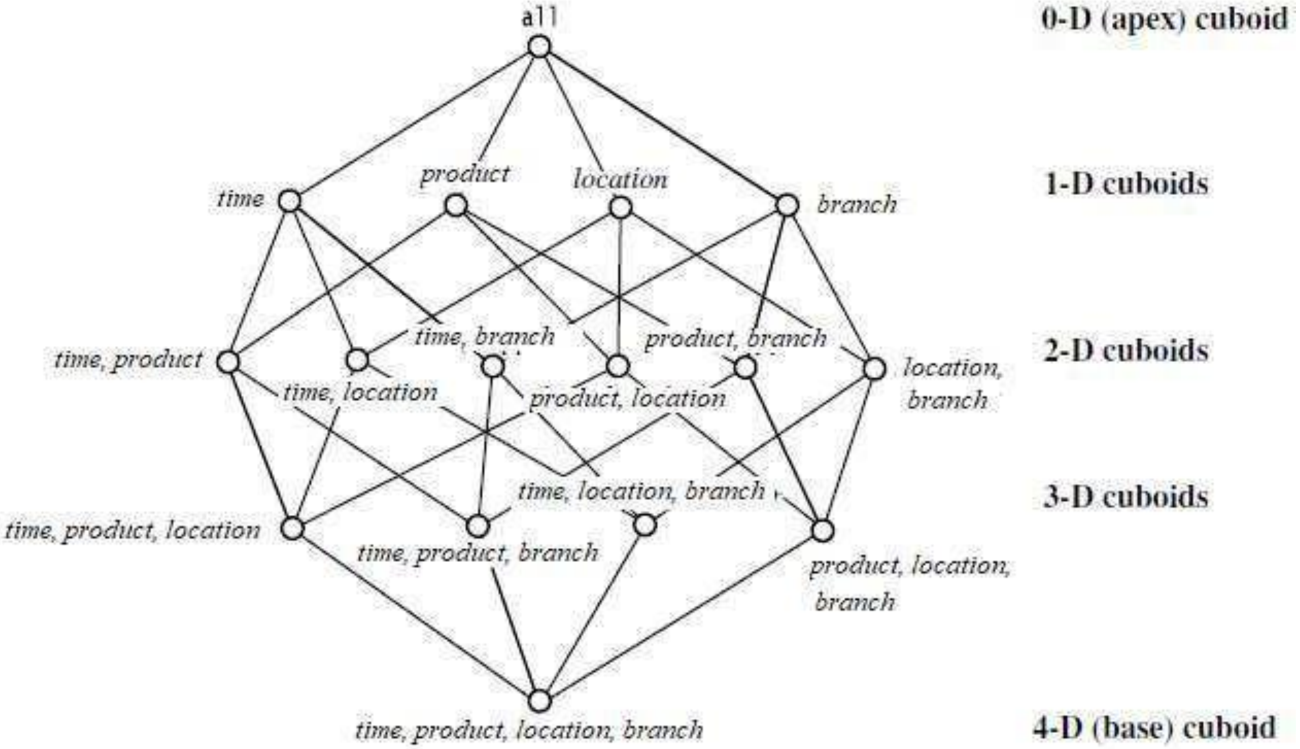


Figure 5.4: Lattice of cuboids, making up a 4-D data cube for the dimensions time, product, location, and branch. Each cuboid represents a different degree of summarization.

CHECK YOUR PROGRESS

Q.1: The core of the multidimensional model is the . ..., which consists of a large set of facts and a number of dimensions.

a) Multidimensional cube b) Dimensions cube

c) Data cube d) Data model

Q.2: Which of the following is not a kind of data warehouse application?

a) Information processing b) Analytical processing

c) Data mining d) Transaction processing

Q.3: Data cube can grow n number of dimensions, thus becoming:

a) dimensional cube b) solid cubes

c) star cubes d) hyper cubes

Q.4: Which of the following describes the data contained the data warehouse?

a) Relational data b) Meta data

c) Informational data d) Operational data

Q.5: Data that can be modelled as dimension attributes and measured attributes are called as ....

a) Multidimensional b) Single dimensional

c) Measured data d) dimensional data

Q.6: OLAP stands for–

a) Online analysis processing

b) Online transaction processing

c) Online analytical processing

d) Online aggregate processing

5.5 DATA WAREHOUSE SCHEMA

Multidimensional schema is especially designed to model data warehouse systems. Schema is a logical description of the entire database. The entity-relationship (ER) data model is generally used in the design of relational databases. A database schema consists of a set of entities and the relationships between them. It includes the name and description of records of all record types including all associated data-items and aggregates. Similar to databases, data warehouses also need to maintain a schema. There are various types of data warehouse schema; they are: *Star schema*, *Snowflake schema*, and *Fact Constellation schema*.

5.5.1 Star Schema

The star schema architecture is the simplest data warehouse schema and is widely used to develop or build a data warehouse and dimensional data marts. This schema is widely used to develop or build a data warehouse and dimensional data marts. It is called a star schema because the diagram resembles a star. It consists of a single large central *fact table* and a set of smaller *dimension table,* one for each dimension. The fact table contains the detailed summary data with no redundancy. It's primary key has one key per dimension. Each dimension table is joined with the fact table using a primary or foreign key. The fact table contains the *fact or subject* of interest for each corresponding dimension in the dimension table. It also stores numerical measures for those co-ordinates which are non-dimensional attributes. The relationship between fact table and dimensional table is 1:N. An example of star schema is shown in figure 5.5.

In the star schema shown in figure 5.5, **SALES** is a fact table having attributes *product_id, time_id, customer_id, employee_id* and *location_id* which references to the dimension tables *product, time, customer, employee* and *location* respectively. It also contains two numerical measures: *price_sold* and *qty_sold*. **Employee** dimension table contains the attributes: *employee_id, first_name, mid_name, last_name* and *dob*. **Time** dimension table contains attributes: *time_id, day, month, quarter* and *week***.** **Location** dimension table contains attributes: *location_id, district, street_number* and *city_name.* **Product** dimension table contains the attributes: *product_id, product_name, product_type, category* and *unit_price.* **Customer dimension** table contains the attributes: *customer_id, customer_fname, customer_lname, city, state, pin_no* and *contact_no.*

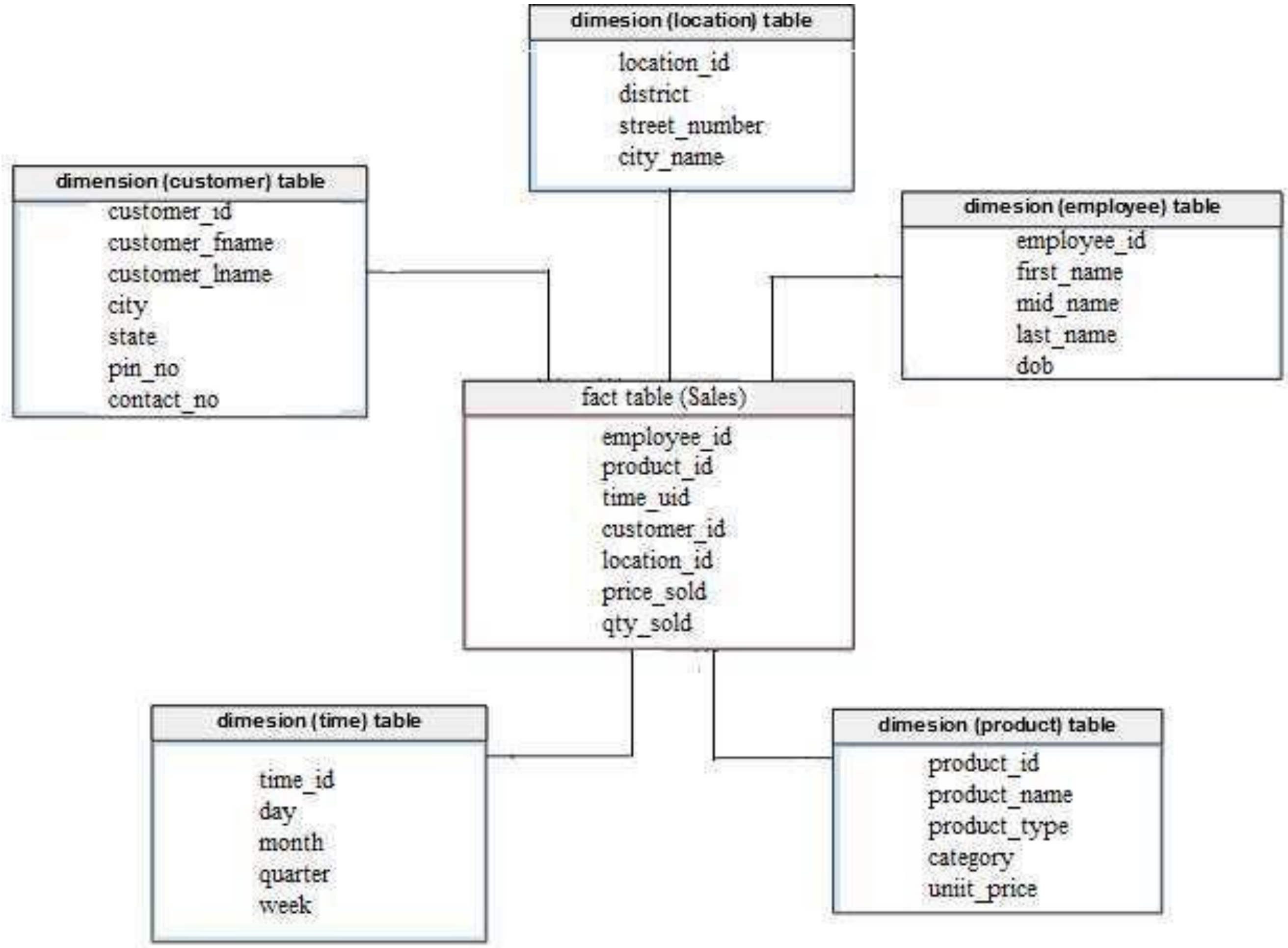


Figure 5.5: Star Schema

Advantages of Star Schema Data Warehouses:

- Easy to understand
- Easy to define dimension hierarchies
- Reduces number of physical joins
- Easy to maintain
- Used very simple metadata

Disadvantages of Star Schema Data Warehouses:

- Data integrity is not enforced well since it is in a highly de-normalized state

- Is not flexible in terms of analytical needs
- Normally do not reinforce many-to-many relationships within business entities.

EXERCISE 5.1

Draw a Star schema for a Library management Data warehouse.

5.5.2 Snowflake Schema

The *snowflake* schema is more complex compared to star schema because the dimension tables of the snowflake are normalized to support attributed hierarchies. A sample snowflake schema is shown in figure 5.6.

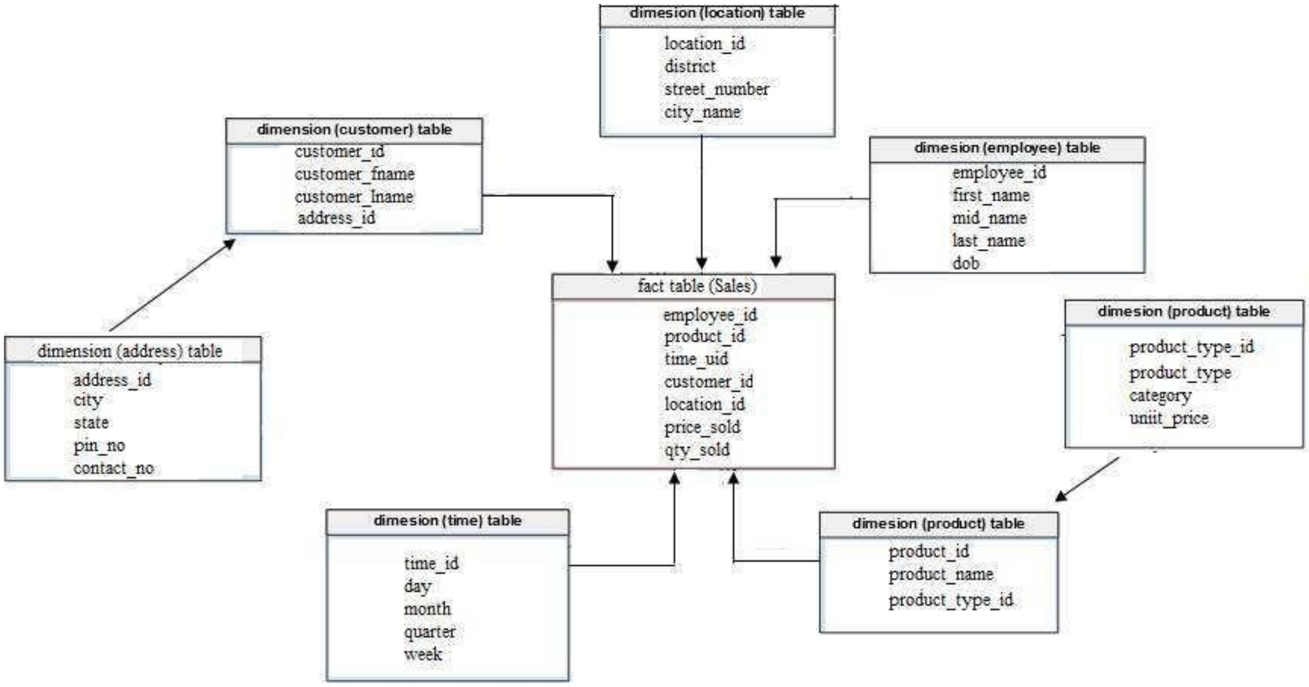


Figure 5.6: Snowflake Schema

The snowflake schema consists of a centralized fact table which is connected to multiple dimension tables and the dimension tables can be normalized into additional dimension tables. Similar to star schema, the fact table in snowflake schema also contains the *fact or subject* of interest for each corresponding dimension in the dimension table and stores numerical measures for those co-ordinates as well. In contrast to star schema, the dimension table in snowflake schema is normalized.

Advantages of Snowflake Schema Data Warehouses:

- Easy to maintain because dimension tables are normalized and normalizing results in saving storage spaces
- Reduces redundant information storage

Disadvantages of Snowflake Schema Data Warehouses:

- Increase in large number of join operations.

EXERCISE 5.2

Draw a Snowflake schema for a Hospital management Data warehouse.

5.5.3 Fact Constellation

Fact constellation schema is more complex than star or snowflake schema. It contains more than one fact table which share

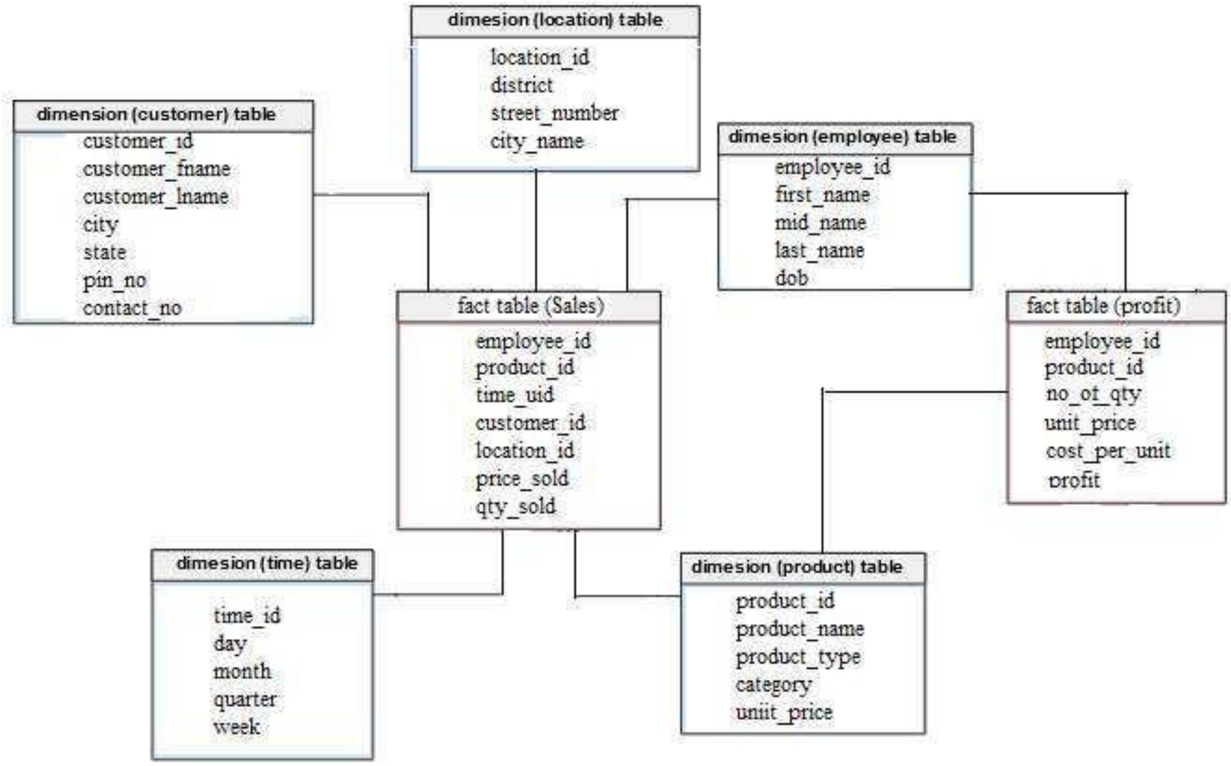


Figure 5.7: Fact Constellation Schema

some dimension tables. It is also referred to a galaxy schema. A sample fact constellation schema is shown in figure 5.7. There are two fact tables *Sales* and *Profit* which share the same dimension tables *employee* and *product*.

Advantages of Fact Constellation Schema Data Warehouses:

- Provides a more flexible schema compared to other schemas
- Different fact tables are explicitly assigned to the dimensions

Disadvantages of Fact Constellation Schema Data Warehouses:

- Hard to maintain
- More Complexity involved due to the involvement of more number of aggregations.

CHECK YOUR PROGRESS

Q.7: Which is a good alternative to the star schema–

a) Star schema b) Snowflake schema

c) Fact constellation d) Star-snowflake schema

Q.8: The type of relationship in star schema is ....

a) many to many b) one to many

c) many to one d) many to many

Q.9: Which statement best describes fact table?

a) fact table describes the transactions stored in a DW

b) fact table is the main store of the descriptions of the transactions

c) fact table describes the granularity of data in a DW

d) fact table is the main store of all of the recorded transactions over time

Q.10: Which of the following is the numeric measurements or values that represents a specific business aspects or activity:

a) dimensions b) schemas

c) facts d) tables

Q.11: Fact tables in a Data Warehouse is:

a) partially normalized b) completely denormalized

c) completely normalized d) partially denormalized

Q.12: In which of the following, a fact table in the centre is directly linked with dimension table?

a) star schema b) snowflake schema

c) fact constellation schema d) relational schema

5.6 LET US SUM UP

- In data warehouse, data is viewed as multidimensional data model which stores data in the form of data cube.
- A data cube allows data to be modelled and viewed in multiple dimensions.
- An OLAP data cube is defined by *dimensions* and *facts*.
- The Dimensions are the perspectives or entities with respect to which an organization wants to keep records and Facts are numerical measures by which it analyzes relationships between dimensions.
- There are various types of data warehouse schema; they are: *Star schema*, *Snowflake schema*, and *Fact Constellation schema*.
- A star schema consists of a single large central *fact table* and a set of smaller *dimension table,* one for each dimension.
- The snowflake schema consists of a centralized fact table which is connected to multiple dimension tables.
- The Fact constellation schema; know as galaxy contains more than one fact table which share some dimension tables.

5.7 FURTHER READING

1) Pudi, V., Krishna R. P. (2008) *Data Mining*. Oxford University Press.

2) Pujari, A. K. (2001). *Data Mining Techniques*. Universities Press.

5.8 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1:	(a)	**Ans. to Q. No. 2:**	(d)
Ans. to Q. No. 3:	(d)	**Ans. to Q. No. 4:**	(b)
Ans. to Q. No. 5:	(a)	**Ans. to Q. No. 6:**	(c)
Ans. to Q. No. 7:	(c)	**Ans. to Q. No. 8:**	(b)
Ans. to Q. No. 9:	(d)	**Ans. to Q. No. 10:**	(c)
Ans. to Q. No. 11:	(c)	**Ans. to Q. No. 12:**	(a)

5.9 MODEL QUESTIONS

Q.1: What is a data model?

Q.2: What is multidimensional data model?

Q.3: What is data cube? What do you mean by lattice of cuboids?

Q.4: What is fact table? How it is related to dimension table?

Q.5: What is data warehouse schema? What are the different types of data warehouse schema?

Q.6: Explain star schema. State the advantages and disadvantages of star schema?

Q.7: Briefly describe Snowflake schema in data warehouse. State the advantages and disadvantages of star schema? How is it different from star schema?

Q.8: What is fact constellation schema? State any two advantages of fact constellation schema.

*** ***** ***

UNIT 6: DATA WAREHOUSE ARCHITECTURE

UNIT STRUCTURE

6.1 LEARNING OBJECTIVES

After going through this unit you will be able to:

- describe the architecture of data warehouse specially three-tier architecture
- explain the different warehouse models
- describe how to design a warehouse
- describe OLAM and the conversion from OLAP to OLAM.

6.2 INTRODUCTION

In this previous unit we have learned about multidimensional view of data, data cubes and different data warehouse schema such as *star schema*, *snowflake schema* and *fact constellation*. In this unit we will learn about data warehouse architecture, data warehouse design. We will also learn about OLAP three-tier architecture in detail along with indexing &

querying in OLAP, OLAM etc. In the next unit we will explore the concept of data, knowledge and different data visualization techniques.

To design a data warehouse, first we collect the operational data from different source database. Process it for consistency and load in data warehouse. Data warehouse gives the advantages like track the trends and patterns over a long period, can gather information quickly and efficiently, helps us to manage customer relationship. We can design three types of data warehouse server namely data mart, virtual data warehouse and enterprise warehouse. And get another data warehouse as metadata. Data warehouse design is the process of building a solution to integrate multiple sources data that support analytical reporting and data analysis. It consists of requirements gathering, physical environment setup, data modeling, ETL, OLAP cube design, front end development, report development. Indexing the data warehouse can reduce the amount of time it takes to see query results. Indexing can be implemented on dimensions as well as fact table. Integration of OLAP and mining is called as OLAP mining (OLAM) and can easily transform from data warehouse to OLAM.

6.3 DATA WAREHOUSE ARCHITECTURE

Conceptual architectures have been proposed for a data warehouse. Operational data is the data collected and available in the transaction processing system. It resides in the different source databases. Before loading in the warehouse, first process the data for consistency from different sources. The detailed transaction level data that have been cleaned and confirmed for consistency is called reconciled data. It is used as base data for all warehouses.

From the data warehouses, the business analyst takes information to measure the performance and make critical adjustments in order to win over other business holders in the market. A data warehouse offers the following advantages–

- it can enhance business productivity because a data warehouse can gather information quickly and efficiently,
- A data warehouse provides a consistent view of customers and items, since it helps us manage customer relationship.

- A data warehouse also helps in bringing down the costs by tracking trends, patterns over a long period.

We need to understand and analyze the business needs to design an effective and efficient data warehouse and construct a **business analysis framework**. Each person has different views about the design of a data warehouse. These views are as follows–

- **The top-down view:** This view allows the selection of relevant information required for a data warehouse.
- **The data source view:** This view presents the operational system information being captured, stored, and managed.
- **The data warehouse view:** It represents the information stored inside the data warehouse with the help of fact tables and dimension tables.
- **The business query view:** It is the view of the data from the end-user viewpoint.

Data Warehouse architecture is presented in figure 6.1.

Single-Tier Architecture: The objective of a single layer is to minimize the amount of stored data. Remove data redundancy is the main goal of single tier architecture. This architecture is not frequently used in practice.

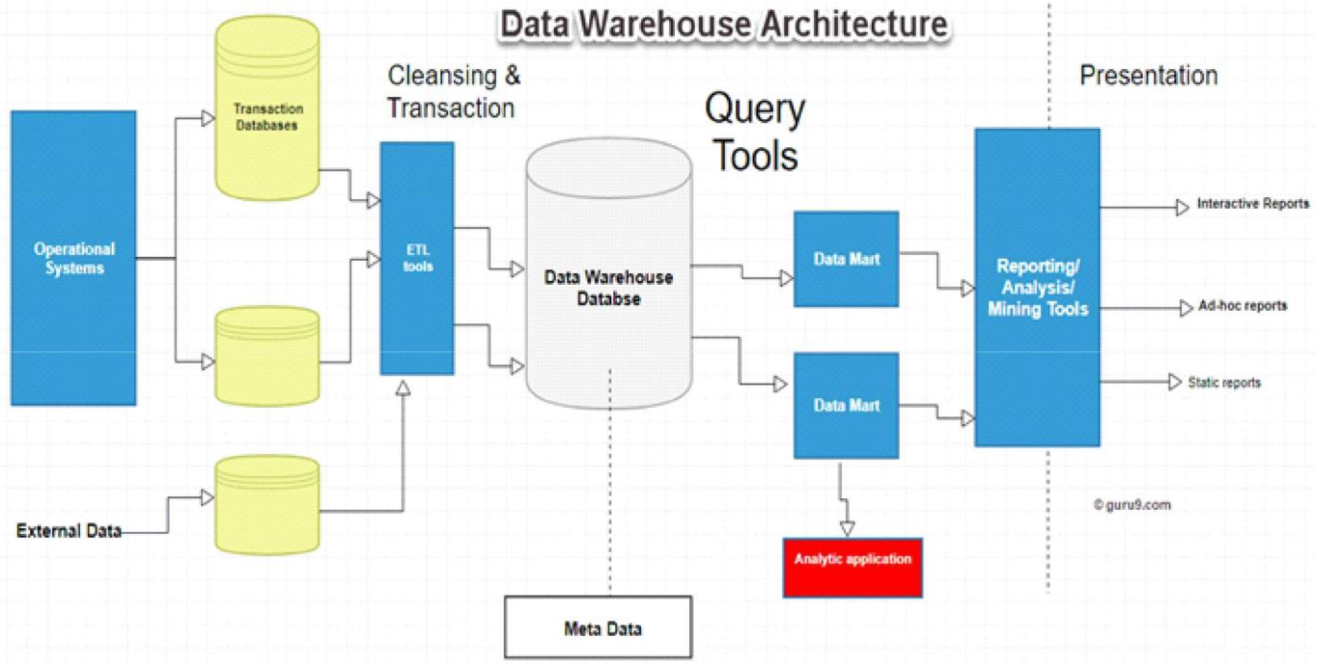


Figure 6.1: Data Warehouse Architecture

Two-Tier Architecture: Two-layer architecture separates physical sources and data warehouse. It is not expandable and also end-user does not support this architecture. Due to network limitations, it has connectivity problems.

Three-Tier Architecture: Generally a data warehouses adopts a three-tier architecture. It has three different tiers namely bottom tier as database server, middle tier as OLAP server and top tier as client or front end.

6.3.1 Data Warehouse Models

From the perspective of data warehouse architecture, we have the following data warehouse models–

- Virtual Warehouse
- Data mart
- Enterprise Warehouse

- **Virtual Warehouse:** The operational data warehouse view is known as a virtual warehouse. It is easy to build a virtual warehouse. Building a virtual warehouse needs excess capacity on operational database servers.
- **Data Mart:** Data mart contains a subset of organizational data. This subset of data is important to specific groups of an organization. In other words, we can claim that group specific data contain in a data mart. For example, data related to items, customers, and sales can be contained in the marketing data mart. Data marts are confined to subjects.

  Points to remember about data marts–

  - Window-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.
  - The implementation data mart cycles is measured in short periods of time, i.e., in weeks rather than months or years.
  - The life cycle of a data mart may be complex in long run, if its planning and design are not organization-wide.
  - Data marts are small in size.

- 0 Data marts are customized by department.
- 0 The source of a data mart is departmentally structured data warehouse.
- 0 Data marts are flexible.

► **Enterprise Warehouse:**

- O All the information and the subjects spanning an entire organization by an enterprise warehouse.
- 0 It provides us integration of enterprise data.
- 0 The data is integrated from operational systems and external information providers.
- 0 This information can be small or large. It can be vary from a few gigabytes to hundreds of gigabytes, terabytes or beyond.

6.3.2 Metadata

Metadata is simply defined as data about data. For example, the index of a book serves as a metadata for the contents in the book. In other words, we can say that metadata is the summarized data that leads us to detailed data. In terms of data warehouse, we can define metadata as follows.

► Metadata is the road-map to a data warehouse.

► Metadata in a data warehouse defines the warehouse objects.

► Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Note: In a data warehouse, we create metadata for the data names and definitions of a given data warehouse.

Categories of Metadata: Metadata can be broadly categorized into three categories–

► **Business Metadata:** It refers to the data ownership information, business definition, and changing policies.

► **Technical Metadata:** It consists of database system names, table and column names and sizes, data types and allowed values. Technical metadata also contains structural information such as primary and foreign key attributes and indices.

- **Operational Metadata:** It contains currency of data and data lineage. Currency of data means whether the data is active, archived, or purged.

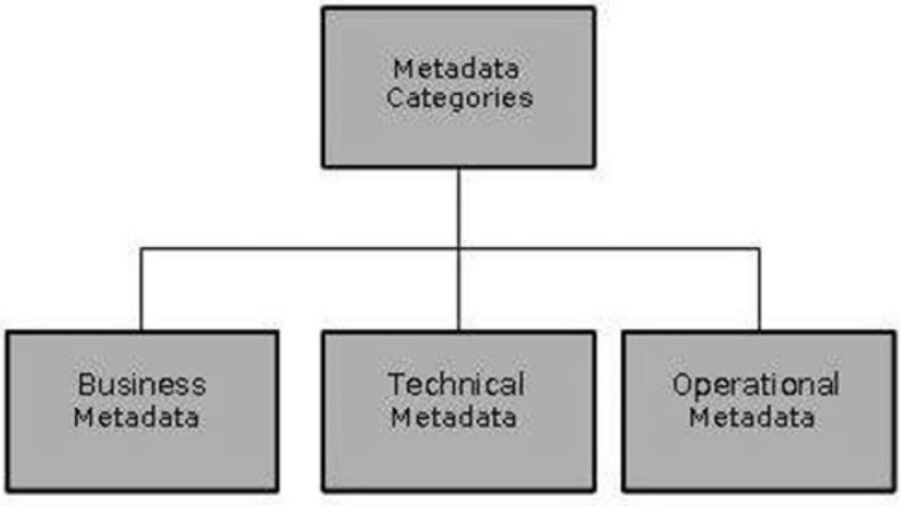


Figure 6.2: Categories of Metadata

Role of Metadata: The role of metadata in a warehouse is different from the warehouse data, and it plays an important role. The various roles of metadata are explained below.

- Metadata acts as a directory.
- This directory helps the decision support system to locate the contents of the data warehouse.
- Metadata helps in decision support system for mapping of data when data is transformed from operational environment to data warehouse environment.
- Metadata helps in summarization between current detailed data and highly summarized data.
- Metadata also helps in summarization between lightly detailed data and highly summarized data.
- Metadata is used for query tools.
- Metadata is used in extraction and cleansing tools.
- Metadata is used in reporting tools.
- Metadata is used in transformation tools.
- Metadata plays an important role in loading functions.

The following diagram shows the roles of metadata.

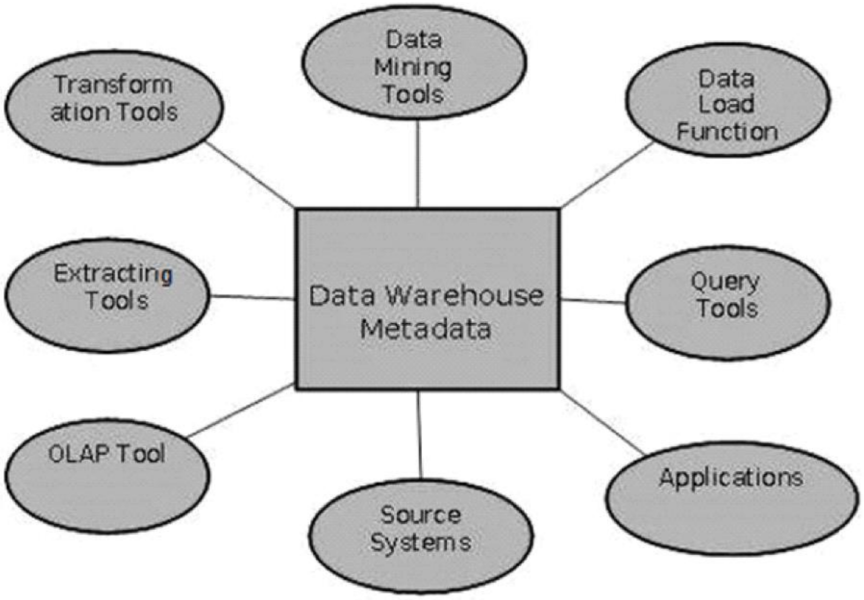


Figure 6.3: Data Warehouse Metadata

Metadata Repository: Metadata repository is an integral part of a data warehouse system. It has the following metadata–

- **Definition of data warehouse:** It includes the structure of data warehouse. Schema, view, hierarchies, derived data definitions, and data mart locations and contents are defined in the structure.
- **Business metadata:** It includes the data ownership information, business definition, and changing policies.
- **Operational metadata:** It contains currency of data and data lineage.
- **Data for mapping from operational environment to data warehouse:** It contains the source databases and their contents, data extraction, data partition cleaning, transformation rules, data refresh and purging rules.
- **Algorithms for summarization:** It consists of dimension algorithms, data on granularity, aggregation, summarizing, etc.

6.4 DATA WAREHOUSE DESIGN

Good Business Intelligence (BI), allows the organization to query data obtained from trusted sources and use the answers to gain a competitive edge in the industry. The first step to effective BI is a well-designed warehouse. Data warehouse design is the process of building a solution to integrate multiple sources data that support analytical reporting and data analysis. A poorly designed data warehouse can result in acquiring and using inaccurate source data that negatively affect the productivity and growth of the organization.

Requirements Gathering: The first step of the data warehouse design process is gathering requirements. The goal of this phase is to determine the criteria for a successful implementation of the data warehouse. An organization’s long-term business strategy is important as the current business and technical requirements. User analysis and reporting requirements is identified as well as hardware, development, testing, implementation, and user training. Once the business and technical strategy has been decided the next step is to address how the organization will back up the data warehouse and how it will recover if the system fails.

Physical Environment Setup: Once the business requirements are set, next step is to determine the physical environment for the data warehouse. There should be separate physical application and database server separate ETL/ELT, OLAP, cube, and reporting processes set up for development, testing, and production. The IT staff can investigate the issue without negatively impacting the production environment if integrity is suspected.

Data Modeling: Once requirements gathering and physical environments have been defined, the next step is to define how data structures will be accessed, connected, processed, and stored in the data warehouse through the process of data modeling. In this phase of data warehouse design, data sources are identified. Once the data sources have been identified, the data warehouse team can begin building the logical and physical structures to fulfill the established requirements.

ETL: The ETL process takes the most time to develop major implementation. Identifying data sources during the data modeling phase is reduce ETL development time. The goal of ETL is to provide optimized load speeds.

OLAP Cube Design: On-Line Analytical Processing (OLAP) provides the infrastructure for ad-hoc user query and multi-dimensional analysis.

Against the query of data OLAP design specification should come. OLAP cube dimensions are specified in the documentation and measures should be obtained during the beginning of data warehouse design process. The three critical elements of OLAP design include:

- Grouping measures– numerical value that want to analyze such as revenue, number of customers, how many products customers purchase, or average purchase amount.
- Dimension– where measures are stored for analysis such as geographic region, month, or quarter.
- Granularity– the lowest level of detail that you want to include in the OLAP dataset.

The OLAP cube process is optimized during development.

Front End Development: This step is executed to work on how users will access the data warehouse. Front end development is defined how users will access the data for analysis and run reports.

Report Development: For most end users, the only contact they have with the data warehouse is through the generated reports. An essential feature for data warehouse report generation is users' ability to select their report criteria quickly and efficiently is. Delivery options are other criteria. A well-designed data warehouse able to handle the new reporting requests with little to no data warehouse system modification.

CHECK YOUR PROGRESS

Q.1: The..................... exposes the information being captured, stored, and managed by operational systems.

a) top-down view b) data warehouse view
c) data source view d) business query view

Q.2: .... describes the data contained in the data warehouse.

a) Relational data b) Operational data
c) Metadata d) Informational data.

Q.3: ....databases are owned by particular departments or business groups.

a) Informational b) Operational

c) Both informational and operational d) Flat

Q.4: What is the full form of ETL?

....

Q.5: Write name of the process in data warehouse design.

....

....

6.5 OLAP THREE TIER ARCHITECTURE

Generally a data warehouses adopts three-tier architecture. Following are the three different tiers of the data warehouse architecture.

- **Bottom Tier:** The data warehouse database server is the bottom tier of the architecture. It is the relational database system which is uses the back end tools and utilities to feed data into the bottom tier. These back end tools and utilities perform the Extract, Clean, Load, and refresh functions.
- **Middle Tier:** In the middle tier, we have the OLAP Server that can be implemented in either of the following ways.
  - By Relational OLAP (ROLAP), which is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.
  - By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations.
- **Top-Tier:** This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.

The following figure 6.4 depicts the three-tier architecture of data warehouse.

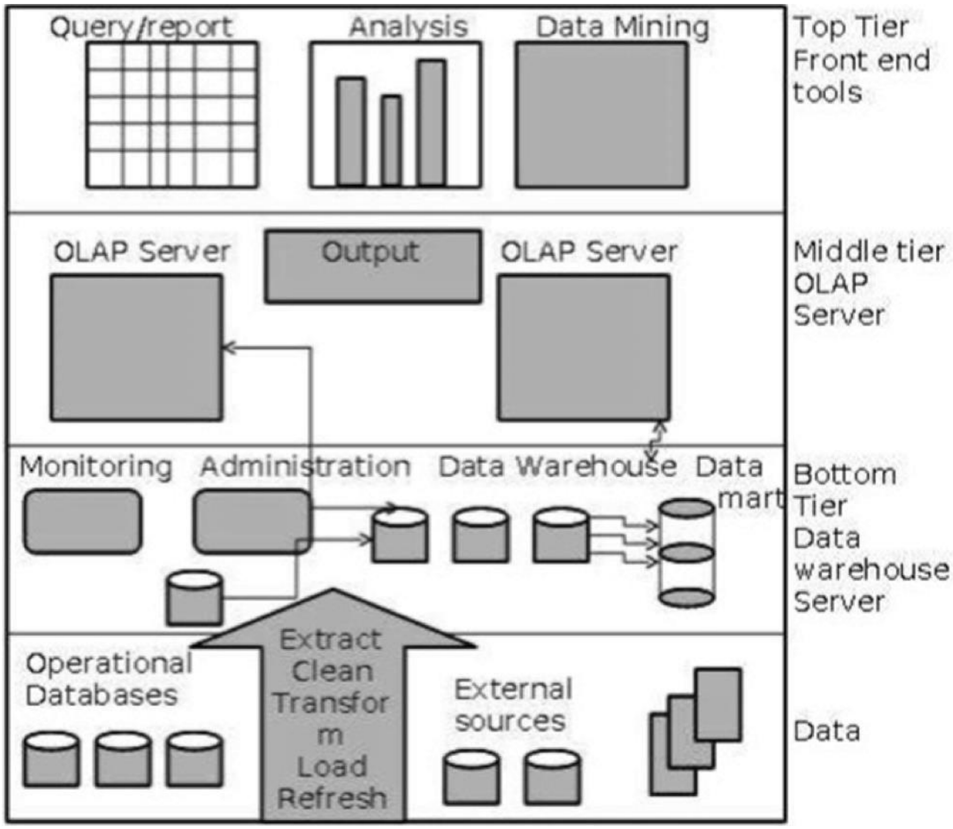


Figure 6.4: Three-tier Architecture of Data Warehouse

6.6 INDEXING AND QUERYING IN OLAP

Indexing the data warehouse can reduce the amount of time to see query results. We can apply indexing on dimensions and on fact table. If too few indexes are applied, the data loads quickly but the query response is slow. If applied indexes are too many, the data loads slowly and your storage requirements go through the roof but the query response is good. Indexing in any database, transactional or warehouse, most often reduces the length of time to see query results. This is especially true with large tables and complex queries that involve in joins operation.

Some of the variables that you'll want to take into account when indexing the data warehouse are the type of data warehouse you have, how large the dimensions and fact tables are, who will be accessing the data and how they'll do so, and whether access will be ad hoc or via structured application interfaces. These variables will determine how indexing scheme should be structured.

Indexing Dimensions: If you want to index the dimension key (primary key), which is not a "natural" or transactional key such as customer name or customer ID where we can not apply clustering.

Indexing the Fact Table: Indexing the fact table is similar to indexing a dimension, although you must account for partitioning.

Modifying Your Indexing Scheme: Over time, you'll have to modify your indexing scheme to show the changes to accommodate what's happening in your organization. And most data warehouse/BI systems will access directly relational tables, so you can use tried-and-true transactional methods for tuning indexes, such as evaluating the query and data mix and adjusting it accordingly.

Querying in OLAP: Online Analytical Processing (OLAP) databases facilitate business-intelligence queries. OLAP is a database technology that has been optimized for querying and reporting, instead of processing transactions. The source data for OLAP is Online Transactional Processing (OLTP) databases stored in data warehouses. OLAP data is derived from this historical data, and aggregated into structures that permit sophisticated analysis for multidimensional structure. OLAP data is also organized hierarchically and stored in cubes. The organization displays high-level summaries using a PivotTable report or PivotChart report, such as sales totals across an entire country or region, and also display the details for sites where sales are particularly strong or weak.

6.7 OLAM

OLAP Mining (OLAM) is an Integration of Data Mining and Data Warehousing–

- On-line analytical mining of data warehouse data is represent as integration of mining and OLAP technologies.
- Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing etc.
- OLAM is Interactive characterization, comparison, association, classification, clustering, and prediction.

- Integration of data mining functions, e.g., first clustering and then association.

Importance of OLAM: OLAM is important for the following reasons–

- **High quality of data in data warehouses:** The data mining tools are required to work on integrated, consistent, and cleaned data. These steps are very costly in the preprocessing of data. The data warehouses constructed by such preprocessing are valuable sources of high quality data for OLAP and data mining as well.
- **Available information processing infrastructure surrounding data warehouses:** Information processing infrastructure refers to accessing, integration, consolidation, and transformation of multiple heterogeneous databases, web-accessing and service facilities, reporting and OLAP analysis tools.
- **OLAP–based exploratory data analysis:** Exploratory data analysis is required for effective data mining. OLAM provides facility for data mining on various subset of data and at different levels of abstraction.
- **Online selection of data mining functions:** Integrating OLAP with multiple data mining functions and online analytical mining provide users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

6.8 IMPLEMENTATION FROM DATA WAREHOUSING (OLAP) TO DATA MINING (OLAM)

Online Analytical Mining integrates with Online Analytical Processing with data mining and mining knowledge in multidimensional databases. The figure 6.5 shows the integration of both OLAP and OLAM–

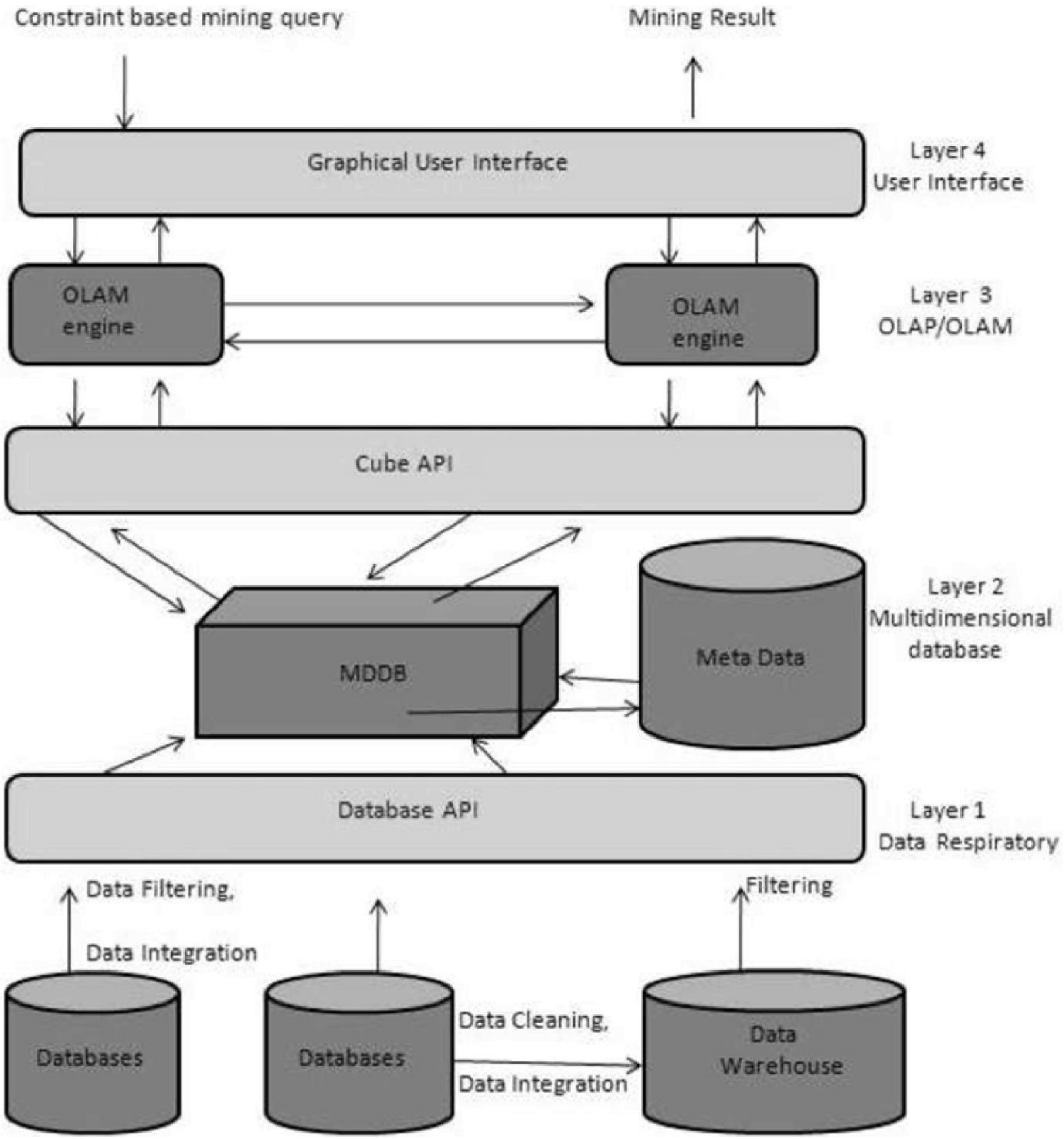


Figure 6.5: Integration of OLAP and OLAM

CHECK YOUR PROGRESS

Q.6: The load and index is–

a) A process to reject data from the data warehouse and to create the necessary indexes.

b) A process to load the data in the data warehouse and to create the necessary indexes.

c) A process to upgrade the quality of data after it is moved into a data warehouse.

d) A process to upgrade the quality of data before it is moved into a data warehouse.

Q.7: The active data warehouse architecture includes–

a) at least one data mart

b) data that can extracted from numerous internal and external sources

c) near real-time updates

d) all of the above.

Q.8: Reconciled data is–

a) data stored in the various operational systems throughout the organization.

b) current data intended to be the single source for all decision support systems.

c) data stored in one operational system in the organization.

d) data that has been selected and formatted for end-user support applications.

Q.9: What are advantages of OLAM?

....

....

....

....

6.9 LET US SUM UP

- Metadata is simply defined as data about data.
- The operational data warehouse view is known as a virtual warehouse.
- Data mart contains a subset of organizational data.
- Data warehouse design is the process of building a solution to integrate multiple sources data that support analytical reporting and data analysis.
- OLAP is a database technology that has been optimized for querying and reporting, instead of processing transactions.
- The source data for OLAP is Online Transactional Processing (OLTP) databases stored in data warehouses.

6.10 FURTHER READING

1) Han, J., Pei, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
2) Ponniah, P. (2011). *Data Warehousing Fundamentals for IT Professionals*. John Wiley & Sons.
3) Pujari, A. K. (2001). Data Mining Techniques. Universities Press.

6.11 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: (c)

Ans. to Q. No. 2: (c)

Ans. to Q. No. 3: (b)

Ans. to Q. No. 4: Extract, Transform and Load

Ans. to Q. No. 5: Requirement gathering, Physical environment setup, Data modeling, ETL, OLAP cube design, Front end development, Report development.

Ans. to Q. No. 6: (b)

Ans. to Q. No. 7: (d)

Ans. to Q. No. 8: (b)

Ans. to Q. No. 9: High quality of data in data warehouses, Available information processing infrastructure surrounding data warehouses, OLAP–based exploratory data analysis, Online selection of data mining functions.

6.12 MODEL QUESTIONS

Q.1: Define Indexing.

Q.2: Write about the different models of data warehouse.

Q.3: Define data warehouse.

Q.4: Write about the component of data warehouse.

Q.5: Define OLAM.

Q.6: Discuss the importance of OLAM.

Q.7: List out the steps of data warehouse design.

Q.8: Describe the three-tier architecture of data warehouse.

Q.9: Define Metadata.

*** ***** ***

UNIT 7: DATA MINING KNOWLEDGE REPRESENTATION

UNIT STRUCTURE

7.1 LEARNING OBJECTS

After going through this unit, you will be able to:

- describe different primitives of data mining task
- represent data and knowledge
- describe basic interestingness measures
- describe different visualization techniques.

7.2 INTRODUCTION

In this previous unit we have learned data warehouse architecture and data warehouse design. We have also learned about OLAP three -tier architecture in detail along with indexing & querying in OLAP, OLAM etc. In this unit, we will learn about data mining tasks and different visualization techniques.

Generally, we use data mining for a long process of research and product development. Also, we can say this evolution was started when business data was first stored on computers. We can also navigate through their data in real time. Data Mining is also popular in the business community. As this is supported by three technologies that are now mature: Massive data collection, Powerful multiprocessor computers, and Data mining algorithms.

A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during discovery in order to direct the mining process. Data mining is based on different task relative primitive like which portion of the transactional database to be mined, what kind of data to be mined, what background knowledge are important to represent the output knowledge, which would be visualize with different techniques. In the next unit, we will explore the concept of attribute generalization, attribute relevance and discuss many statistical measures.

7.3 TASK RELEVANT DATA

Each user will have a data mining task, that is, some form of data analysis that he or she would like to have performed. Data mining query decides the data mining task, which is input to the data mining system. A data mining query is basically defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths.

Here is the list of data mining task primitives–

- **Set of task relevant data to be mined:** This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest (referred to as the relevant attributes or dimensions).
- **Kind of knowledge to be mined:** This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.
- **Background knowledge to be used in discovery process:** This knowledge about the domain to be mined is useful for guiding the

knowledge discovery process and for evaluating the patterns found. Concept hierarchies are a popular form of background knowledge, which allow data to be mined at multiple levels of abstraction.

- **Interestingness measures and thresholds for pattern evaluation:** They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures. For example, interestingness measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.
- **Representation for visualizing the discovered patterns:** This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes.

7.4 BACKGROUND KNOWLEDGE

Background knowledge consists both of domain-specific knowledge as well as general knowledge about the behavior of the world. Use of background knowledge in the process of identifying general patterns within a database leads to patterns that are more useful and significant. Many data mining problems can be solved better if more background knowledge is added: predictive models can become more accurate, and descriptive models can reveal more interesting endings. Collecting and integrating background knowledge is a manual work.

Data mining deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of functions involved in data mining–

1) Descriptive
2) Classification and Prediction

1) Descriptive Function: The descriptive function deals with the general properties of data in the database. Here is the list of descriptive functions–

- Class/Concept Description
- Mining of Frequent Patterns

- Mining of Associations
- Mining of Correlations
- Mining of Clusters

a) **Class/Concept Description:** Class/Concept refers to the data to be associated with the classes or concepts. For example, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by the following two ways–

- ► **Data Characterization:** This refers to summarizing data of class under study. This class under study is called as Target Class.
- ► **Data Discrimination:** It refers to the mapping or classification of a class with some predefined group or class.

b) **Mining of Frequent Patterns:** Frequent patterns are occur frequently in transactional data. Here is the list of kind of frequent patterns–

- ► **Frequent Item Set:** It refers to a set of items that frequently appear together, for example, milk and bread.
- ► **Frequent Subsequence:** A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.
- ► **Frequent Sub Structure:** Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with item-sets or subsequences.

c) **Mining of Association:** Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to the process of uncovering the relationship among data and determining association rules. For example, a retailer generates an association rule that shows that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

d) **Mining of Correlations:** It is a kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs or between two item sets to analyze that if they have positive, negative or no effect on each other.

e) **Mining of Clusters:** Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.

2) **Classification and Prediction:** Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of sets of training data. The derived model can be presented in the following forms–

- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

The list of functions involved in these processes are as follows–

a) **Classification:** It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The derived model is based on the analysis set of training data i.e. the data object whose class label is well known.

b) **Prediction:** It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.

c) **Outlier Analysis:** Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.

d) **Evolution Analysis:** Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.

CHECK YOUR PROGRESS

Q.1: Background knowledge referred to–

a) Additional acquaintance used by a learning algorithm to facilitate the learning process.

b) A neural network that makes of a hidden layer.

c) It is form of automatic learning

d) None of these

Q.2:..........................is not a data mining functionality.

a) Clustering and Analysis

b) Selection and Interpretation

c) Classification and Regression

d) Characterization and Discrimination

Q.3: Classification is–

a) A subdivision of a set of examples into a number of classes.

b) A measure of accuracy, of the classification of a concept that is given by certain theory

c) The task of assigning a classification to a set of examples

d) None of these

Q.4: Prediction is–

a) The result of the application of a theory or a rule in a specific case

b) One of several possible enters within a database table that is chosen by the designer as the primary means of accessing the data in the table.

c) Discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces.

d) None of these

7.5 INTERESTINGNESS MEASURE

Interestingness measures have an important role in data mining, regardless of the kind of patterns being mined. These measures are using for selecting and ranking patterns according to their potential interest to the user. Good measures also allow the time and space costs of the mining process to be reduced.

Measuring the interestingness of discovered patterns is an active and important area of data mining. Based on the diversity of definitions presented to-date, interestingness is perhaps best treated as a broad concept that emphasizes conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility, and action ability. These nine specific criteria are used to determine whether or not a pattern is interesting. They are described as follows:

- **Conciseness:** If pattern contains relatively few attribute-value pairs then it is concise, while a set of patterns is concise if it contains relatively few patterns. A concise pattern or set of patterns is relatively easy to understand and remember and thus is added more easily to the user's knowledge
- **Generality/Coverage:** If a pattern covers a relatively large subset of a dataset then it is general. Generality (or coverage) measures the comprehensiveness of a pattern, that is, the fraction of all records in the dataset that matches the pattern. If a pattern characterizes more information, it tends to be more interesting. Frequent item sets are the most studied general patterns in the data mining literature. Generality frequently coincides with conciseness because concise patterns tend to have greater coverage.
- **Reliability:** A pattern is reliable if the relationship described by the pattern occurs in a high percentage of applicable cases. For example, a classification rule is reliable if its predictions are highly accurate, and an association rule is reliable if it has high confidence.
- **Peculiarity:** A pattern is peculiar if it is far away from other discovered patterns according to some distance measure. Peculiar patterns are generated from peculiar data (or outliers), which are relatively few in number and significantly different from the rest of the data
- **Diversity:** A pattern is diverse if its elements differ significantly from each other, while a set of patterns is diverse if the patterns in the set

differ significantly from each other. Diversity is a common factor for measuring the interestingness of summaries.

- **Novelty:** A pattern is novel to a person if he or she did not know it before and is not able to infer it from other known patterns. No known data mining system represents everything that a user knows, and thus, novelty cannot be measured explicitly with reference to the user's knowledge.
- **Surprisingness:** A pattern contradicts a person's existing knowledge or expectation is termed as surprising (or unexpected). A pattern that is an exception to a more general pattern which has already been discovered can also be considered surprising. Surprising patterns are interesting because they identify failings in previous knowledge and may suggest an aspect of the data that needs further study.
- **Utility:** A pattern uses by a person contributes to reaching a goal is called utility. Different people have different goals concerning the knowledge that can be extracted from a dataset. This kind of interestingness is based on user-defined utility functions in addition to the raw data.
- **Actionability/Applicability:** A pattern is actionable in some domain if it enables decision making about future actions in this domain. Action ability is sometimes associated with a pattern selection strategy.

7.6 REPRESENTING INPUT DATA AND OUTPUT KNOWLEDGE

i) **Concept:** This concept is introduced as what things are to be mined using following categories of mining:

   - **Classification mining/learning:** predicting a discrete class, a kind of supervised learning, success is measured on new data for which class labels are known (test data).
   - **Association mining/learning:** detecting associations between attributes, can be used to predict any attribute value and more than one attribute values, hence more rules can be generated, therefore we need constraints.

- **Clustering:** grouping similar instances into clusters, a kind of unsupervised learning, success is measured subjectively or by objective functions.
- **Numeric prediction:** predicting a numeric quantity, a kind of supervised learning, success is measured on test data.
- **Concept description:** output of the learning scheme.

ii) Instance: Instances are defined as what things to be classified, associated, or clustered. Individual and independent examples of the concept to be learned (target concept). Instance is described by predetermined set of attributes. Input to the learning scheme is defined as set of instances (dataset), represented as a single relation (table), independence assumption, positive and negative examples are taking for a concept.

iii) Attributes: Attributes (features) of input data are predefined set of features to describe an instance. They are nominal (distinct and no relation between them), structured and numeric.

iv) Output knowledge: Output knowledge is represented as the output of Association rules, Classification rules, Rules with relation, used the different prediction schemes like nearest neighbor, Bayesian classification, Neural networks, Regression. And output are represented as decision trees where knowledge is portioned as cluster on the basis of structure, concept or statistics.

7.7 VISUALIZATION TECHNIQUE

Visual Data Mining is the process of discovering implicit but useful knowledge from large data sets using visualization techniques.

Data visualization aims to communicate data clearly and effectively through graphical representation. Data visualization has been used extensively in many applications. For e.g., at work for reporting managing business operations and tracking progress of tasks. More popularly, we can take advantage of visualization techniques to discover data relationships that are otherwise not easily observable by looking at the raw data.

Different Data Visualization techniques are as follows:

1) **Pixel oriented visualization techniques:**

- A simple way to visualize the value of a dimension is to use a pixel where the color of the pixel reflects the dimension's value.
- For a data set of m dimensions pixel oriented techniques create m windows on the screen, one for each dimension.
- The m dimension values of a record are mapped to m pixels at the corresponding position in the windows.
- The color of the pixel reflects other corresponding values.
- Inside a window, the data values are arranged in some global order shared by all windows
- Example: All Electronics maintains a customer information table, which consists of 4 dimensions: income, credit_limit, transaction_volume and age. We analyze the correlation between income and other attributes by visualization.
- We sort all customers in income in ascending order and use this order to layout the customer data in the 4 visualization windows as shown in figure 7.1.
- The pixel colors are chosen so that the smaller the value, the lighter the shading.

Using pixel based visualization we can easily observe that credit_limit increases as income increases customer whose income is in the middle range are more likely to purchase more from All Electronics, there is no clear correlation between income and age.

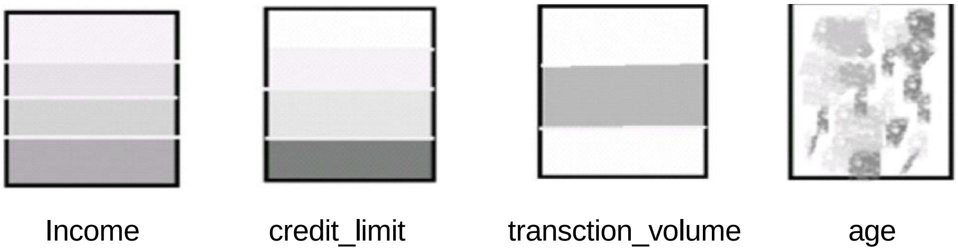


Figure 7.1: Pixel oriented visualization of 4 attributes by sorting all customers in income Ascending order

2) **Geometric Projection visualization techniques:**

- A drawback of pixel-oriented visualization techniques is that they cannot help us much in understanding the distribution of data in a multidimensional space.
- Geometric projection techniques help users find interesting projections of multidimensional data sets.
- A scatter plot displays 2-D data point using Cartesian co-ordinates. A third dimension can be added using different colors of shapes to represent different data points.
- Eg: Where x and y are two spatial attributes and the third dimension is represented by different shapes.

  Through this visualization, we can see that points of types "+" & ''X" tend to be collocated.

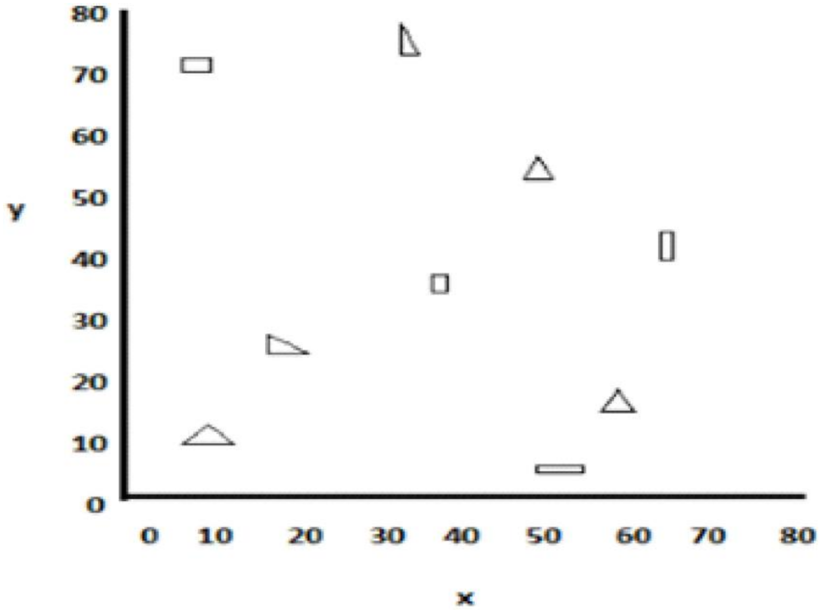


Figure 7.2: Visualization of 2D data set using scatter plot

3) **Icon based visualization techniques:**

- It uses small icons to represent multidimensional data values.
- Two popular icon based techniques are listed below:
  - **Chernoff faces:** it was introduced in 1973 by Herman Chernoff. They display multidimensional data of up to 18 variables as a cartoon human face. Chernoff faces helps to reveal trends in data. Component of face.

- **Stick figures:** It maps multidimensional data to five-piece stick figure, where each figure has 4 limbs and a body. Two dimensionsare mapped to the display axes and the remaining dimensions are mapped to the angle and/or length of the limbs.

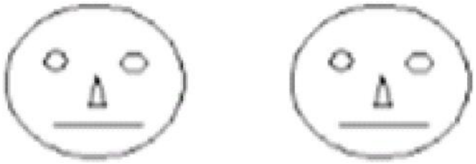
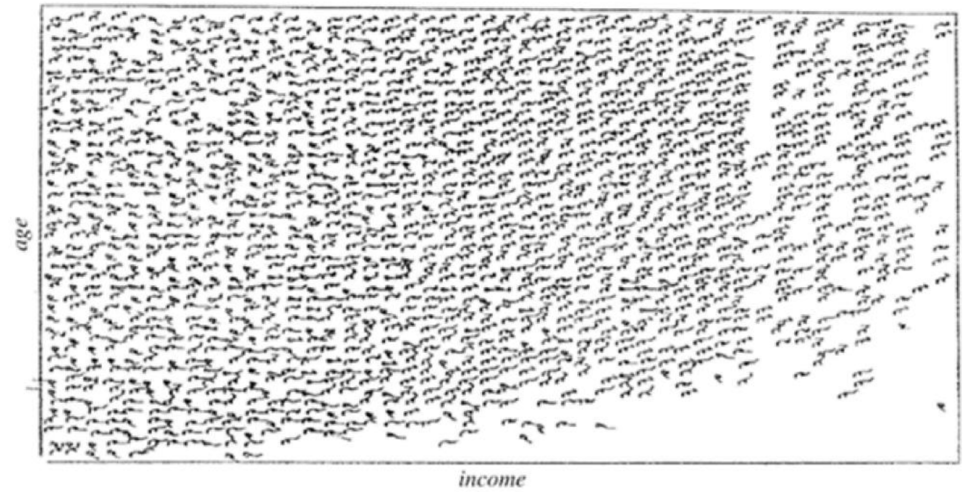


Figure 7.3: Chernoff faces each face represents an 'n' dimensional data points (n<18)



4) **Hierarchical visualization techniques: (i.e.,subspaces):** These techniques focus on visualizing multiple dimensions simultaneously. A large data set of high dimensionality, it would be difficult to visualize all dimensions at the same time. Hierarchical visualization technique makes subset of the dimensions. "Worlds-within-Worlds" also known as n-vision is representing by Hierarchical visualization technique.

5) **Visualizing Complex data and relation:** There are many new techniques dedicated to non-numerical data. For example, many people on the web tag, blog entries and product reviews. A **tag cloud** is a visualization of statistics of user generated tag where tags are listed alphabetically or user preferred order. Important tags are listed with color. In addition, relation among complex data entries also raises challenges for visualization.

CHECK YOUR PROGRESS

Q.6: An objective measure of pattern interestingness in data mining is/are:

a) Support rule b) Confidence rule

c) Both (a) & (b) d) Neither (a) nor (b)

Q.7: Explain Syntax for Interestingness Measures Specification.

....

....

....

Q.8: Explain Syntax for Pattern Presentation and Visualization Specification.

....

....

....

Q.9: What is tag cloud?

....

....

7.8 LET US SUM UP

- Different data mining tasks are the core of data mining process. Different prediction and classification data mining tasks actually extract the required information from the available data sets.
- Background knowledge are either domain specific or based on general knowledge.
- Data mining functions are classified into two main categories: Descriptive, Classification and prediction.
- Interestingness measure is not depends on kind of pattern being mind. Good measure reduces the time and space cost.

- Input data are introduced on the basis of classification, association, clustering and numerical prediction.
- Large data can be easily represent with different visualization techniques.

7.9 FURTHER READING

1) Han, J., Pei, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

2) Pujari, A. K. (2001). Data Mining Techniques. Universities Press.

7.10 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: (a)

Ans. to Q. No. 2: (b)

Ans. to Q. No. 3: (a)

Ans. to Q. No. 4: (a)

Ans. to Q. No. 6: (c)

Ans. to Q. No. 7: Interestingness measures and thresholds can be specified by the user with the statement–

with <interest_measure_name> threshold = threshold_value.

Ans. to Q. No. 8: Generally, we have a syntax, which allows users to specify the display of discovered patterns in one or more forms. display as <result_form>

Ans. to Q. No. 9: A tag cloud is a visualization of statistics of user generated tag where tags are listed alphabetically or user preferred order.

7.11 MODEL QUESTIONS

Q.1: Name some data mining techniques?

Q.2: What is the foundation of data mining?

Q.3: Why is background knowledge required for data mining?

Q.4: What are the task related primitives used in data mining?

Q.5: What is input data for data mining?

Q.6: What are attributes used for representing the input data in data mining?

Q.7: Describe the different aspects of the interestingness measures.

Q.8: Write about the different techniques to visualize large data.

*** ***** ***

UNIT 8: ATTRIBUTE-ORIENTED ANALYSIS

UNIT STRUCTURE

8.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define attribute generalization and attribute relevance
- describe data cube approach
- describe class comparison
- describe the different statistical measures.

8.2 INTRODUCTION

In the previous unit we have discussed topics like data, knowledge and different data visualization techniques. In this unit, we will learn about attribute generalization and attribute relevance along with class comparison in detail. We will also explore the different statistical measures like mean, median, mode etc in detail in this unit.

Data mining usually says about knowledge discovery from data. To know about the data it is necessary to go through the data objects, data attributes and types of data attributes. Mining data also includes relation between data. Data objects are the essential part of a database. A data object represents the entity. Data Objects are like group of attributes of a entity. For example, a sales data object may represent customer, sales or purchases. When a data object is listed in a database they are called data tuples. A set of attributes used to describe a given object are known as attribute vector. Attributes are mainly categorized as qualitative and quantitative.

In general, *data generalization* summarizes data by replacing relatively low-level values (e.g., numeric values for an attribute *age*) with higher-level concepts (e.g., *young*, *middle-aged*, and *senior*), or by reducing the number of dimensions to summarize data in concept space involving fewer dimensions.

In many applications, users may not be interested in having a single class (or concept) described or characterized, but prefer to mine a description that compares or distinguishes one class (or concept) from other comparable classes (or concepts). Class discrimination or comparison (hereafter referred to as **class comparison**) mines descriptions that distinguish a target class from its contrasting classes. We start with *measures of* ***central tendency***, which measure the location of the middle or center of a data distribution. The most common data dispersion measures are the *range*, *quartiles*, and *inter-quartile range*; the *five-number summary* and *boxplots* and the *variance* and *standard deviation* of the data. Most statistical or graphical data presentation software packages include bar charts, pie charts, and line graphs. Other popular displays of data summaries and distributions include *quantile plots*, *quantile-quantile plots*, *histograms*, and *scatter plots*. In the next unit, in the next block we will explore the concept of association rule mining in detail.

8.3 ATTRIBUTE GENERALIZATION

8.3.1 Attribute

It can be seen as a data field that represents characteristics or features of a data object. For a customer object, attributes can be customer Id, address etc. We can say that a set of attributes used to describe a given object are known as attribute vector or feature vector.

Type of attributes: There are two different types of attributes. The attribute types are:

1) Qualitative [Nominal (N), Ordinal (O), Binary (B)].
2) Quantitative (Discrete, Continuous)

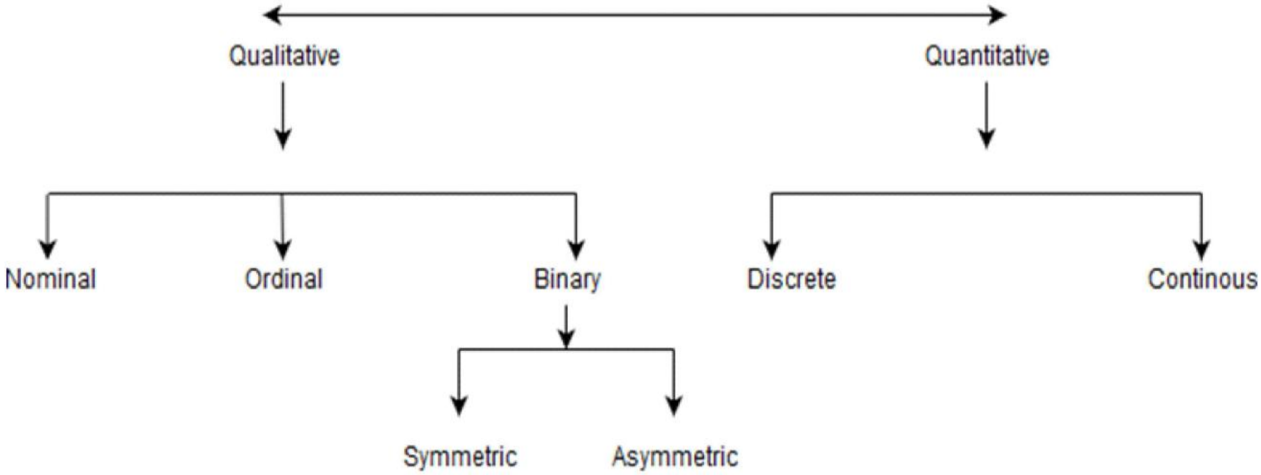


Figure 8.1: Different Types of Attributes

Nominal Attributes-related to names: The values of a Nominal attribute are name of things, some kind of symbols. Values of Nominal attributes represents some category or state. So these attribute are referred as **categorical attributes** and there is no order (rank, position) among values of nominal attribute. E.g., black, blue are the value of color attribute.

Binary Attributes: Binary data has only two values or states. For example, yes or no, affected or unaffected, true or false.

i) **Symmetric:** Both values are equally important (Gender).
ii) **Asymmetric:** Both values are not equally important (Result).

Ordinal Attributes: The Ordinal Attributes contain values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don't indicate how important it is. E.g. grade values are A, B, C, D, E.

Numeric: A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of two types, **interval** and **ratio**.

i) An **interval-scaled** attribute has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point or we can call zero point. Data can be added and subtracted at interval scale but cannot be multiplied or divided. Temperature of two days not comparable.

ii) A **ratio-scaled** attribute is a numeric attribute with a fix zero-point. If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value. Mean, median values are the example of this type.

- **Discrete:** Discrete data have finite values it can be numerical and can also be in categorical form. E.g. teacher, business man, peon are the value of profession attribute.
- **Continuous:** Continuous data have infinite no of states. Continuous data is of float type. There can be many values between 2 and 3. E.g., 5.4, 6.3 are the values of height attribute. For data generalization, there are two approaches namely: data cube(OLAP) approach and attribute oriented induction approach. The general idea behind attribute relevance analysis is to compute some measure which is used to quantify the relevance of an attribute with respect to a given class.

8.3.2 Attribute Generalization

Conceptually, the data cube can be viewed as a kind of multidimensional data generalization. In general, *data generalization* summarizes data by replacing relatively low-level values (e.g., numeric values for an attribute *age*) with higher-level concepts (e.g., *young*, *middle-aged*, and *senior*), or by reducing the number of dimensions to summarize data in concept space involving fewer dimensions (e.g., removing *birth date* and *telephone number* when summarizing the behavior of a group of students). Given the large amount of data stored in databases, it is useful to be able to describe concepts in concise and succinct terms at generalized (rather than low) levels of abstraction. Data generalization is a process that abstracts a large set of task relevant data in a database from relatively conceptual level to high conceptual levels. The generalization of large data sets can be categorized according to two approaches.

1) The data cube(OLAP) approach
2) The attribute oriented induction approach

The Data Cube Approach: For example, *All Electronics* database, sales managers may prefer to view the data generalized to higher levels, such as summarized by customer groups according to geographic regions, frequency of purchases per group, and customer income.

This leads us to the notion of *concept description*, which is a form of data generalization. A concept typically refers to a data collection such as *frequent_buyers, graduate_students*, and so on. **Concept description** generates descriptions for data *characterization* and *comparison*. When concept refers to a class, it is called **class description**. **Characterization** provides a concise and succinct summarization of the given data collection.

We have studied data cube (or OLAP) approaches to concept description using multidimensional, multilevel data generalization in data warehouses. *"Is data cube technology sufficient to accomplish all kinds of concept description tasks for large data sets?"* Consider the following cases.

- **Complex data types and aggregation:** Data warehouses and OLAP tools are based on a multidimensional data model that views data in the form of a data cube, consisting of dimensions (or attributes) and measures (aggregate functions). Furthermore, the aggregation of attributes in a database may include sophisticated data types such as the collection of non-numeric data, the merging of spatial regions, the composition of images, the integration of texts, and the grouping of object pointers. Therefore, OLAP, with its restrictions on the possible dimension and measure types, represents a simplified model for data analysis. Concept description should handle complex data types of the attributes and their aggregations, as necessary.

- **User control versus automation:** Online analytical processing in data warehouses is a user-controlled process. The selection of dimensions and the application of OLAP operations (e.g., drill-down, roll-up, slicing, and dicing) are primarily directed and controlled by users. The control in most OLAP systems is quite user-friendly. Furthermore, in order to find a satisfactory description of the data, users may need to specify a long sequence of OLAP operations.

  This section presents an alternative method for concept description, called *attribute-oriented induction*, which works for complex data types and relies on a data-driven generalization process.

Attribute-Oriented Induction for Data Characterization: The **attribute-oriented induction (AOI)** approach to concept description was first proposed in 1989, a few years before the introduction of the data cube approach. The data cube approach is essentially based on *materialized views* of the data, which typically have been pre-computed in a data warehouse. In general, it performs offline aggregation before an OLAP or data mining query is submitted for processing. On the other hand, the attribute-oriented induction approach is basically a *query-oriented*, generalization-based, online data analysis technique.

The general idea of attribute-oriented induction is to first collect the task-relevant data using a database query and then perform generalization based on the examination of the number of each attribute's distinct values in the relevant data set. This process is called **data focusing.** Then generalization is performed either by *attribute removal* or *attribute generalization*. Aggregation is performed by merging identical generalized tuples and accumulating their respective counts.

Attribute removal is based on the following rule: *If there is a large set of distinct values for an attribute of the initial working relation, but either (case 1) there is no generalization operator on the attribute (e.g., there is no concept hierarchy defined for the attribute), or (case 2) its higher-level concepts are expressed in terms of other attributes, then the attribute should be removed from the working relation*.

Attribute generalization is based on the following rule: *If there is a large set of distinct values for an attribute in the initial working relation, and there exists a set of generalization operators on the attribute, then a generalization operator should be selected and applied to the attribute*.

Both rules– *attribute removal* and *attribute generalization* – claim that if there is a *large* set of distinct values for an attribute, further generalization should be applied. This raises the question: How large is *"a large set of distinct values for an attribute"* considered to be?

Depending on the attributes or application involved, a user may prefer some attributes to remain at a rather low abstraction level while others are generalized to higher levels. The control of how high an attribute should be generalized is typically quite subjective. The control of this process is called **attribute generalization control**. There are many possible ways to control a generalization process. We will describe here two common approaches. The first technique, called **attribute generalization threshold control**, either sets one generalization threshold for all of the attributes, or sets one threshold for each attribute. If the number of distinct values in an attribute is greater than the attribute threshold, further attribute removal or attribute generalization should be performed. Data mining system have a default attribute threshold value ranging from 2 to 8. The second technique, called **generalized relation threshold control**, sets a threshold for the generalized relation. If the number of (distinct) tuples in the generalized relation is greater than the threshold, further generalization should be performed. Otherwise, no further generalization should be performed. For data mining system, this threshold value is ranging from 10 to 30.

8.4 ATTRIBUTE RELEVANCE

The general idea behind attribute relevance analysis is to compute some measure which is used to quantify the relevance of an attribute with respect to a given class. Such measures include the information gain, gini index, uncertainty, and correlation coefficients.

Let S be a set of training object (or tuple) where the class label of each tuple is known. Suppose that there are m classes. Let S contain S_i objects of class C_i, for i = 1, ..., m. An arbitrary object belongs to class C_i with probability S_i/s, where s is the total number of objects in set S. The expected information needed to classify given tuple is:

$$I(s_1, s_2, ..., ..., ..., s_m) = -\sum_{i=1}^{m} s_i \big/ s \log_2 \frac{s_i}{s} \qquad \textbf{(8.1)}$$

If an attribute A with values {a1, a2, ..., av} is used to partition S into the subsets {S1, S2, ..., Sv}, where S_j contains those objects in S that have value a_j of A. Let S_j contain S_{ij} objects of class C_i. The expected information based on this partitioning by A is known as the entropy of A. It is the weighted average:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots\ldots\ldots\ldots + s_{mj}}{s} I(s_{1j} + \ldots\ldots\ldots + s_{mj}) \qquad \textbf{(8.2)}$$

The information gained by branching on A is defined by:

$$Gain(A) = I(s_1, s_2, ..., ..., ..., s_m) - E(A) \qquad \textbf{(8.3)}$$

The attribute which maximizes gain(A) is selected. Attribute relevance analysis for class description is performed as follows.

1) **Data Collection:** Collect data for both the target class and the contrasting class by query processing. Notice that for class comparison, both the target class and the contrasting class are provided by the user in the data mining query. For class characterization, the target class is the class to be characterized, whereas the contrasting class is the set of comparable data which are not in the target class.

2) **Preliminary Relevance analysis using conservative AOI:** Attribute-oriented induction (AOI) can be used to perform some preliminary relevance analysis on the data by removing or generalizing attributes having a large number of distinct values (such as name and phone#). Such attributes are unlikely to be meaningful for concept description. To be conservative, the AOI should employ attribute generalization thresholds that are set reasonably large. (so as to allow more attributes to be considered in further relevance analysis by selected measure performed in step-3). The relation obtained by such an attribute removal and attribute generalization process is called the candidat relation of the mining task.

3) **Remove irrelevant or weakly relevant attributes using the selected measure:** The selected relevance measure is used to evaluate (or rank) each attribute in the candidate relation. For example, the information gain measure described above may be used. The attributes are then sorted (i.e., ranked) according to their computed relevance measure value. Attribute that are not relevant or weakly relevant are then removed based on the set threshold. The resulting relation is called “Initial Target/Contrast class Working Relation”.

CHECK YOUR PROGRESS

Q.1: Identify the example of nominal attribute?

a) Gender b) Temperature

c) Mass d) Salary

Q.2: What is Attribute removal?

....

....

....

....

....

Q.3: What is data focusing?

....

....

....

....

Q.4: Name the process of data relevance analysis.

....

....

....

Q.5: What is nominal attribute?

....

....

....

8.5 CLASS COMPARISON

In many applications, users may not be interested in having a single class (or concept) described or characterized, but prefer to mine a description that compares or distinguishes one class (or concept) from other comparable classes (or concepts). Class discrimination or comparison (hereafter referred to as **class comparison**) mines descriptions that distinguish a target class from its contrasting classes. Notice that the target and contrasting classes must be *comparable* in the sense that they share similar dimensions and attributes. For example, the three classes *person, address*, and *item* are not comparable. However, sales in the last three years are comparable classes, and so are, for example, computer science students versus physics students.

Suppose, for instance, that we are given the *All Electronics* data for sales in 2009 and in 2010 and want to compare these two classes. Consider the dimension *location* with abstractions at the *city*, *province_or_state*, and *country* levels. Data in each class should be generalized to the same *location* level. *"How is class comparison performed?"* In general, the procedure is as follows:

1) **Data collection:** The set of relevant data in the database is collected by query processing and is partitioned respectively into a *target class* and one or a set of *contrasting classes*.
2) **Dimension relevance analysis:** If there are many dimensions, then dimension relevance analysis should be performed on these classes to select only the highly relevant dimensions for further analysis. Correlation or entropy-based measures can be used for this step.
3) **Synchronous generalization:** Generalization is performed on the target class to the level controlled by a user-or expert-specified dimension threshold, which results in a **prime target class relation**. The concepts in the contrasting class(es) are generalized to the same level as those in the prime target class relation, forming the **prime contrasting class(es) relation**.
4) **Presentation of the derived comparison:** The resulting class comparison description can be visualized in the form of tables, graphs, and rules. This presentation usually includes a "contrasting" measure such as count% (percentage count) that reflects the comparison between the target and contrasting classes. The user can adjust the comparison description by applying drill-down, roll-up, and other OLAP operations to the target and contrasting classes, as desired.

8.6 STATISTICAL MEASURES

For a successful data preprocessing, it is essential to have an overall picture of your data. Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

We start with *measures of* ***central tendency***, which measure the location of the middle or center of a data distribution. The most common data dispersion measures are the *range*, *quartiles*, and *inter-quartile range*; the *five-number summary* and *boxplots* and the *variance* and *standard deviation* of the data. Most statistical or graphical data presentation software packages include bar charts, pie charts, and line graphs. Other popular displays of data summaries and distributions include *quantile plots*, *quantile–quantile plots*, *histograms*, and *scatter plots*.

Measuring the central tendency: mean, median, and mode

There have various ways to measure the central tendency of data. Suppose that we have some attribute *X*, like *salary*, which has been recorded for a set of objects. Let $x_1, x_2, ..., ..., ..., x_N$ be the set of *observations* for *X*. Measures of central tendency include the mean, median, mode, and midrange.

The most common and effective numeric measure of the "center" of a set of data is the *(arithmetic) mean*. Let be a set of *N* values or *observations*, such as for some numeric attribute *X*, like *salary*. The **mean** of this set of values is:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_i + x_2 + ...+ x_N}{N} \qquad \textbf{(8.4)}$$

This corresponds to the built-in aggregate function, *average* (avg() in SQL), provided in relational database systems.

Example 8.1: Mean

Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

Using Eq. (8.4), we have,

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12}$$

$$= \frac{696}{12} = 58$$

Thus, the mean salary is \$58,000.

Sometimes, each value x_i in a set may be associated with a weight w_i. The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, we can compute:

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + ...+ w_N x_N}{w_1 + w_2 + ... + w_N} \qquad \textbf{(8.5)}$$

This is called the **weighted arithmetic mean** or the **weighted average**.

For skewed (asymmetric) data, a better measure of the center of data is the **median**, which is the middle value in a set of ordered data values. It is the value that separates the higher half of a data set from the lower half.

Example 8.2: Median

Let's find the median of the data from Example 8.1. The data are already sorted in increasing order. There is an even number of observations (i.e., 12); therefore, the median is not unique. It can be any value within the two middlemost values of 52 and 56 (that is, within the sixth and seventh values in the list). By convention, we assign the average of the two middlemost values as the median; that is the median is $54,000.

Suppose that we had only the first 11 values in the list. Given an odd number of values, the median is the middlemost value. This is the sixth value in this list, which has a value of $52000.

We can approximate the median of the entire data set (e.g., the median salary) by interpolation using the formula,

$$\text{median} = L_1 + \left(\frac{N/2 - (\sum \text{freq})_1}{\text{freq}_{\text{median}}} \right) \text{width} \qquad \textbf{(8.6)}$$

where, L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and *width* is the width of the median interval.

The *mode* is another measure of central tendency. The **mode** for a set of data is the value that occurs most frequently in the set. Therefore, it can be determined for qualitative and quantitative attributes. Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal**, and **trimodal**. In general, a data set with two or more modes is **multimodal**.

Example 8.3: Mode

The data from Example 8.1 are bimodal. The two modes are $52,000 and $70,000.

For unimodal numeric data that are moderately skewed (asymmetrical), we have the following empirical relation:

mean – mode ;: 3 × (mean – median) **(8.7)**

This implies that the mode for unimodal frequency curves that are moderately skewed can easily be approximated if the mean and median values are known.

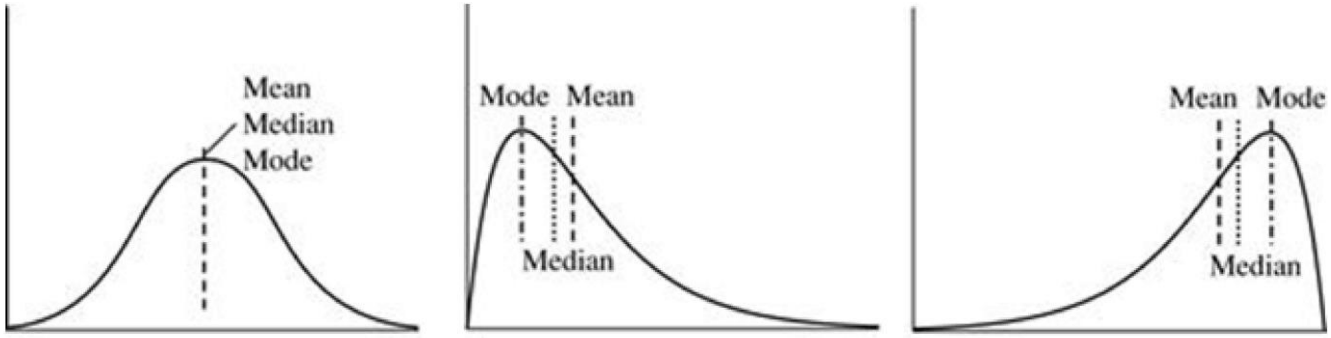
The **midrange** can also be used to assess the central tendency of a numeric data set. It is the average of the largest and smallest values in the set.

Example 8.4: Midrange

The midrange of the data of Example 8.1 is,

$$\frac{30,000+110,000}{2} = \$70,000$$

In a unimodal frequency curve with perfect **symmetric** data distribution, the mean, median, and mode are all at the same center value, as shown in Figure 8.1(a).



(a) Symmetric data (b) Positively skewed data (c) Negatively skewed date

Figure 8.2: Mean, median, and mode of symmetric versus positively and negatively skewed data.

Data in most real applications are not symmetric. They may instead be either **positively skewed**, where the mode occurs at a value that is smaller than the median (Figure 8.2b), or **negatively skewed**, where the mode occurs at a value greater than the median (Figure 8.2c).

Measuring the dispersion of data: range, quartiles, variance, standard deviation, and interquartile range: We now look at measures to assess the dispersion or spread of numeric data. The measures include range, quantiles, quartiles, percentiles, and the interquartile range. The five-number summary, which can be displayed as a boxplot, is useful in identifying outliers. Variance and standard deviation also indicate the spread of a data distribution.

Range, Quartiles, and Interquartile Range: To start off, let's study the *range*, *quantiles*, *quartiles*, *percentiles*, and the *interquartile* range as measures of data dispersion. Let x_1, x_2, …, ..., ..., x_N be a set of observations for some numeric attribute, X. The range of the set is the difference between the largest (max()) and smallest (min()) values.

Suppose that the data for attribute *X* are sorted in increasing numeric order. Imagine that we can pick certain data points so as to split the data distribution into equal-size consecutive sets, as in figure 8.3. These data points are called *quantiles*. **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal-size consecutive sets.

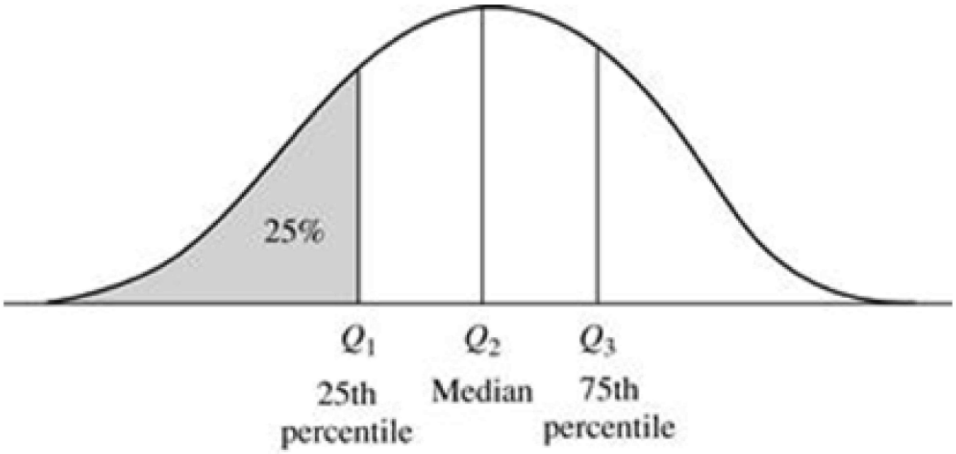


Figure 8.3: A plot of the data distribution for some attribute *X*

The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size consecutive subsets. The second quartile corresponds to the median.

The 2-quantile is the data point dividing the lower and upper halves of the data distribution. It corresponds to the median. The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as **quartiles**. The 100-quantiles are more commonly referred to as **percentiles.** The **first quartile**, denoted by Q_1, is the 25th percentile. It cuts off the lowest 25% of the data. The **third quartile**, denoted by Q_3, is the 75th percentile– it cuts off the lowest 75% (or highest 25%) of the data. The second quartile is the 50th percentile. As the median, it gives the center of the data distribution.

The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the **interquartile range** (**IQR**) and is defined as–

$$\text{IQR} = Q_1\text{-}Q_3 \qquad \textbf{(8.8)}$$

Example 8.5: Interquartile range

The quartiles are the three values that split the sorted data set into four equal parts. The data of Example 8.1 contain 12 observations, already sorted in increasing order. Thus, the quartiles for this data are the third, sixth, and ninth values, respectively, in the sorted list. Therefore, Q_1 = \$47,000 and Q_3 is \$63,000. Thus, the interquartile range is *IQR* = 63 – 47 = \$16,000.

Five-Number Summary, Boxplots, and Outliers: In the symmetric distribution, the median (and other measures of central tendency) splits the data into equal-size halves. This does not occur for skewed distributions. Therefore, it is more informative to also provide the two quartiles Q_1 and Q_3, along with the median. A common rule of thumb for identifying suspected **outliers** is to single out values falling at least 1.5 × IQR above the third quartile or below the first quartile.

Because Q_1, the median, and Q_3 together contain no information about the endpoints (e.g., tails) of the data, a fuller summary of the shape of a distribution can be obtained by providing the lowest and highest data values as well. This is known as the *five-number summary*. The **five-number summary** of a distribution consists of the median (Q_2), the quartiles Q_1 and Q_3, and the smallest and largest individual observations, written in the order of *Minimum*, Q_1, *Median*, Q_3, *Maximum*.

Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:

- Typically, the ends of the box are at the quartiles so that the box length is the interquartile range.
- The median is marked by a line within the box.

- Two lines (called *whiskers*) outside the box extend to the smallest (*Minimum*) and largest (*Maximum*) observations.

For exambple, let us take *All Electronics* data during a given time period. For branch 1, we see that the median price of items sold is \$80, Q_1 is \$60, and Q_3 is \$100. Notice that two outlying observations for this branch were plotted individually, as their values of 175 and 202 are more than 1.5 times the IQR here of 40.

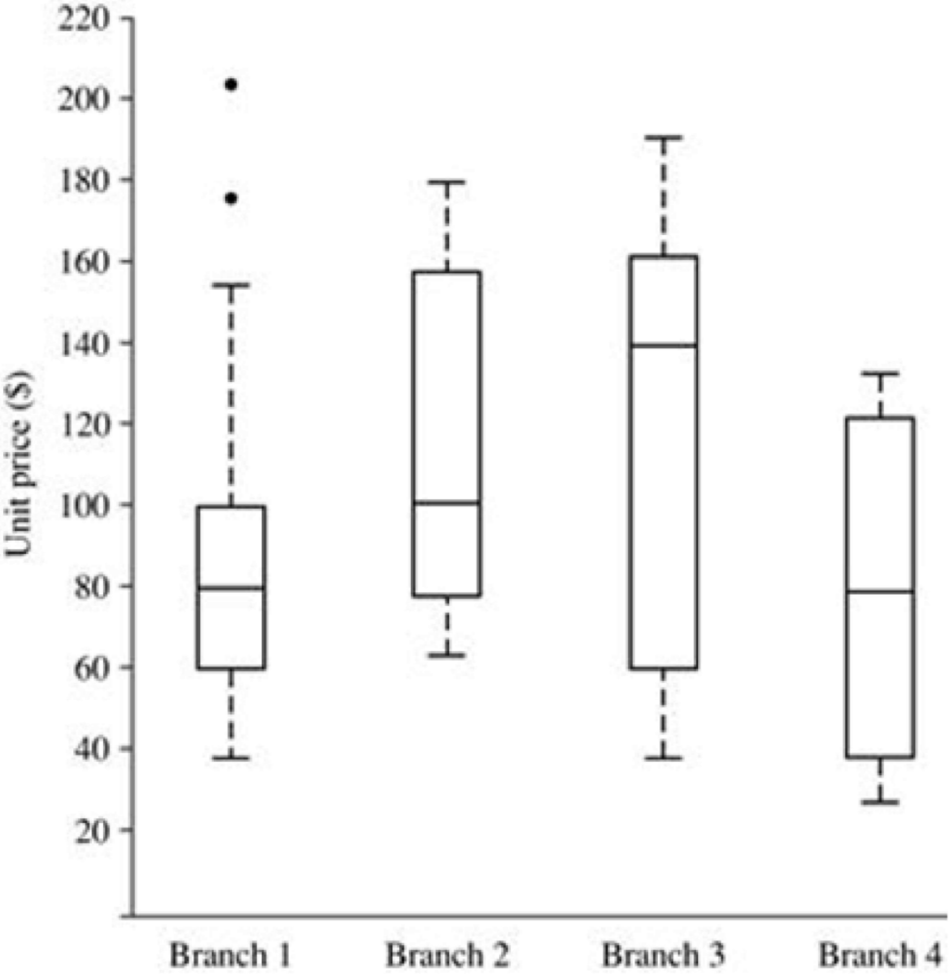


Figure 8.4: Boxplot for the unit price data for items sold at four branches of All Electronics during a given time period

Variance and Standard Deviation: Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is. A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

The **variance** of *N* observations, $x_1, x_2, ..., ..., ..., x_N$, for a numeric attribute *X* is:

$$Cr^{2=\frac{1}{N}} L_{i=1}^{N}\left(x_i - \bar{x}\right)^2 = I\left(\frac{1}{N} L_{i=1}^{N} x_2^{2}\right] / - \bar{x}^2 \qquad \textbf{(8.9)}$$

where, $\bar{x}$ is the mean value of the observations. The **standard deviation**, *σ*, of the observations is the square root of the variance, σ^2.

The basic properties of the standard deviation, *σ*, as a measure of spread are as follows:

- *σ* measures spread about the mean and should be considered only when the mean is chosen as the measure of center.
- $\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise, $\sigma > 0$.

The standard deviation is a good indicator of the spread of a data set. The computation of the variance and standard deviation is scalable in large databases.

Graphic displays of basic statistical descriptions of data: These include *quantile plots*, *quantile-quantile plots*, *histograms*, and *scatter plots*. Such graphs are helpful for the visual inspection of data, which is useful for data preprocessing. The first three of these show univariate distributions (i.e., data for one attribute), while scatter plots show bivariate distributions (i.e., involving two attributes).

Quantile Plot: A **quantile plot** is a simple and effective way to have a first look at a univariate data distribution. First, it displays all of the data for the given attribute. Second, it plots quantile information. Let x_i, for $i = 1$ to N, be the data sorted in increasing order so that x_1 is the smallest observation and x_N is the largest for some ordinal or numeric attribute X.

Let, fi = (i – 0.5) / N **(8.10)**

These numbers increase in equal steps of 1/*N*, ranging from $\frac{1}{2N}$ (which is slightly above 0) to $1 - \frac{1}{2N}$ (which is slightly below 1). On a quantile plot, x_i is graphed against f_i. This allows us to compare different distributions based on their quantiles. For example, given the quantile plots of sales data for two different time periods, we can compare their Q_1, median, Q_3, and other f_i values at a glance.

Figure 8.5 shows a quantile–quantile plot for *unit price* data of items sold at two branches of *All Electronics* during a given time period. Each point corresponds to the same quantile for each data set and shows the unit price of items sold at branch 1 versus branch 2 for that quantile.

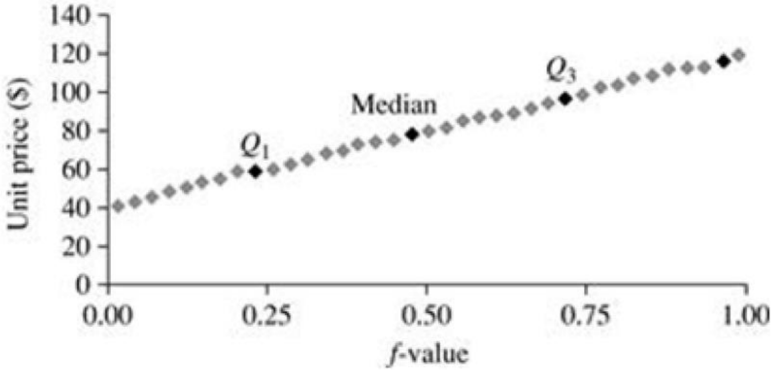


Figure 8.5: A quantile plot for the unit price data of Table 8.1.

Table 8.1: A Set of Unit Price Data for Items Sold at a Branch of All Electronics

Unit Price ($)	Count of Units Sold
40	2765
43	300
47	250
–	–
74	360
75	515
78	540
–	–
115	320
117	270
120	350

Quantile-Quantile Plot: A **quantile-quantile plot**, or **q-q plot**, graphs 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.

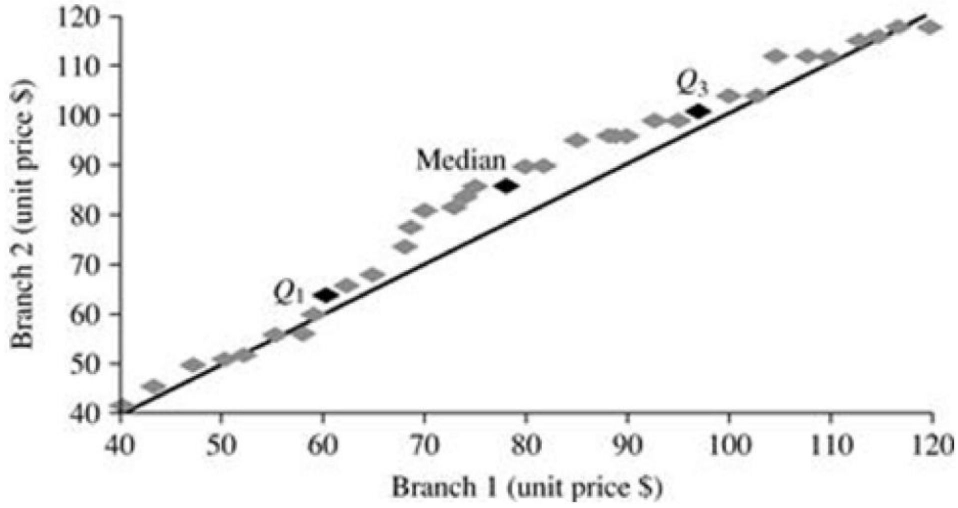


Figure 8.6: A q-q plot for unit price data from two *All Electronics* branches

Histograms: Histograms (or **frequency histograms**) are at least a century old and are widely used. “Histos” means pole or mast, and “gram” means chart, so a histogram is a chart of poles. Plotting histograms is a graphical method for summarizing the distribution of a given attribute, X. If X is nominal, such as *automobile_model* or *item_type*, then a pole or vertical bar is drawn for each known value of X. The height of the bar indicates the frequency (i.e., count) of that X value. The resulting graph is more commonly known as a **bar chart**.

. Figure 8.7 shows a histogram for the data set of Table 8.1, where buckets (or bins) are defined by equal-width ranges representing $20 increments and the frequency is the count of items sold.

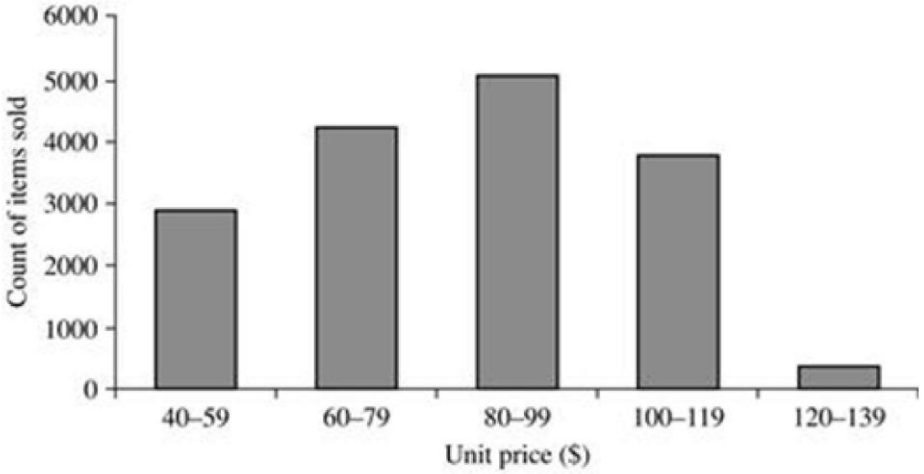


Figure 8.7: A histogram for the Table 8.1 data set

Scatter Plots and Data Correlation: A **scatter plot** is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes. To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane. Figure 8.8 shows a scatter plot for the set of data in Table 8.1.

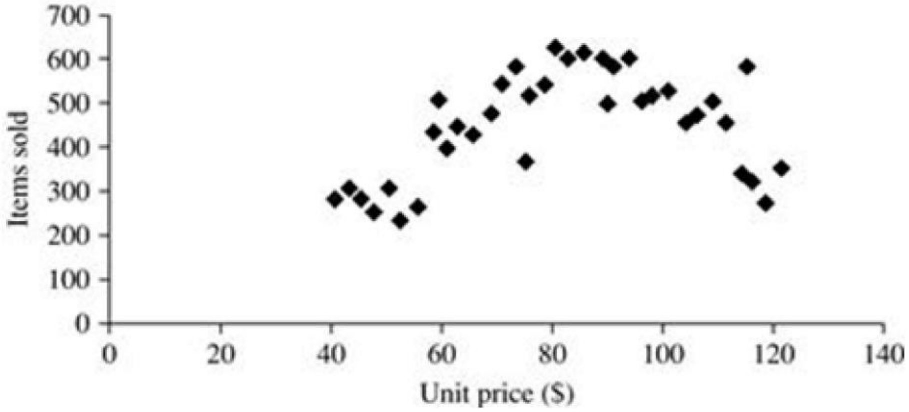


Figure 8.8: A scatter plot for the Table 8.1 data set

Two attributes, *X*, and *Y*, are **correlated** if one attribute implies the other. Correlations can be positive, negative, or null (uncorrelated). Figure 8.9 shows examples of positive and negative correlations between two attributes.

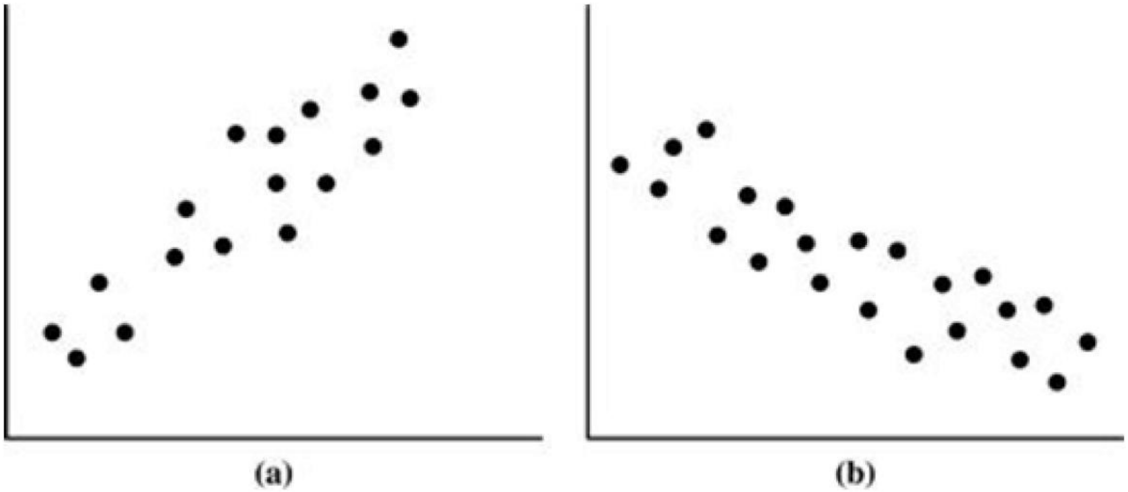


Figure 8.9: Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

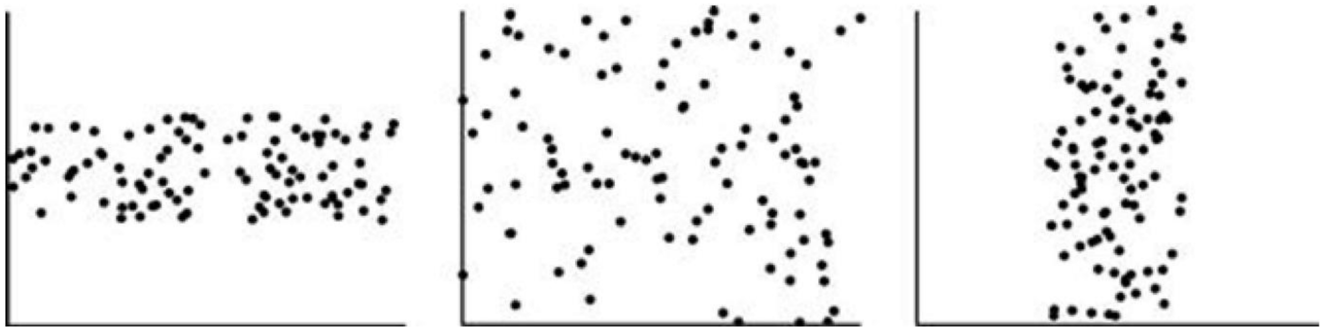


Figure 8.10: Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.

CHECK YOUR PROGRESS

Q.6: What is Class compassion?

....

....

....

Q.7: Define the properties of standard deviation.

....

....

....

....

....

Q.8: What is histogram?

....

....

....

Q.9: What are methods used for graphic display of statistical data?

....

....

8.7 LET US SUM UP

- Attributes are used to describe the objects.
- Quantitative and qualitative are the two main categories of attributes.
- Data Generalization have two approaches. Namely data cube approach and attribute oriented induction approach.
- With attribute relevance, we can quantify the attributes.
- With attribute relevance, information gain, gini index, uncertainty, correlation coefficient are calculated.
- Class comparison describes the difference between targeted class from its constructing class.
- Mean, mode, median are the measure of central tendency of statistical data.
- Dispersion of data we get by calculating range, quartiles, variance, standard deviation and interqurative range.
- Quartile plot, quartile-quartile plot, histogram and scattered plot are the methods used for graphics display of the statistical data.

8.8 FURTHER READING

1) Han, J., Pei, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

2) Pujari, A. K. (2001). Data Mining Techniques. Universities Press.

8.9 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: (a) Gender

Ans. to Q. No. 2: **Attribute removal** is a process based on the following rule: *If there is a large set of distinct values for an attribute of the initial working relation, but either (case 1) there is no generalization* *operator on the attribute or (case 2) its higher-level concepts are expressed in terms of other attributes, then the attribute should be removed from the working relation*.

Ans. to Q. No. 3: The general idea of attribute-oriented induction is to first collect the task-relevant data using a database query and then perform generalization based on the examination of the number of each attribute's distinct values in the relevant data set. This Process is called **data focusing.**

Ans. to Q. No. 4: Data collection, Preliminary Relevance analysis using conservative AOI, Remove irrelevant or weakly relevant attributes using the selected measure.

Ans. to Q. No. 5: The values of a Nominal attribute are name of things, some kind of symbols. Values of Nominal attributes represents some category or state. So its called as categorical attribute.

Ans. to Q. No. 6: Class discrimination or comparison (hereafter referred to as **class comparison**) mines descriptions that distinguish a target class from its contrasting classes

Ans. to Q. No. 7: Properties of standard deviation are:

- σ measures spread about the mean and should be considered only when the mean is chosen as the measure of center.
- $\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise, $\sigma > 0$.

Ans. to Q. No. 8: **Histograms** (or **frequency histograms**) are at least a century old and are widely used. "Histos" means pole or mast, and "gram" means chart, so a histogram is a chart of poles.

Ans. to Q. No. 9: Q*uantile plots*, Q*uantile–quantile plots*, *histograms*, and *scatter plots*.

8.10 MODEL QUESTIONS

Q.1: Define attribute.

Q.2: What is data generalization?

Q.3: Describe OLAP and attribute oriented induction approach of data generalization.

Q.4: Explain the reason behind attribute relevance.

Q.5: Explain the process of attribute relevance.

Q.6: Why is class comparison required?

Q.7: Define the different types of central tendency.

Q.8: Describe the different types of dispersion of data.

Q.9: Write about the methods to represent statistical data graphically.

*** ***** ***

UNIT 9: ASSOCIATION RULE MINING

UNIT STRUCTURE

9.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define association rule
- describe market basket analysis
- describe Apriori algorithm for association rule mining
- define multilevel association rule mining
- describe how correlation analysis is performed.

9.2 INTRODUCTION

In the previous unit, in the first block, we have discussed topics like attribute generalization and attribute relevance. We have also learned about different statistical measures in the previous unit. In this unit, we will learn about association rule mining and market basket analysis. We will also learn how apriori algorithm works. We will also cover the different multilevel association rules. The unit also covers how correlation analysis is performed from association mining. In the next unit, we will explore the concept of classification in detail.

9.3 ASSOCIATION RULE MINING

Association rule mining is a data mining technique that discovers the probability of the co-occurrence of items in a collection of data. The relationships between co-occurring items are expressed as association rules. Association rules are often used to analyse sales transactions. For example, it might be noted that customers who buy 'bread' at the grocery store often buy 'butter' at the same time. In fact, association analysis might find that 85% of the buyers that include 'bread' in their buying list also include 'butter'. This relationship could be formulated as the following rule.

'Bread' implies 'butter' with 85% confidence

This application of association modelling is called market-basket analysis. It is valuable for direct marketing, sales promotions, and for discovering business trends. Market-basket analysis can also be used effectively for store layout, catalogue design and cross-sell.

Association modelling has important applications in other domains as well. For example, in e-commerce applications, association rules may be used for web page personalization. An association model might find that a user who visits KKHSOU website and MHRD website is likely to also visit UGC website in 70% times in the same session. Based on this rule, a dynamic link could be created for users who are likely to be interested in the 3rdpage. The association rule could be expressed as follows.

KKHSOU website and MHRD website imply UGC website with 70% confidence.

9.4 MARKET BASKET ANALYSIS

Market Basket Analysis is the most typical example of association mining. Customers Invoices (Bill) are recorded in all the supermarkets. This database, known as the "market basket" database, consists of a large number of records on past transactions. A single record lists all the items bought by a customer in one sale. Knowing which items are sold together with which set of items gives these shops the freedom to adjust the store layout and the store catalogue to place the optimally concerning one another.

Example 9.1: Database with 4 items and 5 transactions

Table 9.1: Market Basket Dataset

Transaction ID	Milk	Bread	Butter	Beer
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

9.4.1 Some Useful Definitions

Before we discuss more about Association Rule Mining we need to be familiar with few terms that are going to be used in association rule mining

ITEMSET: An item set I is a set of n items represented as I = {i_1, i_2, i_3, …, ..., ..., i_n}. In the above example (Table 9.1) milk, bread, butter and beer are the items in the itemset I.

TRANSACTION: A transaction 't' consists of a transaction_id and a set of items which are present in that transaction. In the table 9.1 first transaction (Transaction_id 1) consist of only milk and butter.

DATASET: A dataset **D** is a set of 'n' transactions, where each transaction 't' contains a non-empty set of items represented as **D** = {t_1, t_2, t_3, …, ..., ..., t_n}

SUPPORT: Support refers to the frequency of an itemset. It is measured by the number of transactions in which the itemset appears. It is referred as Supp(X) where X is the set of items. In the Table 9.1, the support of {bread} is $\frac{3}{5}$, or 60%. Itemsets can also contain multiple items. For instance, the support of {milk, bread, butter} is $\frac{1}{5}$, or 20%.

CONFIDENCE: Confidence refers to how likely item Y is purchased when item X is purchased, expressed as {X ➔ Y}. This is measured by the number of transactions with item X, in which item Y also appears.

Confidence is defined as Conf (X ➔ Y) = $\frac{\text{Supp}(X \cup Y)}{\text{Supp}(X) \text{ x } \text{Supp}(Y)}$.

In Table 9.1, the confidence of {milk ➔ bread} is 100%

i.e., Supp (milk U bread) / Supp (milk) =>(40% / 40%) and confidence of {milk, bread ➔ butter} is 50%.

LIFT: The lift value is a measure of importance of a rule. The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule. The expected confidence of a rule is defined as the product of the support values of the rule body and the rule head divided by the support of the rule body. Lift is defined as Lift (X ➔ Y) = $\frac{\text{Supp}(X \cup Y)}{\text{Supp}(X) \text{ x } \text{Supp}(Y)}$

In Table 9.1, the lift of {milk ➔ bread} is 10/6 = 1.6

i.e., Supp (milk U bread) / [Supp (milk) x Supp (bread)]

$$= \frac{\frac{2}{5}}{\frac{2}{5} \text{x} \frac{3}{5}} = \frac{\frac{2}{5}}{\frac{6}{25}} = \frac{2}{5} \text{x} \frac{25}{6} = \frac{10}{6} = 1.6$$

A lift value greater than 1 means that item Y is *likely* to be bought if item X is bought, while a value less than 1 means that item Y is *unlikely* to be bought if item X is bought.

9.5 APRIORI ALGORITHM

Apriori is a frequent item set mining and association rule learning algorithm over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger item sets. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the transactional database. The Apriori algorithm was proposed by Agrawal and Srikant in 1994. Apriori is designed to operate on databases containing transactions. Given a threshold T, the Apriori algorithm identifies the item sets which are subsets of at least T transactions in the database.

Apriori uses a “bottom up” approach, where frequent subsets are combined to generate the candidates for next iteration as one item at a time. The algorithm terminates when no further successful extensions are possible.

Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length‘k’ from item sets of length ‘k-1’. Then it prunes the candidates which have an infrequent sub pattern in the pruning step by using the downward closure property. According to downward closure property “All Subsets of a frequent itemsets must be frequent”. After that, it scans the transaction database to determine frequent item sets among the candidates.

The algorithm is given below for a transaction database ‘T’, and a support threshold of ‘Th’. At each step, the algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma.

The Apriori algorithm is a collection of three different procedures, the main apriori algorithm, candidate generation and pruning. Bellow all three algorithms are discussed with details.

Algorithm 9.1: Apriori algorithm

Input: Transaction database and a user-defined minimum supportTh

Output: All frequent itemsets

```
Step 1:  $L_0$:= 0; $k$ := 1;
Step 2:  $C_1$:= {{$i$} | $i$ E $I$ }
Step 3:  Answer := 0
Step 4:  while $C_k$ -:t 0
Step 5:      read database and count supports for $C_k$
Step 6:      $L_k$:= {frequent itemsets in $C_k$}
Step 7:      $C_{k+1}$ := Apriori-gen ($L_k$)
Step 8:      $k$ := $k$ + 1
Step 9:   end while
Step 10:      Answer := Answer u $L_k$
Step 11:  return Answer
```

Apriori-gen procedure used above in the Apriori is combination of two procedures, *candidate generation*and a *prune* procedure. The *candidate generation* procedure combines two frequent *k*-itemsets, which have the same (*k*–1)-prefix, to generate a (k+1) itemset as a new preliminary candidate. Following the *candidate generation* procedure, the *prune* procedure is used to remove infrequent itemsets from the preliminary candidate set all itemsets '*c*' such that some *k*-subset of '*c*' is not a frequent itemset.

Algorithm 9.2: The *candidate generation* procedure of the Apriori-gen algorithm

Input: L_k, the set containing frequent item sets found in pass *k.*

Output: Preliminary candidate set C_{k+1}.

```
Step 1: for i from 1 to $|L_k–1|$
Step 2:     for j from i + 1 to $|L_k|$
Step 3:         if$L_k$.itemset$_i$and $L_k$.itemset$_j$ have the same(k–1)-prefix
Step 4:             $C_{k+1}$ := $C_{k+1}$ u { $L_k$.itemset$_i$ u $L_k$.itemset$_j$}
Step 5: break
```

Algorithm 9.3: The *prune* procedure of the Apriori-gen algorithm

Input: Preliminary candidate set C_{k+1}generated from the *join* procedure above.

Output: final candidate set C_{k+1} which does not contain any infrequent itemset.

```
Step 1:   for all itemsetsc in $C_{k+1}$
Step 2:       for all k-subsets  s of c
Step 3:       if s � $L_k$
Step 4:       delete c from $C_{k+1}$
```

Example 9.2: Consider the following market basket dataset and minimum threshold value as 3 (60%). If we apply the Apriori Algorithm then we get the following results.

Transaction ID	Items Bought
T1	{Mango, Onion, Nintendo, Key-chain, Eggs, Yo-yo}
T2	{Doll, Onion, Nintendo, Key-chain, Eggs, Yo-yo}
T3	{Mango, Apple, Key-chain, Eggs}
T4	{Mango, Umbrella, Corn, Key-chain, Yo-yo}
T5	{Corn, Onion, Key-chain, Ice-cream, Eggs}

For simplify the example we use M- for Mango, O- for Onion, N- for Nintendo etc. After simplifying the dataset looks as follows

Transaction ID	Items Bought
T1	{M, O, N, K, E, Y }
T2	{D, O, N, K, E, Y }
T3	{M, A, K, E}
T4	{M, U, C, K, Y }
T5	{C, O, K, I, E}

The above dataset can also be converted to binary dataset as follows:

Transaction ID	M	O	N	K	E	Y	D	A	U	C	I
T1	1	1	1	1	1	1	0	0	0	0	0
T2	0	1	1	1	1	1	1	0	0	0	0
T3	1	0	0	1	1	0	0	1	0	0	0
T4	1	0	0	1	0	1	0	0	1	1	0
T5	0	1	0	1	1	0	0	0	0	1	1

Using the same procedure we can find the frequent itemsets from binary datasets. In experiments we use only the binary datasets since it is easy to calculate frequencies from binary datasets.

Each pass of Apriori algorithm is consist of four steps: Frequency counting, Frequent itemset selection, Candidate generation and Pruning. In the following workout all these four steps are discussed in detail.

Step 1: *Frequency Counting*– Count the number of transactions in which each item occurs, this is also known as frequency counting for one itemsets.

Item	Number of Transactions
M	3
O	3
N	2
K	5
E	4
Y	3
D	1
A	1
U	1
C	2
I	1

Step 2: *Frequent Item Selection*– Depending on the user provided minimum threshold value frequent items of each pass is selected. Here our predefined threshold value is 3. So frequent one items will be those items which have frequency 3 or above. Following are the frequent one items

Item	Number of Transactions
M	3
O	3
K	5
E	4
Y	3

Step 3: *Candidate Generation*– from the frequent one item sets the candidates for the next pass will be generated by following the *Candidate Generation* algorithm. Here all the frequent one items are combined to generate the probable two item candidate sets. In the following table all such two item candidate sets are listed

Item pairs
MO
MK
ME
MY

OK
OE
OY
KE
KY
EY

Step 4: *Pruning*– Since two item candidate sets are generated only from the one item frequent sets so pruning does not prune any two item candidate sets. So same candidate sets are passed to the second pass for frequency counting.

Step 5: *Frequency Counting*– In the second pass, the frequency counting is performed for two item candidate sets. Here, we count how many times each pair of items are bought together. For example, MO is bought together only in the transaction T1{M,O,N,K,E,Y}

While MK are bought together 3 times in T1{M,O,N,K,E,Y}, T3{M,A,K,E} AND T4{M,U,C,K,Y}

After counting the frequencies of all such item pairs we get,

Item Pairs	**Number of Transactions**
MO	1
MK	3
ME	2
MY	2
OK	3
OE	3
OY	2
KE	4
KY	3
EY	2

Step 6: *Frequent Item Selection*– After observing the frequencies in the above table the frequent two itemsets are found as follows:

Item Pairs	Number of transactions
MK	3
OK	3
OE	3
KE	4
KY	3

These are the pairs of items frequently bought together.

Step 7: *Candidate Generation*– To generate the candidates for next pass (i.e., three itemsets) we have look at the following conditions.

The three item candidates will be generated only if prefix of both the frequent itemsets are same and differ by only the last item for example

OK & OE are frequent itemsets and OK & OE differ by K & E(only the last items), this gives OKE (Common + differ items or in other words {O,K} u {O,E})

Similarly from KE and KY, we get, KEY

After observing all the frequent two items we get candidates for next pass are

Item Pairs
OKE
KEY

Step 8: *Pruning*– In the pruning step all the subsets of a candidates are checked whether they are frequent or not, if not then that item is removed from the candidate sets.

Here, subsets of OKE are {OK},{KE},{OE} and all three subsets are frequent so this three item sets are not pruned.

For KEY the subsets are {KE},{EY}and{KY} here {KE} and{KY} are frequent but {EY} is not frequent so the candidate KEY is removed from candidate list.

The final candidate list after pruning is as follows

Item Pairs
OKE

Step 9: *Frequency Counting*– In the third pass the frequency counting is performed for three item candidate sets.

After counting the frequencies of all such item pairs we get

Item Pairs	Number of transactions
OKE	3

Step 6: *Frequent Item Selection*– After observing the frequencies in the above table the frequent three itemsets are found as follows

Item Pairs	Number of transactions
OKE	3

Since only one frequent item pairs left so no more candidate sets can be generated hence the algorithm stops here.

CHECK YOUR PROGRESS

Q.1: What does Apriori algorithm do?

a) It mines all frequent patterns through pruning rules with lesser support

b) It mines all frequent patterns through pruning rules with higher support

c)

d)

Q.2: What do you mean by support (A)?

a) Total number of transactions containing A

b) Total Number of transactions not containing A

c) Number of transactions containing A / Total number of transactions

d) Number of transactions not containing A / Total number of transactions

Q.3: How do you calculate Confidence (A ➔ B)?

a) Support(A n B) / Support (A)

b) Support(A n B) / Support (B)

c) Support(A u B) / Support (A)

d) Support(A u B) / Support (B)

Q.4: Which of the following is direct application of frequent itemset mining?

a) Social Network Analysis

b) Market Basket Analysis

c) Outlier Detection

d) Intrusion Detection

9.6 MULTILEVEL ASSOCIATION RULES

For many applications,it is difficult to find strong associations among data items at low or primitive levels of abstraction due to the diversity of data in multidimensional space. Strong associations discovered at high concept levels that might represent common sense knowledge. However, what may represent common sense to one user may seem novel to another. Therefore, data mining systems should provide capabilities to mine association rules at multiple levels of abstraction and traverse easily among different abstraction spaces.

Association rules generated from mining data at multiple abstraction levels are called multiple-level or multilevel association rules. Multilevel association rules can be mined efficiently using concept hierarchies under a support-confidence framework.

9.6.1 Approaches to Mining Multilevel Association Rules

In general a top-down strategy is employed,where counts are accumulated for the calculation of frequent itemsets at each concept level, starting at the highest concept level and working towards the lower, more specific concept levels, until no more frequent itemsets can be found. That is, when all frequent itemsets at all the concept levels are mined and no more frequent itemsets can be generated the algorithm stops. For each level,any algorithm for discovering frequent itemsets may be used, such as Apriori or its variations by using following two approches.

- **Using uniform minimum support for all levels:** The same minimum support threshold is used when mining at each level of abstraction. When a uniform minimum support threshold is used, the search procedure is simplified. The method is also simple in that users are required to specify only one minimum support threshold. An optimization technique can be adopted, based on the knowledge that an ancestor is a superset of its decedents;the search avoids examining itemsets containing any item whose ancestors do not have minimum support.

The uniform support approach,however, has some drawbacks. It is unlikely that items at lower abstraction levels will occur as frequently as those at higher abstraction levels. If the minimum support threshold is set too high, it could miss some meaningful associations occurring at low abstraction levels. If the threshold is set too low, it may generate many uninteresting associations occurring at high abstraction levels. This leads to the next approach.

- **Using reduced minimum support at lower levels:** Each level of abstraction has its own minimum support threshold. The threshold value is smaller for lower abstraction level, and higher for higher level of abstraction. For mining multiple-level associations with reduced support,there are a number of alternative search strategies such as Level-by-Level independent where a breadth first search is performed without considering the knowledge of previous frequent itemsets. Here, each data is checked regardless of whether or not its parent is frequent or not. Next approach is level-cross-filtering by one item where an item at the i^{th} level is examined if and only if its parent node at the $(i-1)^{th}$ level is frequent. If a node is frequent then its children are examined; otherwise, its descendants are not considered in the next search. Next one is level-cross filtering by -K-itemset where instead of one item in the i^{th} level k-itemsets are examined, if and only if the corresponding parent k-itemset at the $(i-1)^{th}$ level is frequent then only the candidate k-itemsets are considered at i^{th} level.

- **Using item or group-based minimum support:** Because of users or experts often have the insight knowledge about a particular item groups, which one is more important than others, it is sometimes more beneficial to set up a user-specific item group-based minimal support thresholds when mining multilevel rules. For example, a user could set up the minimum support thresholds based on product of interest such as by setting particularly low support thresholds for frying pan and rice cooker to pay particular attention on the association patterns containing these two items.

  For mining patterns with mixed items from groups with different support thresh-olds, usually the lowest support threshold among all the participating groups is taken as the support threshold for mining. This will avoid filtering out valuable patterns containing items from the group with the lowest support threshold. In the meantime, the minimal support threshold for each individual group should be kept to avoid generating uninteresting association rules.

  For mining these types of datasets with different groups of data efficient methods can be developed or extend the exsisting methods to mine these type of datasets.

9.6.2 Checking for Redundant Multilevel Association Rules

A serious side effect of mining multilevel association rules is that it generates many redundant rules across all the multiple abstraction levels. Also concepts of hierarchy are useful in data mining since they permit the discovery of knowledge at different levels of abstraction, such as multilevel association rules. However, when multilevel association rules are mined,some of the rules found will be redundant due to the ancestors at the previous level. So special precisions are needed to be taken up to reduce generation of such redundant rules.

9.7 ASSOCIATION MINING AND CORRELATION ANALYSIS

For filtering uninteresting association rules support and confidence are not sufficient. To overcome this correlation measure is used along with support and confidence for mining association rules.

Correlation rule is used measure not only from the support and confidencebut also it uses the correlation between itemsets. There are many different correlation measures from which we can choose for rule mining.

Lift is a simple correlation measure. The definition of lift is discussed in section 9.4.1. If the value of lift is less than 1, then the occurrence of A is negatively correlated with the occurrence of B, meaning that the occurrence of one likely leads to the absence of the other one. If the resulting value is greater than 1, then A and B are positively correlated, which means 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. Lift is usefull to filter out misleading "strong" associations rules of the form A ⇒ B.

The second correlation measure that we study is the x^2 measure. In (chi-square) test also correlation relationship between two attributes, A and B can be discovered. To compute the x^2 value, we take the squared difference between the observed and expected value for a pair in the contingency table, divided by the expected value. This amount is summed for all such pairs in the contingency table.

Suppose A has c distinct values, namely a_1, a_2, ..., ..., ..., a_c. B has r distinct values, namely b_1, b_2, ..., ..., ..., b_r. The data tuples described by A and B can be shown as a contingency table, with the c values of A making up the columns and the rvalues of B making up the rows. Let (A_i, B_j) denote the joint event that attribute A takes on value a_i and attribute B takes on value b_j. Each and every possible (A_i, B_j) joint event has its own cell in the table. The x^2 value (also known as the Pearson x^2 statistic) is computed as:

$$x^2 = \sum_{i=1}^{c}\sum_{j=1}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed frequency (i.e., actual count) of the joint event (A_i, B_j) and e_{ij} is the expected frequency of (A_i, B_j) which can be computed as:

$$e_{ij} = \frac{\mathbf{count(A = a_i) x count(B = b_j}}{\mathbf{) n}}$$

Here, n is the number of data tuples, *count* (A = a_i) is the number of tuples having valuea_i for A, and *count* (B = b_j) is the number of tuples having value b_j for B. The sum in the equation for computing x^2 is computed over all of the *r* × *c* cells. Note that the cells that contribute the most to the value are those for which the actual count is very different from the expected.

The statistic tests the hypothesis that A and B are independent, that is, there is nocorrelation between them. The test is based on significance level, with *(r* – 1*)* × *(c* – 1)degrees of freedom.

CHECK YOUR PROGRESS

Q.5: Why is correlation analysis important?

a) To make apriori memory efficient

b) To weed out uninteresting frequent itemsets

c) To find large number of interesting itemsets

d) To restrict the number of database iterations

Q.6: Lift value 0 means?

a) A is negatively correlated with the occurrence of B

b) A and B are positively correlated

c) A and B are not related to each other

d) None of the above

Q.7: For multilevel association rule mining

a) Threshold is not used

b) Minimum threshold is computed automatically

c) Multiple minimum threshold is used

d) None of the above

Q.8: x^2 (chi-square) test is used to–

a) check the validity of a rule

b) compute the frequency count

c) compute the validity of an itemset

d) check whether an itemset is in the transaction or not

9.8 LET US SUM UP

- Association rule mining is a data mining technique that discovers the probability of the co-occurrence of items in a collection of data.
- Association rules are often used to analyse sales transactions.
- One of the applications of association modelling is market-basket analysis. Which is valuable for direct marketing, sales promotions, and for discovering business trends.
- Association modelling has important applications in other domains as well. For example, in e-commerce applications, association rules may be used for Web page personalization.
- Apriori is a frequent item set mining and association rule learning algorithm over transactional databases.
- Apriori proceeds by identifying the frequent individual items in the database and extending them to larger item sets.
- The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the transactional database.
- Apriori uses a “bottom up” approach, where frequent subsets are combined to generate the candidates for next iteration as one item at a time.
- The Apriori algorithm is a collection of three different procedures, the main apriori algorithm, candidate generation and pruning.
- Association rules generated from mining data at multiple abstraction levels are called multiple-level or multilevel association rules. Multilevel association rules can be mined efficiently using concept hierarchies under a support-confidence framework.

- For mining multilevel association rule mining one of these three methods are used, using uniform minimum support for all levels, using reduced minimum support at lower levels or using item or group-based minimum support
- A serious side effect of mining multilevel association rules is that it generates many redundant rules across all the multiple abstraction levels.
- For filtering uninteresting association rules support and confidence are not sufficient. To overcome this correlation measure is used along with support and confidence for mining association rules.
- Correlation rule is used measure not only from the support and confidence but also it uses the correlation between itemsets.
- Lift and x^2 (chi-square) test are the two common approach for correlation measure.

9.9 FURTHER READING

1) Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.
2) Han, J., Pei, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
3) Tan, P. N. (2018). *Introduction to Data Mining*. Pearson Education India.

9.10 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1:	(a)	**Ans. to Q. No. 2:**	(c)
Ans. to Q. No. 3:	(a)	**Ans. to Q. No. 4:**	(b)
Ans. to Q. No. 5:	(b)	**Ans. to Q. No. 6:**	(c)
Ans. to Q. No. 7:	(c)	**Ans. to Q. No. 8:**	(a)

9.11 MODEL QUESTIONS

Q.1: What is support and confidence? Explain with example.

Q.2: What is association rule? Explain briefly.

Q.3: Consider the database in table 9.1 and suppose min support threshold is 2. Use the Apriori algorithm to generate all the frequent itemsets in the dataset.

Q.4: Discuss the different methods used for multilevel association rule mining?

Q.5: What is Lift? Explain with example.

Q.6: Explain the x^2 (chi-square) test? What is the use of it in rule mining?

Q.7: Explain in detail:

a) Market Basket analysis

b) Multilevel association rule mining

c) Correlation analysis

*** ***** ***

UNIT 10: CLASSIFICATION

UNIT STRUCTURE

10.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define classification
- describe the basic techniques of data classification
- explain what is decision tree
- explain how to build decision tree classifiers
- describethe different measures for attributeselection for decision tree.

10.2 INTRODUCTION

In the previous unit, we have learned about association rule mining, market basket analysis and different multilevel association rules. We have also learned about how apriori algorithm works and how correlation analysis is performed from association mining. In this unit, we will discuss about Classification and Decision trees. Different attribute selection measures like entropy and information gain are also discussed in this unit. In the next unit, we will explore the concept of prediction in detail.

10.3 CLASSIFICATION

Classification is a data analysis technique used to classify large amount of data into different predefined classes. This technique is used to predict category of data and assign class labels. For example, we can build a classification model to categorize network traffic data into different classes such as attack or normal data. Such analysis provides us a better understanding of the large amount of data. Many classification methods have been proposed by researchers in machine learning, pattern recognition, and statistics. Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has developed scalable classification and prediction techniques capable of handling large amounts of disk-resident data. Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis etc.

A network administrator is interested to know whether the incoming request is “attack” or “normal data” or a bank loan officer needs analysis which loan applicants are “safe” and which are “risky” for the bank. A medical researcher wants to analyze MRI image for identifying “tumours”. In each of these examples, the data are classified according to some predefined criteria, where a classifier is constructed to predict class (category) of the data, such as “attack” or “normal” in network administrator or “safe” or “risky” for the loan application data or “tumour” or “not a tumour” for the MRI data. These categories can be represented by discrete values, where the ordering among values has no meaning.

Regression analysis is a statistical methodology that is most often used for numeric prediction; hence classification and regression are used synonymously. Some other methods for numeric prediction also exist. Classification and numeric prediction are the two major types of prediction problems.

Data classification is a two-step process, *learning step* and *classification step*. In the learning step the classification model is constructed and in the classification step class labels are predicted by the classification model.

In the *learning step*, a classifier is built to classify data into a predetermined set of data classes or concepts. This phase is also known as training phase, where a classification algorithm builds the classifier by analyzing the training dataset. The training dataset is also similar to normal datasets except that each tuple in the training dataset has a predefined class label. Using this predefined class labels learning phase builds the classifier by using any of the classification algorithms. The class label is discrete-valued and unordered. The class label indicates the data category of the tuple.

Since the class label of each training tuple is provided in the training dataset classification problem is termed as *supervised learning*. It contrasts clustering as an unsupervised learning method where no information about the class label and number of class is provided beforehand. For example, in a network traffic data if we did not have the class label as which data is normal or which one is attack for the training data, we may use clustering to try to groups the data based on the clustering algorithm. Clustering is discussed in UNIT 13.

In the first step of classification(learning step), data or tuples of training data sets are mapped with the associated class labels. In the second step, these mappings are used to predict the associated class label of a tuple of which class label is unknown. In the first level, we design the classification function to specify the data classes. Typically, this mapping or functions are represented in the form of classification rules, decision trees, or mathematical formula etc. These rules are used to categorize the future data tuples, as to well as provide deeper insight into the data contents.

Second step classification model is used for classification of data. First, the predictive accuracy of the classifier is estimated to check the correctness of the classifier. If we use the training set to measure the accuracy of the classifier’s, this is likely be come up with 100% accuracy, because the classifier is designed from the same dataset.Therefore, in some datasets, apart from the training dataset one more dataset is provided, known as testing data. They are independent of the training datasets, meaning that they were not used in the construction of the classifier. By using this dataset the classification model is tested for its accuracy and correctness.

The accuracy of a classifier on a given test set is the ratio of number of data that are correctly classified and the total number of data in the test set. For computation of correctness, the associated class label of each test tuple is compared with the computed class label obtained by the classifier. If the accuracy of the classifier is acceptable, then the classifier can be used for future classification of data.

CHECK YOUR PROGRESS

Q.1: Classification is ....

a) Supervised learning b) Unsupervised learning

c) Semi supervised learning d) None of the above

Q.2: Classification classifies the data and assigns the .... of the data.

a) Data type b) File type

c) Class label d) Classification algorithm

Q.3: Classification and regression are used synonymously.

a) True b) False

Q.4: Testing dataset is used to design the classifier

a) True b) False

Q.5: Data classification is a two-step process,........................ and a classification step.

a) Testing Phase b) Training Phase

c) Learning Phase d) None of the above

10.4 DECISION TREE

Decision Tree Analysis is a general, predictive modelling tool that has applications spanning a number of different areas. In general, decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; edges represent the answers to the question; and the leaves represent the actual output or class label. They are used in non-linear decision making with simple linear decision surface.

Decision trees classify the problem dataset by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the problem dataset. Each node in the tree acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. This process is recursive in nature and is repeated for every subtree.

Let us discuss the decision tree problem with the following example. Let’s assume we want to play badminton on a particular day — how we will decide whether to play or not. Let’s say we go out and check if it’s hot or cold, check the speed of the wind and the humidity, how the weather is i.e. is it sunny, cloudy or rainy. We take all these factors into account to decide whether we will play or not.

Let us assume that we record all these data for last 10 days as in the table 10.1 given below.

Table 10.1 Observations of the last ten days

Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

Now, we may use this table to decide whether to play or not. But, if the weather pattern on that particular day does not match with any of rows in the table then it creates a problem. A decision tree would be a great way to represent data like this because it takes into account all the possible paths that can lead to the final decision by following a tree-like structure.

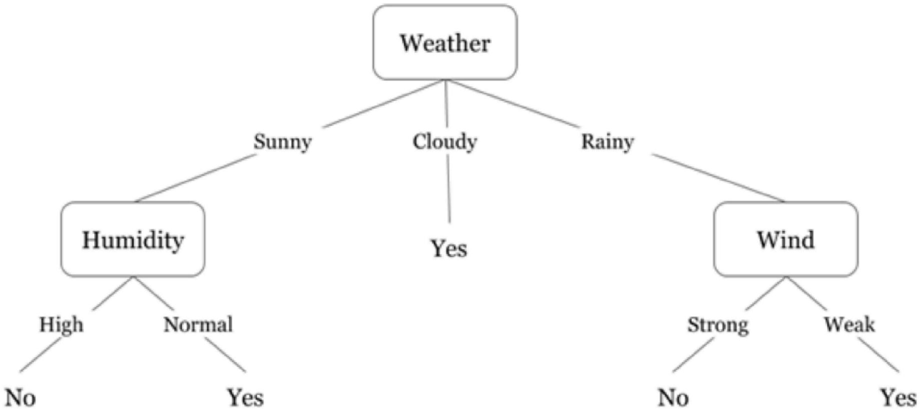


Figure 10.1: A decision tree for the concept play badminton

Figure 10.1 illustrates a learned decision tree. We can see that each node represents an attribute or feature and the branch from each node represents the outcome of that node. In each situation the leaves of the tree represents the final decision. If features are continuous; internal nodes can test the value of a feature against a threshold (see Figure. 10.2)

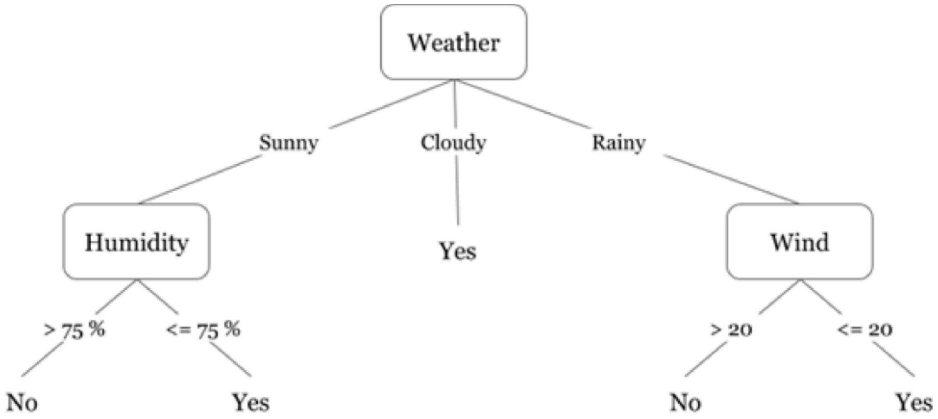


Figure 10.2: Decision tree for Badminton Play (with continuous attributes)

Incorporating continuous valued attributes: Our initial example is restricted to attributes that take on a discrete set of values. One way to make the decision tree more useful is with continuous variables. Let's take our example of *Play Badminton* the temperature is continuous; we could test the information gain of certain partitions of the temperature values, such as temperature > 42.5. Typically, whenever the classification changes from no to yes or yes to no, the average of the two temperatures is taken as a potential partition boundary.

Table 10.2: Observations of the last ten days

Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	80	High	Weak	No
2	Cloudy	66	High	Weak	Yes
3	Sunny	43	Normal	Strong	Yes
4	Cloudy	82	High	Strong	Yes
5	Rainy	65	High	Strong	No
6	Rainy	42	Normal	Strong	No
7	Rainy	70	High	Weak	Yes
8	Sunny	81	High	Strong	No
9	Cloudy	69	Normal	Weak	Yes

In the above data since 42 corresponds to No and 43 corresponds to Yes, 42.5 becomes a candidate. If any of the partitions end up exhibiting the greatest information gain, then it is used as an attribute and temperature is removed from the set of potential attributes to split on.

10.4.1 Decision Tree Algorithms

There are multiple algorithms written to build a decision tree, which can be used according to the problem characteristics we are trying to solve. A few of the commonly used algorithms are listed below:

- ID3
- C4.5
- CART
- CHAID (CHi-squared Automatic Interaction Detector)
- MARS
- Conditional Inference Trees

Though the methods are different for different decision tree building algorithms but all of them work on the principle of greediness.

Algorithms try to search for a variable which gives the maximum information gain or divides the data in the most homogenous way.

The basic algorithm used in decision trees is known as the ID3 (by Quinlan) algorithm. The ID3 algorithm builds decision trees using a top-down, greedy approach. Briefly, the steps to the algorithm are:

- Select the best attribute ➔ A
- Assign A as the decision attribute (test case) for the **NODE**.
- For each value of A, create a new descendant of the **NODE**.
- Sort the training examples to the appropriate descendant node leaf.
- If examples are perfectly classified, then STOP else iterate over the new leaf nodes.

10.5 ATTRIBUTE SELECTION MEASURE

An attribute selection measure is a heuristic for selecting the splitting criterion that “best” separates a given data partition, D, of class-labelled training tuples into individual classes. If we were to split D into smaller partitions according to the outcomes of the splitting criterion, ideally each partition would be pure (i.e., all the tuples that fall into a given partition would belong to the same class). Conceptually, the “best” splitting criterion is the one that most closely results in such a scenario. Attribute selection measures are also known as splitting rules because they determine how the tuples at a given node are to be split.

The attribute selection measure provides a ranking for each attribute describing the given training tuples. The attribute having the best score for the measure is chosen as the splitting attribute for the given tuples. If the splitting attribute is continuous-valued or if we are restricted to binary trees, then, respectively, either a split point or a splitting subset must also be determined as part of the splitting criterion. The tree node created for partition D is labelled with the splitting criterion; branches are grown for each outcome of the criterion, and the tuples are partitioned accordingly. This section describes three popular attribute selection measures– *information gain*, *gain ratio*, and *Gini index*.

Now, the next big question is how to choose the best attribute. For ID3, the best attribute is selected based on the value of which attribute has the most *information gain*, a measure that expresses how well an attribute splits that data into groups based on classification.

10.5.1 Entropy & Information Gain

The word entropy is borrowed from Thermodynamics which is a measure of variability or chaos or randomness. Shannon extended the thermodynamic entropy concept in 1948 and introduced it into statistical studies and suggested the following formula for statistical entropy:

$$H = -\sum_{j=1}^{n} P_j \ln P_j$$

where, H is the entropy in the system which is a measure of randomness.

Assuming, rolling a fair coin and we want to know the entropy of the system. As per the formula given–

entropy would be equals to– *[0.5 ln(0.5) + 0.5 ln(0.5)]* = 0.69; which is the maximum entropy which can occur in the system. In other words, there will be maximum randomness in our dataset if the probable outcomes have the same probability of occurrence.

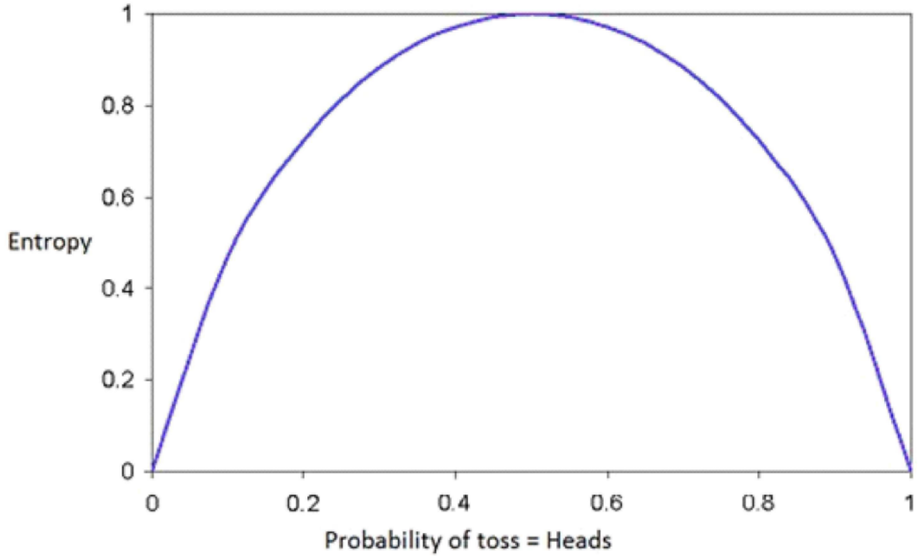


Figure 10.3: Example of Entropy Measure

The graph shown in figure 10.3 shows the variation of Entropy with the probability of a class. We can clearly see that Entropy is maximum when probability of either of the classes is equal. Now, we can understand that when a decision algorithm tries to split the data, it selects the variable which will give us maximum reduction in system Entropy.

Information Gain: The information gain is based on the decrease in entropy after a data-set is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain.

$$\text{Gain}(T, X) = \text{Entropy}(T) - \text{Entropy}(T, X)$$

where *Entropy*(T, X) is the conditional entropy of T given the value of attribute X.

10.5.2 Alternative Measures for Selecting Attributes

The information gain formula used by ID3 algorithm treats all of the variables the same regardless of their distribution and their importance. This is a problem when it comes to continuous variables or discrete variables with many possible values because training examples may be few and far between for each possible value, which leads to low entropy and high information gain by virtue of splitting the data into small subsets but results in a decision tree that might not generalize well.

One way to avoid this is to use some other measure to find the best attribute instead of information gain. An alternative measure to information gain is *gain ratio*. Gain ratio tries to correct the information gain’s bias towards attributes with many possible values by adding a denominator to *information gain* called *split information*. *Split Information* tries to measure how broadly and uniformly the attribute splits the data. One more measure for selecting the split attribute is Gini Index. Both measures are discussed below.

Gain Ratio: Soon after the development of entropy mathematicians realized that information gain is biased toward multi-valued attributes and to conquer this issue, “Gain Ratio” came into picture which is more reliable than information gain. The gain ratio can be defined as:

$$\text{Gain Ratio} = \frac{\text{Information Gain}}{\text{Split Info}}$$

where Split info can be defined as:

$$- \sum_{i=0}^{n} \text{Di}\log_2 \text{Di}$$

Assuming that we are dividing our variable into ‘n’ child nodes and Di represents the number of records going into various child nodes. Hence, gain ratio takes care of distribution bias while building a decision tree.

Gini Index: There is one more metric which can be used while building a decision tree i.e., Gini Index (Gini Index is mostly used in CART). Gini index measures the impurity of a data partition K, formula for Gini Index can be written down as:

$$\text{Gini}(K) = 1 - \sum_{i=1}^{n} P_i^2$$

where m is the number of classes, and P_i is the probability that an observation in K belongs to the class. Gini Index assumes a binary split for each of the attribute in S, let say T_1 & T_2. The Gini index of K given this partitioning is given by:

$$\text{Gini}_S(K) = \frac{T_1}{T}\text{Gini}(T_1) + \frac{T_2}{T}\text{Gini}(T_2)$$

which is nothing but a weighted sum of each of the impurities in split nodes. The reduction in impurity is given by:

$$\text{Gini}(K) - \text{Gini}_S(K)$$

Similar to Information Gain & Gain Ratio, split which gives us maximum reduction in impurity is considered for dividing our data.

Here, we have discussed all 3 commonly used metrics. However, confusion arises when we have to choose any one of them. There are a few drawbacks associated with all 3 of the metrics. These are summarized in the table below:

Table 10.3: Metrics and their drawbacks

Metrics	Drawback
Information Gain	Information Gain is biased towards multivariate attributes.
Gain Ratio	Gain Ratio generally prefers the unbalanced split of data where one of the child node has more number of entries compared to the others.
Gini Index	With more than 2 categories in the dataset, Gini Index gives unfavorable results. Apart from that it favors the split which results into equal sized children.

Many other attribute selection measures have been proposed. CHAID, a decision tree algorithm that is popular in marketing, uses an attribute selection measure that is based on the statistical 2 test for independence. Other measures include C-SEP (which per-forms better than information gain and the Gini index in certain cases) and G-statistic (an information theoretic measure that is a close approximation to 2 distribution).

Attribute selection measures based on the Minimum Description Length (MDL) principle have the least bias toward multivalued attributes. MDL-based measures use encoding techniques to define the "best" decision tree as the one that requires the fewest number of bits to both (1) encode the tree and (2) encode the exceptions to the tree(i.e., cases that are not correctly classified by the tree). Its main idea is that the simplest of solutions is preferred.

Other attribute selection measures consider multivariate splits (i.e., where the partitioning of tuples is based on a combination of attributes, rather than on a single attribute). The CART system, for example, can find multivariate splits based on a linear combination of attributes. Multivariate splits are a form of attribute (or feature) construction, where new attributes are created based on the existing ones.

"Which attribute selection measure is the best?" All measures have some bias. It has been shown that the time complexity of decision tree induction generally increases exponentially with tree height. Hence, measures that tend to produce shallower trees may be preferred. However, some studies have found that shallow trees tend to have a large number of leaves and higher error rates. Despite several comparative studies, no one attribute selection measure has been found to be significantly superior to others. Most measures give quite good results.

10.6 DRAWBACKS OF USING DECISION TREES

Till now we have talked about various benefits of Decision Trees. But there are a few drawbacks or precautions which we should be aware of before going ahead with Decision trees. These are:

- Decision trees are susceptible to change in data; Even a small change in data can result into a completely new tree structure
- Decision trees tend to overfit but this can be overcome by pruning the trees
- We might face a problem when we are trying to do an out of sample testing or prediction.

CHECK YOUR PROGRESS

Q.6: A.......................is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

a) Decision tree b) Graphs

c) Trees d) Neural Networks

Q.7: Which data mining technique is most suitable for classification?

a) Clustering b) Association rule mining

c) Decision Tree d) None of the above

Q.8: ID3 is a ....

a) Clustering algorithm

b) Association rule mining algorithm

c) Decision Tree algorithm

d) Functional Algorithm

Q.9: ID3 uses ....

a) Gini Index b) Information Gain

c) Gain Ratio d) G-statistic

Q.10: Which of the following is not a decision tree building algorithm?

a) CART b) CHAID c) C4.5 d) K-Means

10.7 LET US SUM UP

- Classification is a data analysis technique used to classify large amount of data into different classes.
- This technique is used to predict category of data and assign the class labels.
- Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis etc.
- Regression analysis is a statistical methodology that is most often used for numeric prediction; hence, classification and regression are used synonymously.
- Data classification is a two-step process, learning step and a classification step.
- In the learning step the classification model is constructed and in the classification step predicts class labels for given data.
- In the first step, a classifier is built using a predetermined set of data classes or concepts also known as training step.
- Since the class label of each training tuple is provided in the training dataset so this step is also known as supervised learning.

- In the second step the mappings of first step are used to predict the associated class label of a new tuple of which class label is unknown.
- The accuracy of a classifier on a given test set is the ratio of number of data that are correctly classified and the total number of data in the test set.
- A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; edges represent the answers to the question; and the leaves represent the actual output or class label.
- Decision trees classify the problem dataset by sorting them down the tree from the root to some leaf node, with the leaf node provide the class of the input data.
- There are multiple algorithms available for building a decision tree- ID3, C4.5, CART, CHAID, MARS etc.
- An attribute selection measure is a heuristic for selecting the splitting criterion of the decision tree.
- Three popular attribute selection measures are– *information gain*, *gain ratio*, and *Gini index*.
- Many other attribute selection measures also present like CHAID, C-SEP, G-statistic, MDL etc.
- Disadvantage of Decision Tree includes:
  - ► Decision trees are susceptible to change in data; Even a small change in data can result into a completely new tree structure
  - ► Decision trees tend to overfit but this can be overcome by pruning the trees
  - ► We might face a problem when we are trying to do an out of sample testing or prediction.

10.8 FURTHER READING

1) Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.

2) Han, J., Pei, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

3) Tan, P. N. (2018). *Introduction to Data Mining*. Pearson Education India.

10.9 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1:	(a)	**Ans. to Q. No. 2:**	(c)
Ans. to Q. No. 3:	(a)	**Ans. to Q. No. 4:**	(b)
Ans. to Q. No. 5:	(c)	**Ans. to Q. No. 6:**	(a)
Ans. to Q. No. 7:	(c)	**Ans. to Q. No. 8:**	(c)
Ans. to Q. No. 9:	(b)	**Ans. to Q. No. 10:**	(d)

10.10 MODEL QUESTIONS

Q.1: What is Classification? Explain briefly.

Q.2: What is Decision Tree? List all the different decision tree building algorithms.

Q.3: What is information selection measure? Explain gini index?

Q.4: What is entropy and information gain? What is the use of it? Explain briefly.

Q.5: What are the disadvantages of using decision tree? Explain.

*** ***** ***

UNIT 11: PREDICTIONS

UNIT STRUCTURE

11.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define prediction
- describe prediction methods
- define Bayes' Theorem
- describe Bayesian Classification and Bayesian naïve classification
- describe Bayesian Belief Networks
- describe instance based method like k-nearest neighbor method.

11.2 INTRODUCTION

In the previous unit, we have discussed topics like classification, decision tree and attribute selection measures. In this unit, we go through the definition of prediction, different methods for prediction and where we use prediction. We will go through the Bayes' Theorem, Bayesian Classification, Bayesian naïve classification, Bayesian Belief Networks and what mechanism is in it. We will also discuss Lazy learner method specially k-nearest neighbor method for classification. In the next unit, we will explore the concept of classification in detail.

Prediction in data mining is to identify data points purely on the description of another related data value. We get the different prediction techniques like neural network, Bayesian classifier, decision tree, nearest neighbor, support vector machine, multiple linear regression etc. Bayesian classification is a statistical classifier. They can predict class membership probabilities. Bayesian classification based on Baye's theorem which gives the two types of probability: Posterior Probability [P(H|X)] and Prior Probability [P(H)].

Bayesian belief networks is probabilistic graphical models, which allow the representation of dependencies among subsets of attributes. Bayesian belief networks can be used for classification from which we get directed acyclic graph and conditional probability table.

The *k*-nearest neighbors algorithm (*k*-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification.

11.3 PREDICTION TECHNIQUES

Prediction in data mining is to identify data points purely on the description of another related data value. It is not necessarily related to future events but the used variables are unknown. Prediction derives the relationship between a thing you know and a thing you need to predict for future reference.

With the help of the following example we can explain Prediction. Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example, we are bothered to predict a numeric value, so that upcoming sale amount can be planned accordingly. Therefore, the data analysis task is an example of numeric prediction. In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or an ordered value.

Prediction Techniques:

1) **Neural Network:** Artificial neural networks are based on the operation and structure of the human brain. A main feature of neural

networks is an iterative learning process in which data samples are presented to the network one at a time, and the weights are adjusted to predict the correct class label. Advantages of neural networks include their high tolerance to noisy data as well as their ability to classify patterns on which they have not been trained.

2) **Bayesian Classifier:** It is a statistical classification approach based on the Bayes theorem to calculate probability of A given B, P (B given A) =P(A and B)/P (A) the algorithm counts the number of cases where A and B occurs simultaneously and divides it by the number of cases where A alone occurs. Let X be a data tuple, X is considered "Evidence", in Bayesian terms. Let H be some hypothesis, such that the data tuple X belongs to class C. P (H|X) is posterior probability, of H conditioned on X. P (H) is the prior probability of H in contract.

3) **Decision Tree:** Decision tree uses the simple divide-and conquer algorithm. In these tree structures, leaves represent classes and branches signify conjunctions of features that lead to those classes. The attribute that most effectively splits samples into different classes is chosen, at each node of the tree. A path to a leaf from the root is found depending on the assessment of the predicate at each node that is visited, to predict the class label of an input. Decision tree is a fast and easy method.

4) **Support Vector Machine:** Normally SVM is the classification technique. Initially it is developed for binary type classification but later extended to multiple classifications. This SVM creates the hyper plane on the original inputs for effective separation of data points.

5) **Multiple Linear Regression:** This method is performed on a dataset to predict the response variable based on a predictor variable or is used to study the relationship between a response and predictor variable, for example, student test scores compared to demographic information such as income, education of parents, etc.

6) **k-Nearest Neighbors:** Like the classification method with the same name above, this prediction method divides a training dataset into groups of k observations using a Euclidean Distance measure

to determine similarity between “neighbors”. These groups are used to predict the value of the response for each member of the validation set.

11.4 BAYESIAN (STATISTICAL) CLASSIFICATION

Bayesian classification is a statistical classifier. They can predict class membership probabilities. Bayesian classification is based on Bayes’ theorem. Bayes’ Theorem is named after Thomas Bayes, a nonconformist English clergyman who worked on probability and decision theory in the 18th century. There are two types of probabilities–

- Posterior Probability [P(H|X)]
- Prior Probability [P(H)]

where X as a data tuple belongs to the specified class C and H is some hypothesis.

According to Bayes’ Theorem, P(H|X)= P(X|H)P(H) / P(X)

Naïve Bayesian Classification: There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle “all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable”. The naïve classifier or simple classifier works as follows:

1) Let D be a training set of tuples and their associated class labels, each tuple is represented by an n-dimensional attribute vector, $X=(x_1, x_2, \ldots, \ldots, ..., x_n)$, depicting measurements made on the tuple from n attributes respectively $A_1, A_2, ..., ..., ..., A_n$.
2) Suppose there are m classes $C_1, C_2, \ldots, \ldots, ..., C_m$. given a tuple X the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. The naïve Bayesian classifier predict that tuple X belongs to the class C_i if and only if,

   $P(C_i|X) > P(C_j|X)$ for 1d � jd � m, j -:t i

   Thus, we maximize $P(C_i|X)$. the class C_i for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis. By Bayes’ theorem, $P(C_i|X) = P(X|C_i)P(C_i) / P(X)$

3) As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ needs to be maximized. If the class prior probabilities are not known, then commonly assumed that all classes are equals. So we maximized $P(X|C_i)$, otherwise maximize $P(X|C_i)P(C_i)$.

4) If the dataset consists of many attributes, then it is extremely expensive to compute $P(X|C_i)$. To reduce computation, assume class conditional independence which explain that attributes values are conditionally independent of one another.

$$P(X|C_i)\,P(C_i) = \overset{n}{\underset{k=1}{IT}}\; P(x_k|C_i)\,P(C_i)$$

$$= P(x_1|C_i)x\; P(x_2|C_i)x \;\ldots,\; \ldots,\; \ldots,\; x\; P(x_n|C_i)$$

Here x_k refers to the value of attribute A_k for tuple X. To compute $P(X|C_i)$ we consider the following:

a) If A_k is categorical, then $P(x_k|C_i)$ is the number of tuples of class C_i in D having the value x_k for A_k, divided by $|C_{i,D}|$, the number of tuples of class C_i in D.

b) If A_k is continuous value, then, $P(\, x_k|C_i)=g(x_k,\mu C_i,\, \sigma C_i)$

Considering that continuous value attribute have a Gaussian distribution with a mean μ and standard deviation σ, defined by,

$$g(x, \mu, cr) = \frac{1}{\sqrt{2ncr}}\, e^{-(\,x-\mu)^2\ /\ 2cr^2}$$

5) To predict the class label of X, $P(X|C_i)P(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of tuple X is the class C_i if and only if,

$P(X|C_i)P(C_i) > P(X|C_j)P(C_j)$ for 1d � jd � m, j -:t i

CHECK YOUR PROGRESS

Q.1: What is the use of prediction?

....

....

Q.2: Name the different techniques of prediction.

....

....

....

Q.3: What is the main feature of neural networks?

....

....

....

Q.4: What is the advantage of neural network?

....

....

....

Q.5: What is support vector machine?

....

....

11.5 BAYESIAN NETWORKS

Bayesian belief networks is probabilistic graphical models, which unlike naïve Bayesian classifiers allow the representation of dependencies among subsets of attributes. Bayesian belief networks can be used for classification.

Concepts and Mechanisms: The naïve Bayesian classifier makes the assumption of class conditional independence, that is, given the class label of a tuple, the values of the attributes are assumed to be conditionally independent of one another. This simplifies computation. When the assumption holds true, then the naïve Bayesian classifier is the most accurate in comparison with all other classifiers. In practice, however, dependencies can exist between variables.Bayesian belief networks specify joint conditional probability distributions. They allow class conditional independencies to be defined between subsets of variables. They provide a graphical model of causal relationships, on which learning can be performed. Trained Bayesian belief networks can be used for classification. Bayesian belief networks are also known as belief networks, Bayesian networks, and probabilistic networks. For brevity, we will refer to them as belief networks. A belief network is defined by two components–

1) a directed acyclic graph and
2) a set of conditional probability tables

Directed Acyclic Graph: Each node in the directed acyclic graph represents a random variable. The variables may be discrete or continuous-valued. They may correspond to actual attributes given in the data or to "hidden variables" believed to form a relationship (e.g., in the case of medical data, a hidden variable may indicate a syndrome, representing a number of symptoms that, together, characterize a specific disease). Each arc represents a probabilistic dependence. If an arc is drawn from a node Y to a node Z, then Y is a parent or immediate predecessor of Z, and Z is a descendant of Y. Each variable is conditionally independent of its non-descendants in the graph, given its parents.

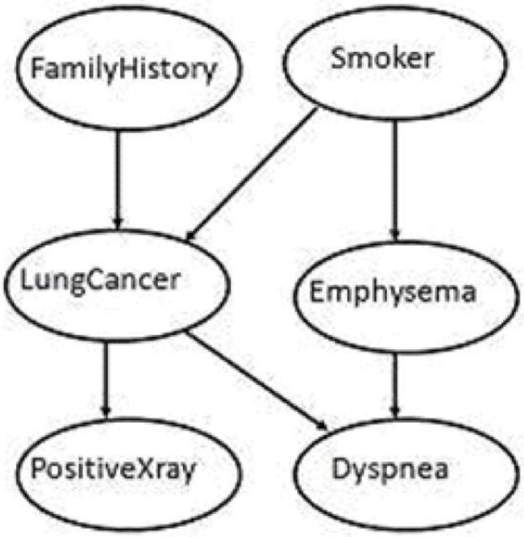
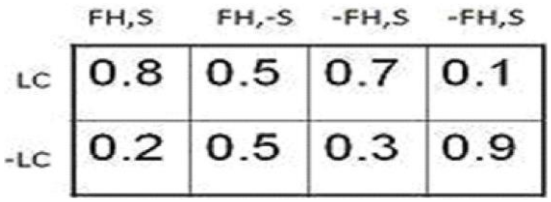


Figure 11.1: Example of Directed Acyclic Graph

For example, having lung cancer is influenced by a person's family history of lung cancer and whether the person is smoker or not. Variable PositiveXRay is independent of whether the patient has a family history of lung cancer or is a smoker, given that we know the patient has lung cancer. In other words, once we know the outcome of the variable LungCancer, then the variables FamilyHistory and Smoker do not provide any additional information regarding PositiveXRay. The arcs also show that the variable

LungCancer is conditionally independent of Emphysema, given its parents, FamilyHistory and Smoker.

Conditional Probability Table: The conditional probability table for the values of the variable LungCancer (LC) showing each possible combination of the values of its parent nodes, FamilyHistory (FH), and Smoker (S) is as follows:



	FH,S	FH,-S	-FH,S	-FH,S
LC	0.8	0.5	0.7	0.1
-LC	0.2	0.5	0.3	0.9

A belief network has one conditional probability table (CPT) for each variable. The CPT for a variable Y specifies the conditional distribution P(Y|Parents(Y)), where Parents(Y) are the parents of Y. The conditional probability for each known value of LungCancer is given for each possible combination of the values of its parents. For instance, from the upper leftmost and bottom rightmost entries, respectively, we see that

P(LungCancer = yes|FamilyHistory = yes, Smoker = yes) = 0.8

P(LungCancer = no |FamilyHistory = no, Smoker = no) = 0.9.

Training Bayesian Belief Networks: In the learning or training of a belief network, a number of scenarios are possible. The network topology (or “layout” of nodes and arcs) may be constructed by human experts or inferred from the data. The network variables may be observable or hidden in all or some of the training tuples. The hidden data case is also referred to as missing values or incomplete data. Several algorithms exist for learning the network topology from the training data given observable variables.

Let D be a training set of data tuples, X1, X2, ..., ..., ..., X|D| . Training the belief network means that we must learn the values of the CPT entries. Let w_{ijk} be a CPT entry for the variable $Y_i = y_{ij}$ having the parents $U_i = u_{ik}$, where $w_{ijk} = P(Y_i = y_{ij}|U_i = u_{ik})$.

For example, if w_{ijk} is the upper leftmost CPT entry of above Conditional Probability Table, then Y_i is LungCancer; y_{ij} is its value, “yes”; U_i lists the parent nodes of Y_i, namely, {FamilyHistory, Smoker}; and u_{ik} lists the values of the parent nodes, namely, {“yes”, “yes”}. The w_{ijk} are viewed as weights. The weights are initialized to random probability values. A gradient descent strategy performs greedy hill-climbing. At each iteration, the weights are updated and will eventually converge to a local optimum solution.

For our problem, we maximize $P_w(D)$. This can be done by following the gradient of $\text{lnP}_w(S)$, which makes the problem simpler. Given the network topology and initialized w_{ijk}, the algorithm proceeds as follows:

1) **Compute the Gradients:** For each i, j, k, compute:

$$\frac{8\text{lnP}_{w}(D)}{8W_{ijk}} = \overset{D}{\underset{d=1}{L}} \frac{P(Y_i = y_{ij}, U_i = u_{ik} \mid X_d)}{W_{ijk}} \qquad (11.1)$$

The probability on the right side of Eq. (11.1) is to be calculated for each training tuple, X_d, in D. For brevity, let's refer to this probability simply as p. When the variables represented by Y_i and U_i are hidden for some X_d, then the corresponding probability p can be computed from the observed variables of the tuple using standard algorithms for Bayesian network inference.

2) **Take a small step in the direction of the gradient:** The weights are updated by,

$$w_{ijk}\ \text{f--}\ w_{ijk} + (l)8\text{lnP}_w(D)/8w_{ijk} \qquad (11.2)$$

where l is the learning rate representing the step size and $8\text{lnP}_w(D)/8w_{ijk}$ is computed from Eq. (11.1). The learning rate is set to a small constant and helps with convergence.

3) **Renormalize the weights:** Because the weights w_{ijk} are probability values, they must be between 0.0 and 1.0, and P j w_{ijk} must equal 1 for all i, k.

11.6 INSTANCE-BASED METHODS (NEAREST NEIGHBOR)

In Lazy Learners method when given a set of training tuples, we construct a generalization (i.e., classification) model before receiving new tuples to classify. We can think of the learned model as being ready and eager to classify previously unseen tuples. Imagine a contrasting lazy approach, in which the learner instead waits until the last minute before doing any model construction to classify a given test tuple. That is, when given a training tuple, a lazy learner simply stores it and waits until it is given a test tuple. Only when it sees the test tuple does it perform generalization to classify the tuple based on its similarity to the stored training tuples. Lazy learners do less work when a training tuple is presented and more work when making a classification or numeric prediction. Because lazy learners store the training tuples or "instances'', they are also referred to as instance-based learners, even though all learning is essentially based on instances. When making a classification or numeric prediction, lazy learners can be computationally expensive. K-nearest neighbor method is example of lazy learner method.

11.6.1 K-nearest Neighbor Method

The k-nearest-neighbor method was first described in the early 1950s. The method is labor intensive when given large training sets, and did not gain popularity until the 1960s when increased computing power became available. It has since been widely used in the area of pattern recognition. Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. In this way, all the training tuples are stored in an n-dimensional pattern space. When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k "nearest neighbors" of the unknown tuple. "Closeness" is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, $X_1 = (x_{11}, x_{12}, ..., ..., ..., x_{1n})$ and $X_2 = (x_{21}, x_{22}, ..., ..., ..., x_{2n})$, is:

$$\text{dist}(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (X_{1i} - X_{2i})^2} \qquad (11.3)$$

In other words, for each numeric attribute, we take the difference between the corresponding values of that attribute in tuple

X_1 and in tuple X_2, square this difference, and accumulate it. The square root is taken of the total accumulated distance count. Typically, we normalize the values of each attribute before using Eq. (11.3). This helps prevent attributes with initially large ranges (e.g., income) from outweighing attributes with initially smaller ranges (e.g., binary attributes). Min-max normalization, for example, can be used to transform a value v of a numeric attribute A to v' in the range [0, 1] by computing–

$$\bar{v} = \frac{v - min_A}{max_A - min_A} \qquad (11.4)$$

where min_A and max_A are the minimum and maximum values of attribute A.

For k-nearest-neighbor classification, the unknown tuple is assigned the most common class among its k-nearest neighbors. When k = 1, the unknown tuple is assigned the class of the training tuple that is closest to it in pattern space. Nearest-neighbor classifiers can also be used for numeric prediction, that is, to return a real-valued prediction for a given unknown tuple. In this case, the classifier returns the average value of the real-valued labels associated with the k-nearest neighbors of the unknown tuple.

Distance calculation method for nominal attributes: It is a simple method used to compare the corresponding value of the attribute in tuple X_1 with that in tuple X_2. If the two are identical (e.g., tuples X_1 and X_2 both have the color blue), then the difference between the two is taken as 0. If the two are different, then the difference is considered to be 1.

Missing values: In general, if the value of a given attribute A is missing in tuple X_1 and/or in tuple X_2, we assume the maximum possible difference. Suppose that each of the attributes has been mapped to the range [0, 1]. For nominal attributes, we take the difference value to be 1 if either one or both of the corresponding values of A are missing. If A is numeric and missing from both tuples X_1 and X_2, then the difference is also taken to be 1. If only one value is missing and the other is present and normalized, then we can take the difference to be either |1 – v' | or |0 – v' | (i.e., 1 – v' or v'), whichever is greater.

Calculation of number of neighbors(k): This can be determined experimentally. Starting with k = 1, we use a test set to estimate the error rate of the classifier. This process can be repeated each time by incrementing k to allow for one more neighbor. The k value that gives the minimum error rate may be selected. In general, the larger thenumber of training tuples means the larger the value of k. Nearest-neighbor classifiers use distance-based comparisons that intrinsically assign equal weight to each attribute. They therefore can suffer from poor accuracy when given noisy or irrelevant attributes. Nearest-neighbor classifiers can be extremely slow when classifying test tuples. If D is a training database of |D| tuples and k = 1, then O(|D|) comparisons are required to classify a given test tuple.

Other techniques to speed up classification time include the use of partial distance calculations and editing the stored tuples. In the partial distance method, we compute the distance based on a subset of the n attributes. If this distance exceeds a threshold, then further computation for the given stored tuple is halted, and the process moves on to the next stored tuple. The editing method removes training tuples that prove useless. This method is also referred to as pruning or condensing because it reduces the total number of tuples stored.

CHECK YOUR PROGRESS

Q.6: Define Multiple Linear Regression method.

....

....

....

Q.7: Define Bayesian network.

....

....

Q.8: What are the components of Bayesian belief networks?

....

....

Q.9: What are the steps of gradient descent strategy?

....

....

....

Q.10: What is the use of Lazy Learner method?

....

....

....

11.7 LET US SUM UP

- Neural network, Bayesian classification, decision tree, nearest neighbor, support vector machine, multiple linear regressions are different prediction methods.
- Bayes' theorem gives posterior and prior probability.
- Naïve Bayesian classification is based on 'the value of particular feature is independent of the value of any other feature given the class value'.
- Decision tree is based on divide and conquer method.
- Bayesian Network is a probabilistic graphical model.
- Directed Acyclic Graph and Conditional Probability Table can achieved fron Bayesian Network.
- DAG represents probability dependency.
- CTP values show each possible combination.
- Calculation of k in k-nearest neighbor classifier is starts from 1 and process will be repeat with increment of k value.

11.8 FURTHER READING

1) Han, J., Pei, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

2) Pujari, A. K. (2001). *Data Mining Techniques*. Universities Press.

11.9 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: Prediction in data mining is to identify data points purely on the description of another related data value.

Ans. to Q. No. 2: Different prediction techniques are neural network, Bayesian classifier, decision tree, nearest neighbor, support vector machine, multiple linear regressions.

Ans. to Q. No. 3: Main feature of neural networks is an iterative learning process in which data samples are presented to the network one at a time, and the weights are adjusted to predict the correct class label.

Ans. to Q. No. 4: Advantages of neural networks include their high tolerance to noisy data, and ability to classify patterns on which they have not been trained.

Ans. to Q. No. 5: Support Vector Machine technique is developed for binary type classification later extended to multiple classifications.

Ans. to Q. No. 6: Multiple Linear Regression method is performed on a dataset to predict the response variable based on a predictor variable or used to study the relationship between a response and predictor variable.

Ans. to Q. No. 7: Bayesian belief networks is probabilistic graphical models, which allow the representation of dependencies among subsets of attributes.

Ans. to Q. No. 8: A belief network is defined by two components– a directed acyclic graph and a set of conditional probability tables.

Ans. to Q. No. 9: Gradient descent strategy are consists of three steps: calculate the gradient, take a small step in the direction of the gradient and renormalized the weight.

Ans. to Q. No. 10: Lazy Learners when given a set of training tuples, will construct a generalization model before receiving new uples to classify.

11.9 MODEL QUESTIONS

Q.1: What is prediction?

Q.2: Write about the different techniques of prediction.

Q.3: Explain the Bayes' Theorem.

Q.4: What is the Naïve Bayesian Classification?

Q.5: What is mechanism used in Bayesian Networks?

Q.6: Explain directed acyclic graph with example.

Q.7: What is conditional probability table? Explain with example.

Q.8: What is training process related to Bayesian Networks?

Q.9: Explain the k-nearest neighbor classification method for prediction.

*** ***** ***

UNIT 12: EVALUATION

UNIT STRUCTURE

12.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define training data and testing data
- describe how to evaluate accuracy of a classifier
- define cross validation
- describe combination of multiple models.

12.2 INTRODUCTION

In the previous unit, we have learned about prediction, different methods for prediction and where we can use prediction. We have also learned in detail about Bayes' Theorem, Bayesian Classification, Bayesian naïve classification, Bayesian Belief Networks along with k-nearest neighbor method for classification. In this unit we will learn what is test data and training data and how they differ from each other. We will also learn how to calculate the accuracy and improve the accuracy of a classifier and how it is effective for future prediction. We will learn topics like cross validation and how we can perform statistical analysis for independent dataset. In the next unit, we will explore the concept of clustering.

When we have to build a classification model, some questions arise like what was the previous data and on the basis of what can we build the classification model. After building the classification model, we have to predict if it will be effective or not for the future business. On basis of that we predict whether the particular classifier is accurate or not. Classification Accuracy means the ratio of number of correct predictions to the total number of input samples. Accuracy can be measured with the help of confusion matrix, area under curve, F1 score etc.

Cross-validation is a technique that is used for the assessment of how the results of statistical analysis generalize to an independent data set. Cross-validation is largely used in settings where the target is prediction and it is necessary to estimate the accuracy of the performance of a predictive model. The prime reason for the use of cross-validation rather than conventional validation is that there is not enough data available for partitioning them into separate training and test sets.

To improve the accuracy of the classification, ensemble methods are introduced. The ensemble for classification is a composite model of classifier. The individual classifiers vote and a class label prediction is returned by the ensemble based on collection of votes. Ensemble tends to be more accurate than their component classifiers.

12.3 TRAINING AND TESTING

When we have to build a classification model, many questions are going through our mind. For example, suppose you used previous sales data to build a classifier to predict the purchase behavior of a customer. Classifier can predict the future customer purchasing behavior, that is, future customer data on which the classifier has not been trained. Even we can build more than one classifier using different methods and wish to compare their accuracy to predict the best classifier.

A training dataset is a dataset of examples used for learning that is to fit the parameters (e.g., weights) of a classifier. Most approaches that search through training data for empirical relationship stend to overfit the data, meaning that they can identify and exploit apparent relationships in the training data that do not hold in general.

A test dataset is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset. A model that fits to the training dataset also fits the test dataset well, minimal over-fitting has taken place. A better fitting of the training dataset as opposed to the test dataset usually points to over-fitting. A test set is therefore a set of examples used only to assess the performance of a fully specified classifier.

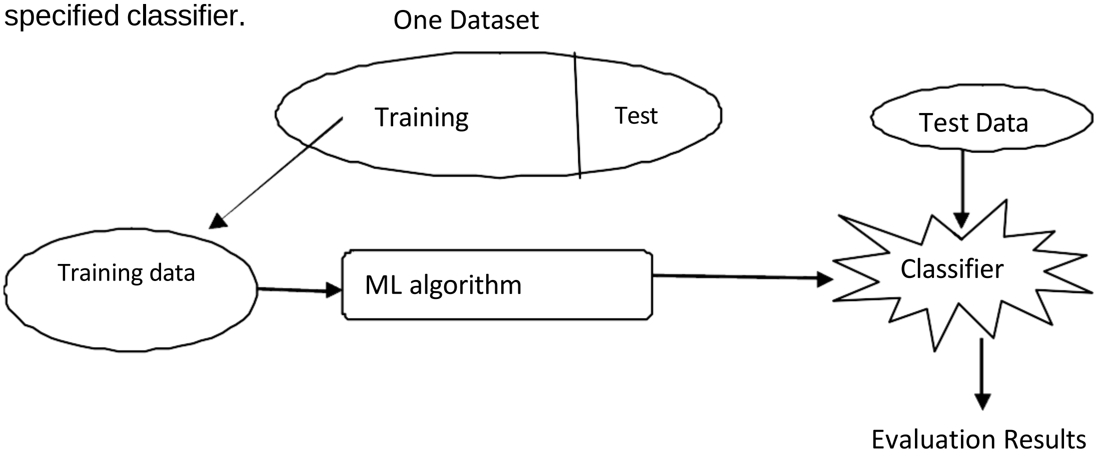


Figure 12.1: Example of Classifier

In data mining, the data is divided into training set (most of data) and testing set (smaller portion) after the model processed by using the training set. We test the model by making prediction against the test set based on the value that determined for training set. If your division sets are n numbers this process is repeated for n times; for example if your data divided into four training set and one testing set, the process repeated for five times alternatively. Now it's easy to know whether your model guesses are correct or not.

12.4 EVALUATIING ACCURACY OF A CLASSIFIER

Classification accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

$$\textbf{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions Made}}$$

It works well only if there are equal number of samples belonging to each class.

For example, consider that there are 98% samples of class A and 2% samples of class B in our training set. Then our model can easily get **98% training accuracy** by simply predicting every training sample belonging to class A.

When the same model is tested on a test set with 60% samples of class A and 40% samples of class B, then the **test accuracy would drop down to 60%**. Classification accuracy is great, but it gives us the false sense of achieving high accuracy.

The real problem arises, when the cost of misclassification of the minor class samples are very high. If we deal with a rare but fatal disease, the cost of failing to diagnose the disease of a sick person is much higher than the cost of sending a healthy person to more tests.

Confusion Matrix: Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model.

Let us assume we have a binary classification problem. We have some samples belonging to two classes: YES or NO. Also, we have our own classifier which predicts a class for a given input sample. On testing our model on 165 samples, we get the following result.

n = 165	**Predicted: NO**	**Predicted: YES**
Actual: NO	**50**	**10**
Actual: YES	**5**	**100**

Figure 12.2: Confusion Matrix

There are 4 important terms:

- **True Positives:** The cases in which we predicted YES and the actual output was also YES.
- **True Negatives:** The cases in which we predicted NO and the actual output was NO.
- **False Positives:** The cases in which we predicted YES and the actual output was NO.

- **False Negatives:** The cases in which we predicted NO and the actual output was YES.

Accuracy for the matrix can be calculated by taking average of the values lying across the **"main diagonal"** i.e.,

$$\textbf{Accuracy} = \frac{\textbf{True Positives + False Negatives}}{\textbf{Total Number of Samples}}$$

$$\therefore \textbf{Accuracy} = \frac{\textbf{100 + 50}}{\textbf{165}} = \textbf{0.91}$$

Confusion Matrix forms the basis for the other types of metrics.

Area Under Curve: Area Under Curve(AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. Before defining AUC, let us understand two basic terms:

- **True Positive Rate (Sensitivity):** True Positive Rate is defined as *TP/ (FN*+TP). True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

$$\textbf{True Positive Rate} = \frac{\textbf{True Positive}}{\textbf{(False Negative + True Positive)}}$$

- **False Positive Rate (Specificity):** False Positive Rate is defined as *FP / (*FP+TN). False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

$$\textbf{False Positive Rate} = \frac{\textbf{False Positive}}{\textbf{(False Positive + True Negative)}}$$

False Positive Rate and True Positive Rate both have values in the range [0, 1]. FPR and TPR both are computed at threshold values such as (0.00, 0.02, 0.04, …, ..., ..., 1.00) and a graph is drawn. AUC is the area under the curve of plot False Positive Rate vs True Positive Rate at different points in [0, 1].

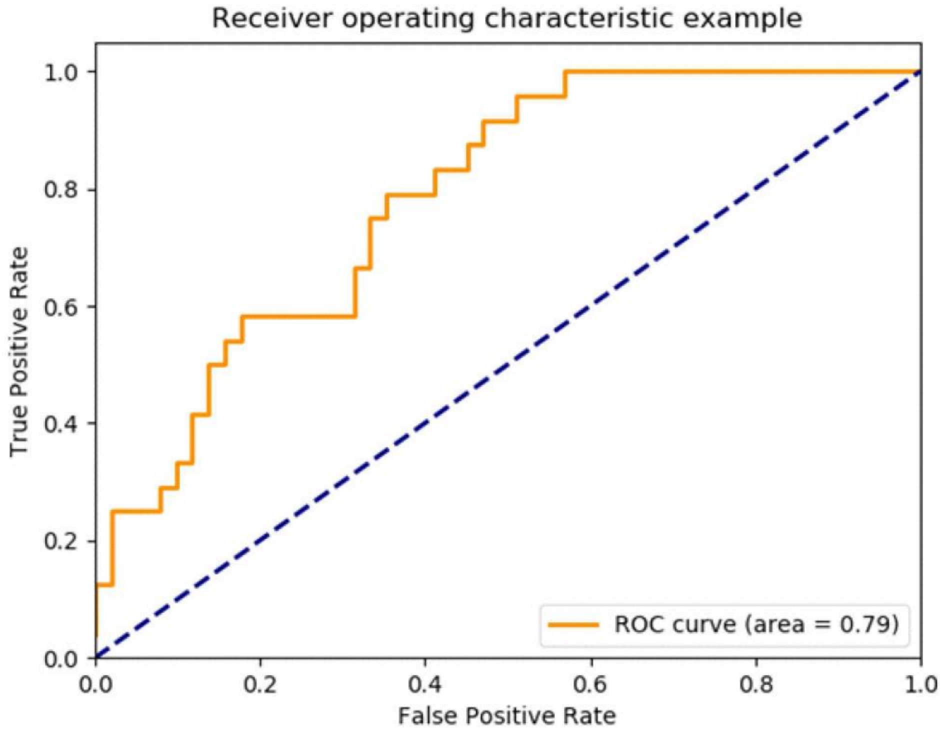


Figure 12.3: Graph of False Positive Rate versus True Positive Rate

As evident, AUC has a range of [0, 1]. The greater the value, the better is the performance of our model.

F1 Score: *F1 Score is used to measure a test's accuracy.* F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as:

$$F1 = 2 * \frac{1}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

F1 Score tries to find the balance between precision and recall.

- **Precision:** It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$\textbf{Precision} = \frac{\textbf{True Positive}}{\textbf{(False Negatives + True Positives)}}$$

- **Recall:** It is the number of correct positive results divided by the number of ***all*** relevant samples (all samples that should have been identified as positive).

$$\textbf{Recall} = \frac{\textbf{True Positive}}{\textbf{(True Positives + False Negatives)}}$$

Classifiers can be also compared with respect to the following aspects:

- **Speed:** This refers to the computational costs involved in generating and using the given classifier.
- **Robustness:** This is the ability of the classifier to make correct predictions given noisy data or data with missing values. Robustness is typically assessed with a series of synthesis data sets representing increasing degrees of noise of missing values.
- **Scalability:** This refers to the ability to construct the classifier efficiently given large amounts of data.
- **Interpretability:** This refers to the level of understanding and insight that is provided by the classifier or predictor. Interpretability is subjective and more difficult to assess.

CHECK YOUR PROGRESS

Q.1: What is test dataset?

....

....

....

Q.2: What is precision?

....

....

....

Q.3: What is confusion matrix?

....

....

Q.4: What is sensitivity and specificity?

....

....

....

....

....

....

....

....

Q.5: How can we compare the classifier?

....

....

12.5 CROSS VALIDATION

Cross-validation is a technique that is used for the assessment of how the results of statistical analysis generalize to an independent data set. Cross-validation is largely used in settings where the target is prediction and it is necessary to estimate the accuracy of the performance of a predictive model. The prime reason for the use of cross-validation rather than conventional validation is that there is not enough data available for partitioning them into separate training and test sets (as in conventional validation). This results in a loss of testing and modeling capability.

For a prediction problem, a model is generally provided with a data set of known data, called the training data set, and a set of unknown data against which the model is tested, known as the test data set. The target is to have a data set for testing the model in the training phase and then provide insight on how the specific model adapts to an independent data set. A round of cross-validation comprises the partitioning of data into complementary subsets, then performing analysis on one subset. After this, the analysis is validated on other subsets (testing sets). To reduce variability, many rounds of cross-validation are performed using many different partitions and then an average of the results is taken. Cross-validation is a powerful technique in the estimation of model performance technique.

K-fold Cross Validation: K-Fold Cross Validation is a common type of cross validation that is widely used in machine learning. K-fold cross validation is performed as per the following steps:

Step 1: Partition the original training data set into k equal subsets. Each subset is called a **fold**. Let the folds be named as f_1, f_2, …, ..., f_k.

Step 2: For i = 1 to i = k

- Keep the fold f_i as validation set and keep all the remaining *k–1* folds in the cross validation training set.
- Train your machine learning model using the cross validation training set and calculate the accuracy of your model by validating the predicted results against the validation set.

Step 3: Estimate the accuracy of your machine learning model by averaging the accuracies derived in all the *k* cases of cross validation.

In the k-fold cross validation method, all the entries in the original training data set are used for both training as well as validation. Also, each entry is used for validation just once.

Generally, the value of *k* is taken to be 10, but it is not a strict rule, and *k* can take any value.

Leave-one-out is special case k-fold cross validation where k is set to the number of initial tuples. That is only one sample is left out at the time of testing. In stratified cross validation, the folds are stratified so that class distribution of the tuples in each fold is approximately same as that of the initial data.

12.6 COMBAINING MULTIPLE MODELS

To improve the accuracy of the classification, ensemble methods are introduced. The ensemble for classification is a composite model of classifier. The individual classifiers vote and a class label prediction is returned by the ensemble based on collection of votes. Ensemble based results tend to be more accurate than their component classifiers.

Bootstrap: The bootstrap method is a re-sampling technique used to estimate statistics on a population by sampling a dataset with replacement.

T*he bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.*

It can be used to estimate summary statistics such as the mean or standard deviation. It is used in applied machine learning to estimate the skill of machine learning models when making predictions on data not included in the training data.

Importantly, samples are constructed by drawing observations from a large data sample one at a time and returning them to the data sample after they have been chosen. This allows a given observation to be included in a given small sample more than once. This approach to sampling is called sampling with replacement.

The process for building one sample can be summarized as follows:

Step 1: Choose the size of the sample.

Step 2: While the size of the sample is less than the chosen size

- Randomly select an observation from the dataset
- Add it to the sample

Example 12.1: The bootstrap method can be used to estimate the quantity of a population. This is done by repeatedly taking small samples, calculating the statistics and taking the average of the calculated statistics. We can summarize this procedure as follows:

Step 1: Choose a number of bootstrap samples to perform

Step 2: Choose a sample size

Step 3: For each bootstrap sample

- Draw a sample with replacement with the chosen size
- Calculate the statistic on the sample

Step 4: Calculate the mean of the calculated sample statistics.

The procedure can also be used to estimate the skill of a machine learning model.

Boosting: Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms that convert weak learners to strong ones. Boosting is based on the question posed by Kearns and Valiant (1988, 1989). “Can a set of **weak learners** create a single **strong learner**?” A weak learner is defined to be a classifier that is only slightly correlated with the true classification. In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the true classification.

Bagging: Bootstrap Aggregation also known as bagging, is a powerful and simple ensemble method to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. Bagging is a special case of the model averaging approach.

An ensemble method is a technique that combines the predictions from many machine learning algorithms together to make more reliable and accurate predictions than any individual model. It means that we can say that prediction of bagging is very strong. The main purpose of using the bagging technique is to improve Classification Accuracy.

Description of the Bagging Technique: Given a standard training set of size *n*, bagging generates *m* new training sets, each of size n', by sampling from *D* uniformly and with replacement. By sampling with replacement, some observations may be repeated in each. If $n' = n$, then for large *n* the set is expected to have the fraction (1 – 1/*e*) (H�63.2%) of the unique examples of *D*, the rest being duplicates. This kind of sample is known as a bootstrap sample. The *m* models are fitted using the above *m* bootstrap samples and combined by averaging the output (for regression) or voting (for classification).

For example, we have 1000 observations and 200 elements. In bagging, we will create several models with a subset of variables and a subset of observations. i.e., we might create 300 trees with 300 random variables and 20 observations in each tree. After that, we can average the results of all the 300 tree’s (models) to get to our final prediction.

CHECK YOUR PROGRES

Q.6: What is the standard value for k in k fold cross validation method?

....

....

Q.7: What is fold?

....

....

Q.8: What is leave-one-out?

....

....

Q.9: What is bootstrap?

....

....

....

....

12.7 LET US SUM UP

- Dataset can be divided as test dataset and training dataset.
- A test dataset is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset.
- Confusion matrix describes the complete performance of model.
- Confusion Matrix is defined on the basis of true positive, true negative, false positive, false negative.
- True positive rate (Sensitivity) is defined as TP/(FN+TP).
- False positive rate (Specificity) is defined as FP/(FP+TN).
- F1 score measures the test’s accuracy and it tries to find balance between recall and precision.
- Cross validation is used where target is prediction.

- To improve the accuracy of the classification, ensemble methods are introduced.

12.8 FURTHER READING

1) Han, J., Pei, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
2) Pujari, A. K. (2001). *Data Mining Techniques*. Universities Press.

12.9 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: A test dataset is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset

Ans. to Q. No. 2: It is the number of correct positive results divided by the number of positive results predicted by the classifier.

Precision = True Positives/ (True Positives + False Positives)

Ans. to Q. No. 3: Confuse matrix gives a matrix as output and describes the complete performance of the model.

Ans. to Q. No. 4: True Positive Rate is defined as *TP/ (FN*+TP). True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

True Positive Rate = True Positive / (False Negative + True Positive)

False Positive Rate is defined as *FP / (*FP+TN). False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

False Positive Rate = False Positive / (False Positive + True Negative)

Ans. to Q. No. 5: Classifiers are compared with respect to the following aspects: Speed, robustness, scalability and interpretability.

Ans. to Q. No. 6: The value of *k* is taken to be 10 in K fold cross validation method, but it is not a strict rule, and *k* can take any value.

Ans. to Q. No. 7: Partition the original training data set into k equal subsets. Each subset is called a **fold.**

Ans. to Q. No. 8: Leave-one-out is special case k-fold cross validation where k is set to the number of initial tuples.

Ans. to Q. No. 9: The bootstrap method is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement. The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.

12.10 MODEL QUESTIONS

Q.1: Differentiate between test dataset and training dataset.

Q.2: How can we calculate the accuracy of a classifier?

Q.3: What are the factors effected in accuracy of a classifier?

Q.4: Explain the k fold cross validation method.

Q.5: What is cross validation?How is cross validation effective for dataset?

Q.6: Explain the combined model of classifier.

Q.7: What is bootstrap?

Q.8: What is boosting and bagging?

Q.9: Differentiate between boosting and bagging.

*** ***** ***

UNIT 13: CLUSTERING

UNIT STRUCTURE

13.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define clustering
- describe various classification of clustering algorithm
- describe partitioning and hierarchical methods
- describe how K-means algorithm works
- describe DBSCAN algorithm.

13.2 INTRODUCTION

In the previous unit, we have learnt about training data and testing data. We have also learned about cross validation and different evaluation measures. In this unit, we will learn about different clustering techniques. We will also learn the various classification of clustering algorithms such as partition algorithm, hierarchical algorithm, divisive algorithm etc. We will also learn two different clustering algorithms K-means and DBSCAN algorithm in detail in this unit. In the next unit, we will explore the topic of web data and web mining.

13.3 CLUSTERING

Clustering is a usefull technique for the discovery of data distribution and pattern of given data. The main aim of clustering technique is to find both distributed and dense region in a data set. Clustering technique includes several classes of clustering algorithms. These algorithm attempts to automatically partition the data space into several region i.e to sparse or dense region. The main goal of clustering is to find all sets of similar examples in the data set. Following are the different objectives of clustering techniques:

- To uncover natural grouping.
- To initiate hypothesis about the data.
- To find consistent and valid organization of data.

Clustering has several applications. Following are some of them:

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers to discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- Clustering also helps in the identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

13.4 CLASSIFICATION OF VARIOUS CLUSTERING ALGORITHMS

Basically there are two main approaches to clustering the data set. Those two approaches are *partitioning approach* and *hierarchical approach*.

Partitioning method partitions the data set into a predefined set of clusters. Suppose, we have a dataset with **n** objects then partitioning method will partition the dataset to **k** number of clusters. When **k<=n,** it means each cluster will have at least one object or more than one and each object will exactly belong to one group. For a given number of partitions (say k), the partitioning method will create an initial partitioning. Then it uses the iterative re-clustering technique to improve the partitioning by moving objects from one group to other.

Hierarchical method creates a hierarchy of cluster from the data set. **Hierarchical clustering are of two types: *Agglomerative Clustering and Divisive Clustering.*** Agglomerative clustering is a bottom up approach. Divisive clustering technique is a top down approach.

13.4.1 Partitioning Methods

Partitioning clustering algorithm construct partition of a database with N objects to K cluster. The partitioning clustering technique adopts iterative optimization technique. Here, the database is iteratively breaking down to K cluster. Following K-means is a partitioning clustering algorithm.

K-means: K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and to associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate ‘k’ new centroids as BabyCenter of the clusters resulting from the previous step. After we have these ‘k’ new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing objective functions known as squared error function given by:

$$J(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}(|| x_i - v_j ||)^2$$

where, '$||x_i – v_j||$' is the Euclidean distance between x_i and $v_{j.}$ 'c_i' is the number of data points in i^{th} cluster. '*c*' is the number of cluster centers.

Algorithm 13.1: K-means clustering algorithm

Let X = {x_1, x_2, x_3, …, …, ..., x_n} be the set of data points and V = {v_1, v_2, …, …, ..., v_c} be the set of centers.

Step 1: Randomly select '*c*' cluster centers.

Step 2: Calculate the distance between each data point and cluster centers.

Step 3: Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

Step 4: Recalculate the new cluster center using the formula:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in i^{th} cluster.

Step 5: Recalculate the distance between each data point and new obtained cluster centers.

Step 6: If no data point was reassigned then stop, otherwise repeat from step 3.

CHECK YOUR PROGRESS

Q.1: State whether the following statements are true (T) or false (F)

i) The main aim of clustering technique is to find both distributed and dense region in a data set.

ii) The partitioning clustering technique does not adopt iterative optimization technique.

iii) The main idea of K mean clustering is to define k centers, one for each cluster.

13.4.2 Hierarchical Methods

Hierarchical method creates a hierarchy of clusters from the data set. Hierarchical clusterings are of two types: ***Agglomerative Clustering and Divisive Clustering.*** Agglomerative clustering is a bottom up approach. In this technique, clustering starts with small cluster where each cluster contains single object and keeps on merging the objects or groups that are close to one another. It keeps on doing so until all of the groups are merged into one or until the termination condition holds. Divisive clustering technique is a top down approach. This technique creates k number of clusters from a single cluster. The following DBSCAN algorithm is a hierarchical algorithm.

Density based DB-SCAN: DBSCAN (Density Based Spatial Clustering of Application of Noise) used a density based concept to discover a new set of cluster. The key idea is that for each object of cluster, the neighborhood of a given radius has to contain a minimum threshold data object. The main parameter of DBSCAN is distance function of data object. Following are some parameter of DBSCAN

c-Neighbourhood of an object: For a given non negative value c, the c-Neighbourhood of an object O_i, denoted by $N_c(O_i)$, is defined by $N_c(O_i) = \{O_j \, c \, D \mid d(O_i, O_j)$ � $c\}$

Core Object: An object is said to be core object if size of the c-Neighbourhood of the object is greater than mean point. (user specified minimum density) i.e., $| N_c (O)| \ 2$ MinPts.

Directly Density Reachable: An object O_i is directly density reachable from an object O_j with respect to c and MinPts, if O_j is a core object and O_i is in its c-Neighbourhood.

Density Reachable: An object O_i is density reachable from an object O_j with respect to c and MinPts in D if there is a chain of objects O_1, $O_{2,}$ O_3, ..., ..., ..., $O_{n,}$ such that $O_1 = O_j$ and $O_i = O_{n,}$ such that O_e c D and O_{e+1} is directly density reachable from O_e.

Density Connected: An object O_i is density connected to an object O_j with respect to c and MinPts in D if there is another object O c D such that both O_i and O_j are density reachable from O.

Cluster: A Cluster C with respect to c and MinPts is a non empty subset of D satisfying the following conditions:

- For all O_i, O_j c D, if O_i c C and O_j is density reachable from O_i with respect to c and MinPts, then O_j c C.
- For all O_i, O_j c C, O_i is density connected to O_j, with respect to c and MinPts.

Noise: Let C_1, C_2, ..., ..., ..., C_k are the clusters of D with respect to c and MinPts. Then the set of objects O are said to be noise if they do not belong to any cluster C_i.

Algorithm 13.2: DBSCAN Algorithm

Input: Database of objects D

Algorithm DBSCAN (D, c, MinPts)

```
Step 1:  Do for all D
             If O is unclassified
             Call function mk_cluster(O, D, c, MinPts)
         End do
```

Function mk_cluster(O, D, c, MinPts)

```
Step 1:  Function mk_cluster(O, D, c, MinPts)
Step 2:  Get the c-neighbourhood of O as $N_c(O)$
Step 3:  If | $N_c(O)$| < MinPts
             Mark O as Noise
             return
         else
Step 4:  Select new cluster_id and mark all object of $N_c(O)$ with this
         cluster_id and put them into candidate objects.
```

```
Step 5:   do while candidate-objects is not empty
            Select an object from candidate-object as current_object
            delete current_object from candidate_objects
            retrieve $N_c$(urrent_object)
            if | $N_c$(current_object)| 2 MinPts
                 select all objects in $N_c$(current_object) not yet
                 classified or marked as noise,
                 mark all of the objects with cluster_id,
                 include the unclassified objects into candidate-objects
                 end do
            return
```

The DBSCAN algorithm maintains three categories of objects. Those are classified, unclassified and noise. The classified and noise are the objects whose c-neighbourhood are already calculated. The unclassified objects are those whose c-neighbourhood is not calculated. Each classified object has an associated cluster-id. Noise has also a dummy cluster-id. The unclassified objects do not have any cluster-id. The algorithm gradually converts the unclassified objects to classified objects and/or noise.

CHECK YOUR PROGRESS

Q.2: Fill in the blanks:

i) Agglomerative clustering is a .... approach.

ii) Divisive clustering technique is a .... approach.

iii) An object is said to be core object if size of the .... of the object is greater than mean point.

iv) The set of objects are said to be noise if they do not belong to any ....

13.5 LET US SUM UP

- Clustering is a usefull technique for the discovery of data distribution and pattern of given data
- The main goal of clustering is to find all sets of similar examples in the data set.
- Two approaches of clustering are partitioning approach and hierarchical approach.
- Partitioning clustering algorithm construct partition of a database with N objects to K cluster.
- K-means is a partitioning clustering algorithm.
- Hierarchical method creates a hierarchy of cluster from the data set.
- DBSCAN(Density Based Spatial Clustering of Application of Noise) used a density based concept to discover a new set of cluster.
- The DBSCAN algorithm maintains three categories of objects, those are classified, unclassified and noise.

13.6 FURTHER READING

1) Han, J., Pei, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
2) Pujari, A. K. (2001). *Data Mining Techniques*. Universities Press.

13.7 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) True; ii) False; iii) True

Ans. to Q. No. 2: i) bottom up; ii) top down; iii) c-Neighbourhood; iv) cluster

13.8 MODEL QUESTIONS

Q.1: What is clustering technique?

Q.2: What are the different applications of clustering?

Q.3: Explain the various classifications of clustering.

Q.4: Differentiate between Partitioning approach Hierarchical approach.

Q.5: Differentiate between Agglomerative and Divisive clustering techniques.

Q.6: Explain K mean clustering algorithm.

Q.7: What is DBSCAN? Explain the DBSCAN algorithm.

*** ***** ***

UNIT 14: INTRODUCTION TO WEB MINING

UNIT STRUCTURE

14.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define web mining
- describe the different types of web mining are
- list the uses of web mining
- describe the different methods of web mining

- define text mining
- describe how episode rules are discovered from text.

14.2 INTRODUCTION

In the previous unit, we have learned about the different clustering techniques such as partition algorithm, hierarchical algorithm, divisive algorithm etc. We have also learned about K-means and DBSCAN algorithm in detail. In this unit, we will learn about web mining and its different types. Different categories of web mining along with the applications of web mining are also covered in this unit. Another topic that has been introduced in this unit is text mining. In the next unit, we will explore the concepts of temporal data mining and spatial data mining.

14.3 WEB MINING

Web Mining is the application of data mining techniques to extract useful knowledge from web data like contents of web documents, hyperlinks structure of documents and web usage logs. In web mining, web data refers to the data about web documents, the web structure and web log. The information on web is availed publicly by many companies and persons. Compare to common text documents, the contents of web pages are often complex. The data in the web is very large and it also expands rapidly. Considering all these factors, the WWW has turned into an area of interest for data mining. On WWW, the web documents do not have specific order of arrangement and also it is also a large collection of such documents. According to requirement and interest, these web documents are often updated regularly as the users can have dissimilar background and purpose of usage. It may be possible that only a small portion of information on web is significant and useful to specific user needs. To pursue mining on web all these factors are important and valuable for the discovery and utilization of resources on World Wide Web.

Mining of web data focuses on good design of web sites and formation of techniques to analyse the behaviour of the user. There are also strong requirements of techniques to help in business decision in e-commerce. By understanding the user behaviour, desired advertising can be displayed on web sites to which the user is likely to respond. Design of web sites is important. A well designed web sites can easily fulfil the objective of business by highlighting the products with good profit margin and its also improves the users navigation experience. Web data mining can be defined in two distinct forms: first, it is defined as chain of order tasks and, second, it is defined considering the type of web data used in web data mining process. There are three types of data generally concerned in web data mining: web contents, web structure and web usage log. According to the kinds of data to be mined, Web Data Mining can be broadly divided into three categories: *Web content mining*, *Web structure mining* and *Web usage mining*.

14.3.1 Types of Web Data

The data on World Wide Web are available in three different formats: web content, web structure and web usage.

- **Web Content Data:** Web content data is the data in the web pages;it can be text, image, audio or video. Web content data are used to provide information to the users. In this category, the HTML pages are common and more familiar form of web content data. Because of variety of internal formats and the browsers way of interpretation, the HTML document may appear differently for different browsers while viewing. The basic document structure of HTML is similar. HTML documents are often considered semi-structured as different elements of documents are not designed according to specific schema.

  XML document is another known form of web content data which enables storing and transporting information. It is having structured information and includes contents as well as information about contents. Each XML document has specific structure and allows identifying document structure and adding information.

  Another type of web content data is dynamic server pages which are processed by the web server and the generated result is sent to web browser. In contrast, without any change, the static contents are sent to browser. Some of the familiar server page languages are JSP(Java Server Page), ASP(Active Server Page), PHP(Pre-Hypertext Processor) etc.

- **Web Structure data:** Web structure data represents linkage and relationship of web contents to others. Two types of structure, namely *intra-page* and *inter-page* structure can be considered. In specific web page, information about the arrangement of different HTML tags is intra-page structure information. The pages are connected with other pages using hyper-links. This is inter-page structure information. Hyperlinks of web pages collectively form a graph called web graph and it describes the whole structure of the web site. Web graph is a common way of showing the links of one web page to another in WWW and depicts the overall structure. Web graph is a representation of a specific site describing the structure of links and relationship to the HTML documents in WWW.

  Web document is depicted as a node in graph and edge is HTML link connecting one page with another. In two different ways, the edges of the graph are presented. A hyperlink stopping at related page is presented as outgoing arcs and the hyperlinks using which related page can be found is presented as incoming arcs. Web graph can be used in some applications like web searching, indexing and web communities detection.

- **Web Usage data:** Web usage data involves the log data collected by web server and application server. When the user interacts with web site, web log data is generated on web server in form of web server log files. Application Server Data is common in commercial application servers. These data are used to track various types of business events and logs. Application Level Data is another source for web usage data. With this type of data it is possible to record various kinds of events in an application. These data are used for generating histories about selected special events. The data in this category can be divided into three categories based on the source of its collection: server side data, client side data, and proxy server side data. Other additional data sources are demographic data, site files, cookies etc. Generally, the web server is assigned a domain name and it has IP address. When the user sends request for page to web server through browser, the request is processed by the server and the page is sent to the user. As a result of user interaction with web site and server, data are generated on server and resource request, success, error etc. information are recorded into server log files. Different types of usage log files are created on the server such as access log, error log, referrer log, agent log.

Information stored in web access log includes IP Address, username, date and time, request. Error information like 'file not found', 'no data', 'aborted transmission' etc. is recorded into Error logs. The information about browser, its version and operating system of user making request is stored into Agent logs.

Using web traffic analyser software, the log files of web server can be analysed and useful information can be derived which helps to improve the structure of the web site.

14.4 CATEGORIES OF WEB DATA MINING

Web Data Mining can be divided into three categories according to the kinds of data to be mined.

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

Web content mining is the process of extracting information from the content of web documents on World Wide Web like mining the content of HTML pages. Web structure mining is the process of extracting knowledge from the link structure of the World Wide Web. Web usage mining is the process of discovering interesting patterns from the web access logs on servers. It is also called web log mining. Figure 14.1 represents the categories of web data mining.

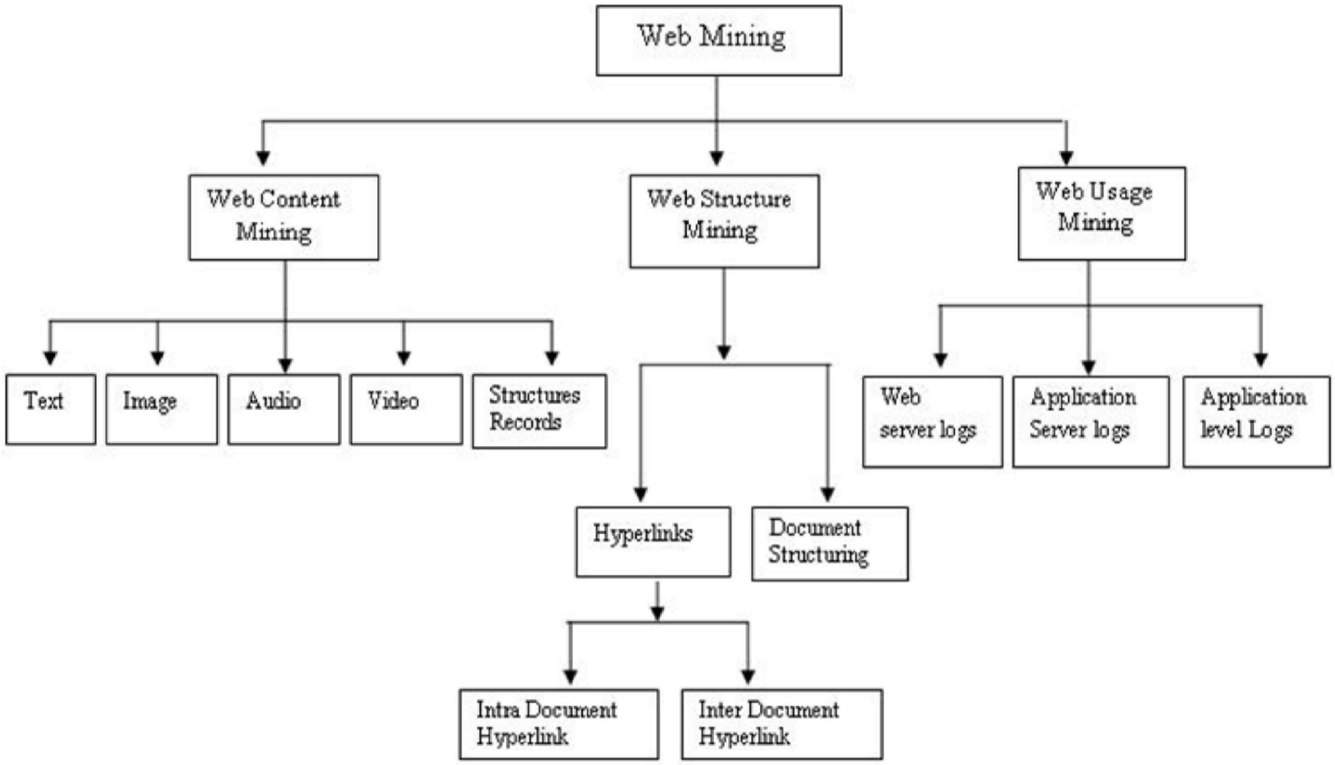


Figure 14.1: Categories of Web Data Mining

14.4.1 Web Content Mining

Web Content Mining is the task of extracting knowledge from the content of documents on World Wide Web. Web Content Mining is a form of text data mining applied to the web domain and has to do with finding the content of documents, classifying documents, and clustering documents. It focuses on finding web documents whose content matches specific criteria. Generally, web document includes the data like text, audio, video, image and hyperlinks. HTML documents are semi structured data. Due to this unstructured nature, web content mining becomes more complex. Web content mining focuses on automatic search of information resources online. The most studied forms of web content mining are text mining and its application including topic discovery, extracting association patterns, clustering and classification of Web Pages.

The multi layered database model was used to transform unstructured data on web into the form relevant to database. To find relevant documents, intelligent tools are used to extract information from web pages and the abstract information of the relevant documents is contained in the database. Using query language, characterizations are queried from the database. In case of other information, search engine is used to query web resources. The crawler searches the web for documents that match with defined classes. The found documents are saved into database and query techniques are used to query the database.

14.4.2 Web Structure Mining

Web Structure Mining focuses on discovering structure information from the Web to identify the relevant documents. It describes the connectivity in the Web subset based on the given collection of interconnected Web documents. The structure of a typical Web graph consists of Web pages as nodes and hyperlinks as edges connecting the related pages. Mining the site structure and Web page structure can help to guide the classification and clustering of pages to find authoritative pages to improve retrieval performance.

Unlike Web content mining, which mainly concentrates on the information of single document, web structure mining tries to discover the link structures of the hyperlinks between documents. By using information of hyperlinks, web structure mining can classify the Web pages and produce results such as the similarity and relationship between different Web sites. The structural data for Web structure mining is the link between the information and the document structure. Given a collection of web pages and topology, interesting facts related to page connectivity can be discovered. Web document contents can also be represented in a tree-structured format, based on the different HTML and XML tags within the page.

In the area of web structure mining, research on Social Network Analysis has been done. The separation of page is made with respect to the number of incoming and outgoing links. The methods have been used to calculate the quality and relevance relation of two web pages.

Web structure mining has another dimension, which is to discover the structure of Web document. This dimension of web structure mining is used to extract the schema of Web pages. This information is beneficial for navigation purpose and it provides comparison and integration of Web page schemes. Another task of web structure mining is to discover the nature of the hierarchy or network of hyperlink in the web sites of a particular domain. This may help to generalize the flow of information in Web sites that may represent some particular domain; therefore the query processing can be performed more easily and more efficiently. Web structure mining has a strong relation with the web content mining.

14.4.3 Web Usage Mining

Web Usage Mining is the process of applying data mining techniques to discover interesting patterns from the Web usage data. Web usage mining provides better a understanding for serving the needs of Web-based applications. Web Usage data keeps information about the identity or origin of the Web users with their browsing behaviour in a web domain. Web usage mining itself can be divided into subcategories based on the type of web usage data used. Web server data and application server data are common forms of web usage data.

Most of the research in Web usage mining is focused on applications using web Server Data. Most of the Web log analysis tools use only the textual data from log files. Web usage mining tries to discover useful information from the data extracted from the interactions of the users while surfing on the Web. It also has focus on the methods predicting user behaviour while the user interacts with Web.

The possible application areas of web usage mining are prediction of the user's behaviour within the site, comparison between expected and actual Web site usage, reconstruction of web site structure based on the interests of its users.

14.5 APPLICATIONS OF WEB DATA MINING IN E-COMMERCE

Using Web mining, deep analysis can be performed by combining other corporate information with Web traffic data. This allows accounting, customer profile, inventory, and demographic information to be correlated with Web browsing that answers complex questions such as:

How many persons purchased something, who visited web site?

Which advertising campaigns resulted in the most purchases?

Do my Web visitors fit a certain profile?

Can I use this for segmenting my market?

Web mining tools can be extended and programmed to answer almost any question. Web mining can provide the companies managerial insight into visitor profiles which help the top management to take strategic actions accordingly. Also, the company can obtain some subjective measurements using Web Mining on the effectiveness of their marketing campaign or marketing research that will help the business to improve and to align their marketing strategies in time.

The company can identify the strength and weakness of its web marketing campaign through Web Mining, and then can make strategic adjustments and obtain the feedback from Web Mining again to see the improvement. This procedure is an on-going continuous process. Some Web Mining applications in e-commerce have been discussed below.

14.5.1 Customer Attraction

The two essential parts of attraction are the selection of prospective new customers and the acquisition of selected potential candidates. One marketing strategy to perform this exercise is to find the common characteristics in already existing visitors' information and behaviour for the classes of profitable and non-profitable customers. These groups are then used as labels for a classifier to discover Internet marketing rules, which are applied online on site visitors. Depending on the outcome, a dynamically created page is displayed, whose contents depend on the found associations between browser information and the offered products or services.

14.5.2 Customer Retention

Customer retention is the step of managing the process of keeping the online shopper as loyal as possible. Due to the non-existence of physical distances between the providers, this is an extremely challenging task in the electronic commerce scenario. One strategy is similar to the acquisition that is dynamically creating web offers based on associations. However, it has been proved more successful to consider associations across time, also known as sequential patterns. The discovered sequence can be used to display special offers dynamically to keep a customer interested in the site, after a certain page sequence with a threshold support and/or confidence value has been visited.

14.5.3 Cross-Sales

The objective of cross-sales is to diversify selling activities horizontally or vertically to an existing customer base. Traditional generic cross-sales methodology has been adopted in order to perform the given task in an electronic commerce environment. For discovering thepotential customers,the characteristic rules of existing cross-sellers have to be discovered. The entire set of discovered rules can then be used as the model to be applied at run-time on incoming actions and requests from existing customers.

14.5.4 Improve the E-Commerce Web Site Design

Attractiveness of the site depends on its reasonable design of content and organizational structure. Web Mining can provide details of the user behaviour, providing web site designers the basis of decision making to improve the design of the site.

CHECK YOUR PROGRESS

Q.1: The main purpose for structure mining is to extract previously unknown relationships between ....

a) Web pages b) Web hyperlinks

c) Web data d) Web contents

Q.2: Web usage mining refers to the discovery of user access patterns from Web usage logs. (Say True / False)

Q.3: Web structure mining is the process of discovering .... information from the web.

a) Semi structured b) Unstructured

c) Structured d) None of the above

Q.4: The main purpose for content mining is to extract previously unknown relationships of ....

a) Web pages b) Web hyperlinks

c) Web use information d) Web contents

14.6 TEXT MINING

Text mining refers to retrieval of information that 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 growing rapidly 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. Now-a-days, most of the information in government, industry, business, and other institutions are stored electronically, in the form of text databases.

The 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 it may also contain some largely unstructured text components, such as abstract and contents.

14.6.1 Process of Text Mining

The process of Text mining involves a series of activities that are performed to mine the information. These activities are shown in the figure 14.2.

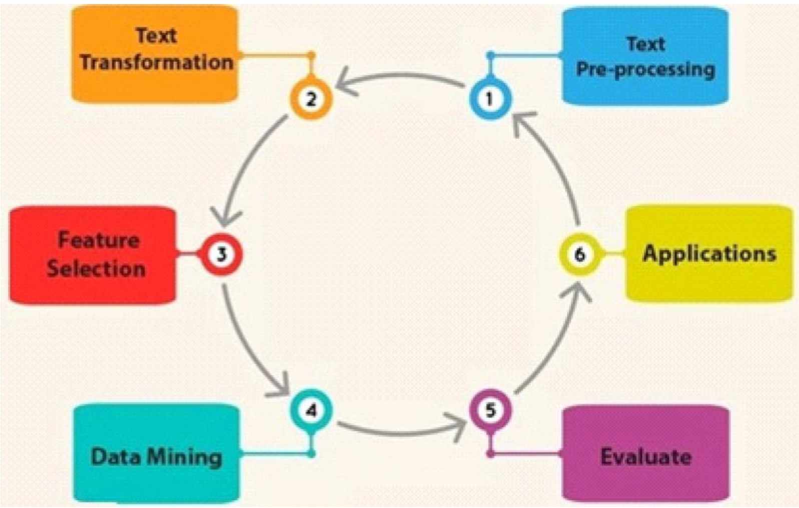


Figure 14.2: Process of Text Mining

Text Pre-processing: It involves a series of steps as shown below:

- **Text Cleanup:** Text Cleanup means removing any unnecessary or unwanted information, such as removing ads from web pages, normalizing text converted from binary formats.
- **Tokenization:** Tokenizing is simply achieved by splitting the text into white spaces.
- **Part of Speech Tagging:** Part-of-Speech (POS) tagging means word class assignment to each token. Its input is given by the tokenized text. Taggers have to cope with unknown words and ambiguous word-tag mappings.

Text Transformation (Attribute Generation): A text document is represented by the words it contains and their occurrences. Two main approaches to document representation are:

i) Bag of words

ii) Vector Space

Feature Selection (Attribute Selection): Feature selection also is known as variable selection. It is the process of selecting a subset of important features for use in model creation by removing the redundant and irrelevant features. Redundant features are the one which provides no extra information. Irrelevant features provide no useful or relevant information in any context.

Data Mining: At this point, the Text mining process merges with the traditional process. Classic Data Mining techniques are used in the text database and model the data as per user requirement

Evaluate: With the help of background knowledge, it is possible to evaluate patterns to find the interesting ones. Normally, a data mining system can discover thousands of patterns. Naturally, many of them are not interesting to the given user, because they represent either common knowledge or lack of novelty. The background knowledge in the form of user-specified constraints or interestingness expectation can be used to guide the discovery process and then reduce the search space.

14.6.2 Applications of Text Mining

Text Mining is applied in a variety of areas. Some of the most common areas are

Risk Management: One of the primary causes of failure in the business sector is the lack of proper or insufficient risk analysis. Adopting and integrating risk management software powered by text mining technologies such as SAS Text Miner can help businesses to stay updated with all the current trends in the business market and also can boost their abilities to mitigate potential risks. Since text mining technologies can gather relevant information from across thousands of text data sources and can create links between the extracted insights, it allows companies to access the right information at the right moment, thereby enhancing the entire risk management process.

Customer Care Service: Text mining techniques, particularly NLP, are finding increasing importance in the field of customer care.

Companies are investing in text analytics software to enhance their overall customer experience by accessing the textual data from varied sources such as surveys, customer feedback, and customer calls, etc. Text analysis aims to reduce the response time of the company and helpsaddress the grievances of thecustomers speedilyand efficiently.

Fraud Detection: Text analytics backed by text mining technologies provides a tremendous opportunity for domains that gather a majority of data in the text format. Insurance and finance companies are harnessing this opportunity. By combining the outcomes of text analyses with relevant structured data these companies are now able to process claims swiftly as well as to detect and prevent frauds.

Business Intelligence: Organisations and business firms have started to leverage text mining techniques as a part of their business intelligence. Apart from providing profound insights into customer behaviour and trends, text mining techniques also help companies to analyse the strengths and weaknesses of their rivals, thus giving them a competitive advantage in the market. Text mining tools such as Cogito Intelligence Platform and IBM text analytics provide insights on the performance of marketing strategies, the latest customer and the market trends etc.

Social Media Analysis: There are many text mining software packages designed exclusively for analysing the performance of social media platforms. These help to track and interpret the texts generated online from the news, blogs, emails, etc. Furthermore, text mining tools can efficiently analyse the number of posts, likes, and followers of our brand on social media, thereby allowing us to understand the reaction of people who are interacting with our brand and online content. The analysis will enable us to understand 'what's hot and what's not' for our target audience.

14.6.3 Approaches to Text Mining

Information Extraction: Information Extraction (IE) refers to the process of extracting meaningful information from vast chunks of textual data. This method focuses on identifying the extraction of entities, attributes, and their relationships from semi-structured or unstructured texts. Whatever information is extracted is then stored in a database for future access and retrieval. The efficacy and relevancyof the outcomes are checked and evaluated using precision and recall processes.

Information Retrieval: Information Retrieval (IR) refers to the process of extracting relevant and associated patterns based on a specific set of words or phrases. IR systems make use of different algorithms to track and monitor user behaviours and to discover the relevant data accordingly. Google and Yahoo search engines are the two most renowned IR systems.

Categorisation: Text categorisation is a form of “supervised” learning wherein normal language texts are assigned to a predefined set of topics depending upon their content. Thus, categorisation or rather Natural Language Processing (NLP) is a process of gathering text documents and processing and analysing them to uncover the right topics or indexes for each document.

Clustering: Clustering is one of the most crucial techniques of text mining. It seeks to identify the intrinsic structures in textual information and organise them into relevant subgroups or ‘clusters’ for further analysis. A significant challenge in the clustering process is to form meaningful clusters from the unlabeled textual data without having any prior information on them. Cluster analysis is a standard text mining tool that assists in data distribution or acts as a pre-processing step for other text mining algorithms running on detected clusters.

Summarisation: Text summarisation refers to the process of automatically generating a compressed version of a specific text that holds valuable information for the end user. The aim here is to browse through multiple text sources to craft summaries of texts containing a considerable proportion of information in a concise format, keeping the overall meaning and intent of the original documents essentially the same. Text summarisation integrates and combines the various methods that employ text categorisation like decision trees, neural networks, regression models, and swarm intelligence.

14.6.4 Mining Result Evaluation of Text Mining

For evaluating the results obtained by using the above techniques for text mining the concept of F measure is used widely. In the following section the concept of F-measure is discussed in detail.

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:

$$\text{precision} = \frac{|\{\text{Relevant}\}\ n\ \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$

Recall: This is the percentage of documents that are relevant to the query and were, in fact,retrieved. It is formally defined as:

$$\text{recall} = \frac{|\{\text{Relevant}\}\ n\ \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|}$$

F-score, which is defined as the harmonic mean of recall and precision:

$$\text{F\_score} = \frac{\text{recall} \times \text{precision}}{(\text{recall} + \text{precision})/2}$$

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

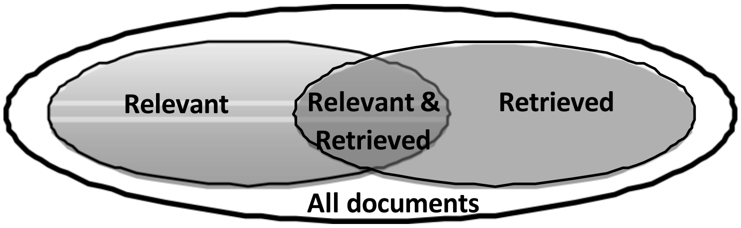


Figure 14.3: Example of relevant and retreived documents

14.7 EPISODE RULE DISCOVERY FOR TEXT

Episode rules and episodes are a modification of the concept of association rules and frequent sets applied to sequential data. Sequential data such as text can be seen as a sequence of pairs where feature vector consists of an ordered set of features and index contains information about the position of the word in the sequence It is a common practice that the sequence is represented in an increasing order of the indices.

Text episodes are as a pair α = (V, �) where V is a collection of feature vectors and � is a partial order on V. Given a text sequence S, a text episode α = (V, �) occurs within S if there is a way of satisfying the feature vectors in V using the feature vectors in S so that the partial order � is respected. Intuitively, this means that the feature vectors of V can be found within S in an order that satisfies the partial order �.

For an occurrence of the episode to be interesting, all feature vectors of the episode must occur close enough in S. The meaning of close enough is represented as *limit W*, also called as window size. Hence, instead of considering all occurrences of the episode in S, only need to examine the occurrences within substrings S' of S where the difference of the indices of the feature vectors in S' is within the range (close enough). Moreover, since there may be several partially differing occurrences of the episode within the substring S', mining can be restricted only to the distinct minimal occurrences of the episode.

The most useful types of partial orders are– (i) *total orders* i.e., the feature vectors of each episode have a fixed order; such episodes are called serial; and (ii) *trivial partial orders* where the order is not significant at all such episodes are called parallel. A typical example of a serial text episode is a phrase consisting of a sequence of related words with a specific meaning. A typical example of a parallel text episode is a collection of co-occurring terms which may describe the contents of a document better than any of the single term.

The support of α in S is defined as the number of minimal occurrences of α in S. Usually, mining process is interested only in episodes with a support exceeding a given support threshold, meaning that they occur in the sequence frequently enough not to be considered accidental.

An episode rule gives the conditional probability that a certain episode occurs or that a sub-episode has occurred. Formally, an episode rule is an expression $\beta[\mathrm{win}_1] = \alpha[\mathrm{win}_2]$,where β and α are episodes, β is a sub-episode of α and win_1 and win_2 are window sizes, win_1 � win_2. The confidence of the rule is the conditional probability that α occurs, given that β occurs under the window size constraints specified by the rule. An example of an episode rule could be:

knowledge discovery in[4] ⇒ databases[5] (85%)

which tells us that in 85 percent of the cases where the three words “knowledge discovery in” occurred within 4 consequent words, also the word databases occurred within 5 words.

CHECK YOUR PROGRESS

Q.5: In text pre-processing which of the following step is performed?

a) Summarization b) Tokenization

c) Data Extraction d) Data Recovery

Q.6: F-measure is used to evaluate the performance of an IR system (Say True/ False)

Q.7: Which of the following is not a text mining method

a) Summarization b) Clustering

c) Information Retrieval

d) Searching Text in Text Document

Q.8: Application of Text Mining can be found in–

a) Fraud Detection b) Video Analysis

c) Documentation d) None of the above

14.8 LET US SUM UP

- Web Mining is the application of data mining techniques to extract useful knowledge from web data like contents of web documents, hyperlinks structure of documents and web usage logs.
- Mining of web data focuses on well design of web sites and formation of techniques to analyse the behaviour of user. There are also strong requirements of techniques to help in business decision in e-commerce.
- Web Data can be broadly divided into three categories: Web content data, Web structure data and Web usage data.

- Web Data Mining can be divided into three categories according to the kinds of data to be mine: Web Content Mining, Web Structure Mining, Web Usage Mining
- Web content mining is process of extracting information from the content of web documents on World Wide Web like mining the content of HTML pages.
- Web structure mining is the process of extracting knowledge from the link structure of the World Wide Web.
- Web usage mining is the process of discovering interesting patterns from web access logs on servers.
- Text mining refers to retrieval of information is stored in text databases or document database, 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.
- Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured.
- The process of Text mining involves a series of activities that are performed to mine the information. These activities are: Text Pre-processing, Text Transformation, Feature Selection, Data Mining, Evaluation of results, Application.
- Application of Text mining includes: Risk Management, Customer care service improvement, Fraud Detection, Business Intelligence, Social Media Analysis etc.
- Approaches of text mining includes: Information Extraction, Information Retrieval, Categorisation, Clustering, Summarisation.
- For evaluate the results obtained by using the text mining techniques we use F-measure.
- Episode rules and episodes are a modification of the concept of association rules and frequent sets applied to sequential data.
- Text episodes are as a pair α = (V, �) where V is a collection of feature vectors and � is a partial order on V.
- An episode rule gives the conditional probability that a certain episode occurs or that a subepisode has occurred.

14.9 FURTHER READING

1) Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.
2) Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
3) Tan, P. N. (2018). *Introduction to Data Mining*. Pearson Education India.

14.10 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: (a)

Ans. to Q. No. 2: True

Ans. to Q. No. 3: (c)

Ans. to Q. No. 4: (d)

Ans. to Q. No. 5: (b)

Ans. to Q. No. 6: True

Ans. to Q. No. 7: (d)

Ans. to Q. No. 8: (a)

14.11 MODEL QUESTIONS

Q.1: What is Web Mining? What are the different types of web data available for mining? Explain briefly.

Q.2: What are the different types of web mining? Explain briefly.

Q.3: Discuss the different applications of web mining.

Q.4: What is text mining? Explain the different steps of text mining.

Q.5: Explain the different applications of text mining.

Q.6: Explain the different techniques for text mining?

Q.7: What is F-measure? What is meant by Precision and Recall? Explain in detail.

Q.8: What is Episode and Episode rule? Why is episode rule discovery from text important? Explain.

*** ***** ***

UNIT 15: INTRODUCTION TO SPATIAL AND TEMPORAL DATA MINING

UNIT STRUCTURE

15.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define temporal data mining
- describe the different temporal data mining tasks
- define temporal association rule mining
- define sequence mining and spatial data mining

- describe the different spatial data mining tasks
- describe spatial clustering algorithms.

15.2 INTRODUCTION

This is the last unit of this course. In the previous unit, we have learned about web mining and text mining. We have also learned about the different categories of web mining along with the applications of web mining. In this unit, we will discuss temporal data mining, its types and different temporal data mining tasks. We will also learn about the different temporal association rules and sequence mining. Spatial data mining and different spatial clustering are also discussed in detail in this unit.

15.3 TEMPORAL DATA MINING

Temporal Data Mining is an important extension to the conventional data mining techniques. *Temporal data mining* is concerned with data mining of large sequential data sets containing the time of occurrence as one of the attributes. In temporal data sets, data are ordered with respect to time. Temporal data mining can be defined as the searchfor interesting correlations or patterns in large sets of temporal data. It has the capability to discover patterns or rules which might be overlooked when the temporal component is ignored or treated as a simple numeric attribute. By taking into account the time aspect, more interesting patterns are extracted. A large volume of research has, therefore, been focused on temporal data mining to discover temporal rules such as sequential patterns, episodes, temporal association rules and inter transaction association rules, etc.

Time series analysis has quite a long history. Techniques for statistical modelling and spectral analysis of real or complex-valued time series have been in use for more than fifty years. Weather forecasting, financial or stock market prediction and automatic process control have been some of the oldest and most studied applications of such time series analysis. Few people contradict temporal data mining with time series analysis. Following are the few differences between these two.

One main difference between temporal data mining and classical time series analysis lies in the size and nature of data sets and the manner in which the data is collected. Often, temporal data mining methods must be capable of analysing data sets that are prohibitively large for conventional time series modelling techniques to handle efficiently.

The second major difference lies in the kind of information that we want to estimate or unearth from the data. The scope of temporal data mining extends beyond the standard forecast or control applications of time series analysis. Very often, in data mining applications, one does not even know which variables in the data are expected to exhibit any correlations or causal relationships.

Most of the work has been done in the field of temporal data mining which is on single timestamp data, where every piece of information is associated with a timestamp. But not all temporal events can be represented by single timestamp for example, in an e-learning system the user's login and logout time or the user's access time or in a cellular phone company, time and length of each phone call made by the customers, cannot be represented by single timestamps. It requires two timestamp, starting time and ending time of the event or the duration of the event. These types of data are referred to as interval data. Time or duration is not the only attribute that can be described by intervals. There are some other attributes like salary, estimated expenditure, weight, height etc. that can be described by intervals.

Temporal data mining is concerned with data mining of large sequential data sets. By sequential data, it is meant that data is ordered with respect to some index. For example, time series constitute a popular class of sequential data, where records are indexed by time. Other examples of sequential data could be text, gene sequences, protein sequences and lists of moves in a chess game. Although there is no notion of time here as such, the ordering among the records is very important and is central to the data description modelling.

Analysis of temporal time varying data facilitates many interesting challenges. For example, there may be many different interpretations for time. The date stored in the record is the date representing when that information becomes current. This is often called the valid time. The valid time for information is the time during which the information is true in the modelled world. This usually consists of a start time and an end time. Another time that could have been used is the transaction time. The transaction time is the timestamp associated with the transaction that inserted this record. This could be different from the start time for the valid time interval. The transaction time interval is the time the tuple actually existed in the database. Other types of times may be used as well.

The temporal data often involve the duration of time; that is, a start time and an end time. In this interpretation, the range of values t_s, t_e is associated with each record. Here t_s is the start time and t_e is the end time. A timestamp of a specific time instance may be used instead of a range.

Satellites continually collect images and sensory data. This information is temporal and is associated with specific points in time. When the data were obtained in a hospital, printouts of heartbeats may be kept for patients which represent a continuous view of temporal data. When an electro encephalograph (EEG) is taken for a patient, several different brain waves are measured in parallel. Each wave represents a continuous set of data over time.

Temporal databases usually do not accept the same types of updates and queries as traditional databases. Instead, the only updates that can be allowed are new corrections and versions. Actual modifications of tuple usually are not allowed. Instead, a new tuple with a different valid time would be added.

15.3.1 Types of Data for Temporal Data Mining

Temporal data mining is an important extension of the data mining and it is non-trivial extraction of implicit, potentially useful and previously unrecorded information with an implicit or explicit temporal content, from large database. Data used in temporal data mining can be categorized as following:

Sequences: Sequences are ordered sequence of the events or transaction. Though there may not be any explicit reference to time, there exists a sort of qualitative temporal relationship between data items. In transactional data a transaction T_i may occur before another transaction T_j or T_i may occur after T_j etc. These types of temporal relationship may be before, after, during, meet and overlap etc. Such relationships are called qualitative relationship between time events.

Time Stamped: This category of the temporal data has explicit time related information. Relationship can be quantitative i.e., we can find the exact temporal distance between data element. The inference made through this type of data may be temporal or non-temporal.

Time Series: Time series data is a special case of the time stamped data. In time series data events have uniform distance on the time scale.

Fully Temporal: Data of this category is fully time dependent. The inferences are also strictly temporal.

15.4 TEMPORAL DATA MINING TASK

The possible objectives of data mining, which are often called tasks of data mining, can be classified into some broad groups. For the case of temporal data mining, these tasks may be grouped as Association rule mining, Prediction, Classification, Clustering, Characterization, Search & retrieval, Pattern discovery, Trend Analysis and Sequence Analysis. These categorization is neither unique nor exhaustive-

1) **Temporal Association Rule Mining:** The discovery of relevant association rules is one of the most important methods used to perform data mining on transactional databases. An effective algorithm to discover association rules is the Apriori. Association rule discovery is an important task in data mining in which we extract the relation among the attribute on the basis of support and confidence. The association rule discovery can be extended to temporal association.

   In general association rules are represented as $X \Rightarrow Y$, which states that if X occurs then Y will occur. The same can be extended to a temporal association rule mining as: $X \Rightarrow (T\ Y)$, which states that if X occurs then Y will occur within time T. Stating a rule in this new form, enables us to control the impact of the occurrence of an event to the other event occurrence, within a specific time interval.

A cyclic rule is one that occurs at regular time intervals, for example heavy rainfalls occur only in the months of June and July. In order to discover this type of rules, an efficient approach needs to be designed. Discovery of cyclic association rules consists of two processes: first, discover the cyclic large item sets and second, generate the rules. Extension to this method consists in allowing the existence of different time units, such as days, weeks or months, and is achieved by defining calendar algebra to define and manipulate groups of time intervals. The rules discovered of such type are called cylindrical association rules.

2) **Prediction:** Prediction has a versatile significance in the data mining. It is the forecasting for future on the basis of the past. The task of time-series prediction has to do with forecasting future values of the time series based on its past samples. For prediction a predictive model is designed for the data. There are many nonlinear models for time series prediction such as Neural Networks, neural networks is good for nonlinear modelling of time series data.

   In many cases, prediction may be formulated as classification, association rule finding or clustering problems. Generative models can also be used effectively to predict the evolution of time series. Prediction problems have some specific characteristics that differentiate them from other problems. Prediction gains the importance in various fields like medical, finance, environmental and engineering with an exponential rate.

3) **Classification:** In classification, one classifies the unknown set of attributes in any one of the predefined classes. In temporal classification, each temporal sequence presented in the database is assumed to be belonging to one of the predefined classes or categories and our goal is to automatically determine the corresponding category/class for the given input temporal set of

attributes. There are many examples of sequence classification applications, like handwriting recognition, speech recognition, gesture recognition, demarcating gene and non-gene regions in a genome sequence, on-line signature verification, etc.

4) **Clustering:** Clustering techniques are used to divide the data in the groups on the basis of similarity measure. There exist several clustering technique algorithms like K-means, K-medoids etc. Clustering of sequences or time series is concerned with grouping a collection of time series or sequences based on their similarity. Clustering is of particular interest in temporal data mining since it provides an attractive mechanism to automatically find some structure in large data sets that would be otherwise be difficult to summarize or visualize. There are many applications where a time series clustering activity is relevant e.g., web activity logs. Clusters can indicate navigation patterns of different user groups. Another example could be clustering of biological sequences like proteins or nucleic acids so that sequences within a group have similar functional properties.

5) **Characterization:** Characterization is nothing but summarization of the general characteristics or features of a target class of data. Characterization can be extended to temporal data. An interesting experiment would be extending to the concept of decision tree construction on temporal attributes. For example a rule could be: the first case of filarial is normally reported after the first pre-monsoon rain and during the month of May-August. The output of characterization can be represented in various forms e.g., pie charts, bar charts, curves, multidimensional data cubes.

6) **Search and Retrieval:** The problem of searching is concerned with efficiently locating subsequences often referred to as queries in large archives of sequences or sometimes in a single long sequence. Query-based searches have been extensively studied in language and automata theory. However the problem of efficiently locating exact matches of substrings is well solved, the situation is different

when looking for approximate matches. In content-based retrieval, a query is presented to the system in the form of a sequence. The task is to search a typically large data base of sequential data and retrieve from it sequences or subsequences similar to the query sequence.

Essentially we need to properly insert 'gaps' in the two sequences or decide which should be corresponding elements in the two sequences. Time warping methods have been used for sequence classification and matching for many years. In speech applications, Dynamic Time Warping (DTW) is a systematic and efficient method based on dynamic programming that identifies which correspondence among the feature vectors of two sequences is the best when scoring the similarity between them. In recent times, there are many situations in which sequence matching are used. For example, many biological sequences such as genes, proteins, etc.

7) **Pattern Discovery:** Unlike in search and retrieval applications, in pattern discovery there is no specific query in hand with which to search the database. The objective is simply to unearth all the patterns of interest. Sequence prediction, Classification, Clustering and Search and retrieval methods originally come from other disciplines like estimation theory, machine learning or pattern recognition. On the other hand, the pattern discovery task has its origins in data mining itself. Pattern discovery is exploratory and an unsupervised nature of operation, which is something of a sole preserve of data mining.

8) **Trend Analysis:** Trend analysis in temporal database is referred to change in attribute due to the change in time. The analysis of one or more time series of continuous data may show similar trends, i.e. similar shape across the time axis. Here we are trying to find the relationships of change in one or more static attributes, with respect to changes in the temporal attributes. Trend analysis is very useful for the decision support system and the decision maker.

15.5 TEMPORAL ASSOCIATION RULES

The problem of the discovery of association rules comes from the need to discover patterns in transaction data in a supermarket. But transaction data in supermarkets are temporal. When gathering data about products purchased in a supermarket, the time of the purchase is registered in the transaction. In large volume temporal data, as used for data mining purposes, information may be found related to products that did not necessarily exist throughout the data collection period. So some, products may be found which, at the moment of mining are already discontinued. There may be also new products that were introduced after the beginning of the data collection. Some of these new products will participate in the associations, but may not be included in association rule as support count of such item is not up to the mark. For example, if the total number of transactions is 3,000,000 and we fix the minimum support as 0.5% which is 150,000. Let a particular product has been sold during the last 30 months and has just the minimum support i.e. it appears on an average 5,000 transactions per month. Now consider another product that was incorporated in the last 6 months and that appears in 20,000 transactions per month. The total number of transactions in which it occurs is 120, 000 so the product is infrequent as it has not attained the minimum support threshold even if it is four times as popular as the first one. However, if we consider the introduction time of the product in the market, its support may be more than the minimum support threshold. So, by introduction of time aspect to traditional association, rule mining is more interesting and potentially useful association rules can be mined.

One more example can be when large volumes of data from supermarket are used for mining. There may be information related to certain products that have not existed throughout the dataset but they appear only in certain period. If we mine such datasets without considering the time aspect we may miss out some interesting association rules. For example, cold drinks are specially sold in the summer and if we consider the whole dataset without considering the time aspect, cold drink may be participated in any of the association rules as it has not attained the minimum support threshold.

One way to solve this problem is by incorporating time in the model of discovery of association rules. We will call these rules Temporal Association Rules. A temporal association rule can be defined as a pair (R, T) where R is an association rule and T is a temporal feature, such as a period or a calendar.

One sub product of this idea is the possibility of eliminating outdated rules, according to the user criteria. Moreover, it is possible to delete obsolete itemsets by considering the lifetime of the itemset which will reduce the amount of work to be done in the determination of the frequent itemsets and hence, in the determination of the rules.

The temporal association rules are the extension of the non-temporal model. The basic idea was to limit the search for frequent sets of items or itemsets to the lifetime of the itemset's members. For that reason, the concept of temporal support was introduced. Thus, each rule has an associated timeframe, corresponding to the lifetime of the itemsparticipating in the rule. If the extent of a rule's lifetime exceeds a minimum stipulated time set by the user, then the rule is analysed for its validity in that period. This concept allows us to find rules with the traditional frequency counting algorithms.

15.5.1 Temporal Rules Discovery

The discovery of all the temporal association rules in a transaction set ***D*** *can be made in two phases*

Phase 1: Finding of all the frequent itemsets in their lifespan i.e. the itemsets whose frequency exceeds the user's specified minimum support cr and temporal support.

Phase 2: Using the above frequent sets to find the association rules.

15.5.2 Cyclic Association Rules

If the association rules are computed over monthly sales data, some seasonal variation may be observed where certain rules are true approximately in the same month each year. Similarly, association rules can also display regular hourly, daily, weekly, etc. which are called cyclic association rules.

15.5.3 Calendric Association Rules

If the temporal association rules with time intervals which follow some user-given calendar schema then it is called calendric association rules. An example of such schema is (year, month, day), which yields a set of calendar based patterns. Calendric association rules are repeated with the calendar schemas.

CHECK YOUR PROGRESS

Q.1: State whether the following statements are True or False:

a) Temporal Data Mining is an important extension of conventional data mining techniques.

b) Temporal data mining and classical time series analysis are same.

c) Temporal association rules are represented as X ⇒ (T Y).

d) Pattern discovery is an temporal data mining task.

Q.2: Which of the following dataset is not temporal?

a) Sequence dataset b) Interval dataset

c) Time series dataset d) None of the above

Q.3: Which of the following is not a temporal data mining task?

a) Classification b) Prediction

c) Select data from dataset d) Pattern Discovery

15.6 SEQUENCE MINING

Sequence mining is a data mining task specialized for analysing sequential data, to discover sequential patterns. More precisely, it consists of discovering interesting sub-sequences in a set of sequences, where the interestingness of a subsequence can be measured in terms of various criteria such as its occurrence frequency, length or profit. Sequence mining has numerous real-life applications due to the fact that data is naturally encoded as sequences of symbols in many fields such as bioinformatics, e-learning, market basket analysis, texts, and webpage click-stream analysis.

Let us discuss the sequence mining with an example. Consider the following sequence database, representing the click/opening of web pages by users.

SID	Sequences
1	<{a,b}, {c}, {f,g}, {g}, {e}>
2	<{a,d}, {c}, {b}, {a,b,e,f}>
3	<{a}, {b}, {f,g}, {e}>
4	<{b}, {f,g}>

This database contains four sequences. Each sequence represents the click stream by users at different times. A sequence is an ordered list of click streams. For example, in this database, the first sequence (SID 1) indicates that an user opens the webpage *a* and then opens page *b* by following hyperlink from page *a*, then in some other time opens the webpage *c*, then opens webpage *f* and *g*, then opens page *g*, and then finally open page *e* in some other time*.*

Traditionally, sequence mining is being used to find sub-sequences that appear often in a sequence database, i.e., that are common to several sequences. Those sub-sequences are called the frequent sequential patterns. For example, in the context of our example, sequence mining can be used to find the sequences of click-streams that is followed by most of the users. This can be useful to understand the behaviour of users and to arrange link to different webpages so that a user can find those webpages easily.

To perform sequence mining, a user must provide a sequence database and specify a parameter called the minimum support threshold. This parameter indicates a minimum number of sequences in which a pattern must appear to be considered frequent, and be shown to the user. For example, if a user sets the minimum support threshold to 2 sequences, the task of sequence mining consists of finding all sub-sequences appearing in at least 2 sequences of the input database. In the example database, 30 sub-sequences met this requirement. These sequential patterns are shown in the table below, where the number of sequences containing each pattern (called the *support*) is indicated in the right column of the table as given below.

Pattern	Support
<{a}>	3
<{a}, {g}>	2
<{a}, {g}, {e}>	2
<{a}, {f}>	3
<{a}, {f}, {e}>	2
<{a}, {c}>	2
<{a}, {c}, {f}>	2
<{a}, {c}, {e}>	2
<{a}, {b}>	2
<{a}, {b}, {f}>	2
<{a}, {b}, {e}>	2
<{a}, {e}>	3
<{a}, {f, g}>	2
<{a, b}>	2
<{b}>	4
<{b}, {g}>	3
<{b}, {g}, {e}>	2
<{b}, {f}>	4
<{b}, {f, g}>	3
<{b}, {f}, {e}>	2
<{b}, {e}>	3
<{c}>	2
<{c}, {f}>	2
<{c}, {e}>	2
<{e}>	3
<{f}>	4
<{f, g}>	3
<{f}, {e}>	2
<{g}>	3
<{g}, {e}>	2

For example, the patterns <{a}> and <{a}, {g}> are frequent and have a support 3 and 2 respectively. In other words, these patterns appear in 3 and 2 sequences of the input database. The pattern <{a}> appears in the sequences 1, 2 and 3, while the pattern <{a}, {g}> appears in sequences 1 and 3. These patterns are interesting as they represent some behaviour common to several users. In real life sequence mining can actually be applied on database containing hundreds of thousands of sequences.

Another example of application of sequence mining is text analysis. In this context, a set of sentences from a text can be viewed as sequence database, and the goal of sequence mining is then to find sub-sequences of words frequently used in the text. If such sequences are contiguous, they are called “n-grams”.

15.7 SPATIAL DATA MINING

Spatial data mining is the application of data mining to spatial models. In spatial data mining, analysts use geographical or spatial information to produce business intelligence or other results. This requires specific techniques and resources to get the geographical data into relevant and useful formats.

Challenges involved in spatial data mining include identifying patterns or finding objects that are relevant to the questions that drive the research project. Analysts may be looking in a large database field or other extremely large data set in order to find just the relevant data, using GIS/GPS tools or similar systems.

One interesting thing about the term “spatial data mining” is that it is generally used to talk about finding useful and non-trivial patterns in data. The core goal of a spatial data mining project is to distinguish the information in order to build real, actionable patterns to present, excluding things like statistical coincidence, randomized spatial modelling or irrelevant results. One way analysts may do this is by combing through data looking for “same-object” or “object-equivalent” models to provide accurate comparisons of different geographic locations.

Spatial data mining is a non-trivial process to extract knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases. For spatial data mining it is required to integrate data mining with spatial database technologies. It can be used to understand spatial data, relationships between spatial and non-spatial data, discovery of spatial relationships, constructing spatial knowledge bases, reorganizing spatial databases, and optimizing spatial queries. It has a wide application in geographic information systems (GIS), vector and raster image database, geo-marketing, remote sensing, image database exploration, environmental studies, and many other areas where spatial data are used. However, in comparison to extracting patterns from traditional numeric and characterized data, it is much difficult to extract interesting and useful patterns from spatial databases because of the complexity of spatial data types, spatial relationships, and spatial autocorrelation.

A huge volume of data is being generated with the growing production of maps which exceed people's capacity to analyse them. Voluminous geographic data have been, and continue to be, collected with modern data acquisition techniques such as global positioning systems (GPS), high-resolution remote sensing, location-aware services and surveys, and internet-based volunteered geographic information. Therefore the knowledge discovery methods like data mining can appropriately be applied to spatial data. This recent concept is an extension of the conventional data mining tasks and is applied to alphanumerical and spatial data. However in spatial data mining the spatial analysis takes into account spatial relations between objects. During the last few years, due to the widespread applications of GPS technology, web-based spatial data sharing and mapping, high-resolution remote sensing, and location-based services, more and more research domains have created or gained access to high-quality geographic data to incorporate spatial information and analysis in various studies, such as social analysis and business applications.

Spatial data mining is becoming quite popular as there are many applicative areas such as ecology, transportation, epidemiology etc, where the concept is being utilized. There are many diversified approaches and algorithms that have been proposed to extract knowledge from such data. However the main challenge to apply any of the algorithms and how to automate the data preparation task, which consumes most of the effort and time required for knowledge discovery in geographic databases. In spatial databases data are stored in different relations and it is required to join them spatially for finding out novel and useful pattern of data.

Since the conventional data mining methods are not suited for spatial data as they neither support location data nor the implicit relationships between spatial objects. Therefore, it is needed to develop some new methods of spatial data mining to find spatial relationships. Calculating the spatial relationship is a time consuming task in spatial database because the data is generated by encoding geometric location. Because of this complexity the global performances will suffer.

However, it is highly desired that the existing methods of data mining are extended and incorporated into spatial data mining methods. Spatial and GIS methods are crucial for spatial join, data access and graphical map display. The conventional data mining methods are able to generate knowledge about alphanumerical properties only.

15.8 SPATIAL DATA MINING TASKS

Spatial data mining tasks are the extension of conventional data mining tasks where spatial data and criteria are combined. The following tasks are performed in spatial data mining:

i) Spatial and non-spatial data are summarized,

ii) Deviations detection is done after looking for general trends,

iii) Classification rules are discovered,

iv) Clusters of similar objects are formed and

v) Associations and dependencies identified to characterize data.

To carry out these tasks different methods are used. Some of these methods are derived from statistics and others from the field of machine learning.

i) DATA SUMMARIZATION: It is the most common approach in which we apply elementary statistics, such as the calculation of mean, average, variance, histogram and pie chart formation etc. To establish spatial relationship between the objects we use the concept of contiguity matrix. Contrary to the autocorrelation measure, density analysis forms part of Exploratory Spatial Data Analysis (ESDA), which does not require any knowledge about data. In this the non-spatial properties are ignored. In geographic data analysis both the alphanumerical property data and spatial data are taken into consideration.

ii) **GENERALISATION:** In this method, abstract level of non-spatial attributes is raised and the details of geometric description are reduced by merging adjacent objects. It is derived from attribute-oriented induction concept. The concept hierarchy is the important representation which can be spatial or the thematic map of non-spatial data. Generalizations are of two types:

1) Non-spatial dominant generalization: In this the thematic hierarchy is used first and then adjacent objects are merged.
2) Spatial dominant generalization: This generalization is based on a spatial hierarchy. After this the aggregation or generalization of non-spatial values for each generalized spatial value is formed.

iii) **CLASSIFICATION:** Classification is about grouping data items into classes or categories according to their properties. Classification is also called supervised learning, it needs a training dataset to train the classification model, a testing dataset is used to evaluate the performance of the trained model. Classification methods include, decision trees, artificial neural networks (ANN), maximum likelihood estimation (MLE), linear discriminant function (LDF), support vector machines (SVM), nearest neighbour methods and case-based reasoning (CBR).

iv) **CLUSTERING:** Cluster analysis is mainly used for data analysis. It organizes a set of data items into groups known as clusters. The items in the same cluster are similar to each other and different from those in other clusters. Many different clustering methods have been developed in various research fields such as statistics, pattern recognition, data mining, machine learning, and spatial analysis.

To consider spatial information in clustering, three types of clustering analysis have been studied, spatial clustering, regionalization, and point pattern analysis. For the first type, spatial clustering, the similarity between data points or clusters is defined with spatial properties. Spatial clustering methods can be partitioning or hierarchical, density-based, or grid-based.

v) **SPATIAL DATA DEPENDENCIES:** It is the local autocorrelation method to work with spatial data which reflects how data are related. Spatial association rule, which works on spatial data dependence, have been adapted to spatial data. Spatial auto-correlation is concerned with the assessment of the degree of spatial dependence. This is done by using the concept of spatial weight matrix. It is equivalent to a residual test in regression analysis. This makes it possible to measure the difference between the actual spatial distribution of variable values and a random one.

15.9 SPATIAL CLUSTERING

Spatial clustering is a process of grouping a set of spatial objects into groups called clusters. Objects within a cluster show a high degree of similarity, whereas the clusters are as much dissimilar as possible. Outlier is a data point that does not conform to the normal points characterizing the data set. Detecting outlier has important applications in data mining as well as in the mining of abnormal points for fraud detection, stock market analysis, intrusion detection, marketing etc.

Important data characteristics, which affect the efficiency of clustering techniques, include the type of data (nominal, ordinal, numeric), data dimensionality (since some techniques perform better in low-dimensional spaces) and error (since some techniques are sensitive to noise).

Clustering of points is a most typical task in spatial clustering, and many kinds of spatial objects clustering can be abstracted as or transformed to points clustering. Here we will discuss about various kinds of spatial clustering. Spatial clustering is the most complicated among all the spatial data mining techniques.

Based on the technique adopted to define clusters, the clustering algorithms can be divided into four broad categories:

a) Hierarchical clustering methods (AGNES, BIRCH etc.),

b) Partitional clustering algorithms (K-means, K-medoids etc.),

c) Density-based clustering algorithms (DBSCAN DENCLUE etc.),

d) Grid based clustering algorithms (STING etc.).

Many of these can be adapted to or are specially tailored for spatial data.

15.9.1 Clustering Algorithms Based on Hierarchical Methods

Hierarchical clustering is a conventional clustering method which has wide variety of applications in different domains. Mainly, it is of two types:

a) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and

b) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).

Both this algorithm are exactly reverse of each other.

Agglomerative Nesting (AGNES): AGNES proceeds by a series of fusions of similar objects. Initially all objects are apart– each object forms a small cluster by itself. At the first step, two closest or minimally dissimilar objects are merged to form a cluster. Then the algorithm finds a pair of objects with minimal dissimilarity. If there are several pairs with minimal dissimilarity, the algorithm picks a pair of objects at random. Picking of pairs of objects, which are minimally dissimilar. For finding out the minimally dissimilar objects dissimilarity measures are used. AGNES computes Agglomerative Coefficient (AC), which measures the clustering structure of the data set.

Divisive Analysis (DIANA): DIANA is a hierarchical clustering technique, but its main difference with the agglomerative method (AGNES) is that it constructs the hierarchy in the inverse order. Initially, there is one large cluster consisting of all n objects. At each subsequent step, the largest available cluster is split into two clusters until finally all clusters, comprise of single objects. Thus, the hierarchy is built in n-1 steps. In the first step of an agglomerative method, all possible fusions of two objects are considered leading to combinations. In the divisive method based on the same principle, there are possibilities to split the data into two clusters. This number is considerably larger than that in the case of an agglomerative method.

When applying AGNES and DIANA to the spatial data, each object in the data set is replaced by spatial points. These spatial points are similar to those in is the 2D plane with x and y coordinates. The main difference is that spatial points are characterized by latitude and longitude. The latitude is the location of a place on the earth, north or south of the equator and longitude is the east – west measurement of position on the earth, measured from a plane running through polar axis. Here, the similarity between objects is expressed in terms of the Euclidian distance between two points in the n-dimension.

15.9.2 Clustering Algorithms for Partitional Clustering

Partitional clustering is another kind of conventional clustering method which decomposes a data set into a set of disjoint clusters. Given a data set of N points, a partitioning method constructs K (N e” K) partitions of the data, with each partition representing a cluster. That is, it classifies the data into K groups by satisfying the following requirements:

1) each group contains at least one point, and
2) each point belongs to exactly one group.

Notice that for fuzzy partitioning, a point can belong to more than one group.

K-Means Algorithm: K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a givendata set through a certain number of clusters fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done. At this point, re-calculate k new centroids of the clusters resulting from the previous step. After having these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. This procedure has to repeat until there is no change in the sets of centroids between two iterations.

K-Medoid Algorithm: The k-medoid algorithm is a clustering algorithm related to the k-means algorithm. Both the k-means and k-medoid algorithms are partitional and both attempt to minimize the distance between points labelled to be in a cluster and a point designated as the centre of that cluster. In contrast to the k-means algorithm, k-medoids chooses datapoints as centres and works with an arbitrary matrix of distances between data points. A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal. In some applications, we need each centre to be one of the points itself. This is where K-medoid comes in.

Even though these algorithms can be used for spatial clustering, but the main disadvantage is that these methods are sensitive to the noise and centre point. Another disadvantage is that these can detect only spherical clusters. So, we prefer density based clustering methods for spatial related applications than hierarchical and partitional.

15.9.3 Clustering Algorithms Based on Density Based Methods

Density-based approaches apply a local cluster criterion. Clusters are regarded as regions in the data space in which the objects are dense, and which are separated by regions of low object density (noise). These regions may have an arbitrary shape and the points inside a region may be arbitrarily distributed. Density based clustering algorithm includes DBSCAN, DENCLUE, OPTICS etc.

DBSCAN: DBSCAN requires two parameters: c (eps) and the minimum number of points required to form a cluster (minPts). It starts with an arbitrary starting point that has not been visited. This point's c-neighbourhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labelled as noise. Note that this point might later be found in a sufficiently sized c-environment of a different point and hence is made part of a cluster. If a point is found to be a dense part of a cluster, its c-neighbourhood is also part of that cluster. Hence, all points that are found within the c-neighbourhood are added. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discoveryof a further cluster or noise.

DENCLUE: DENCLUE (DENsity basted CLUstEring) is a generalization of partitioning, locality based and hierarchical or grid-based clustering approaches. The algorithm models the overall point density analytically using the sum of the influence functions of the points. Determining the density-attractors causes the clusters to be identified. DENCLUE can handle clusters of arbitrary shape using an equation based on the overall density function.

The approach of DENCLUE is based on the concept that the influence of each data point on its neighbourhood can be modelled mathematically. The mathematical function used, is called an impact function. This impact function is applied to each data point and the density of the data space is the sum of the influence function for all the data points. In DENCLUE, since many data points do not contribute to the impact function, local density functions are used. Local density functions are defined by a distance function like Euclidean distance.

The local density functions consider only data points which actually contribute to the impact function. Local maxima, or density-attractors identify clusters. These can be either centre-defined clusters, similar to k-means clusters, or multi-centre-defined clusters, that is a series of centre-defined clusters linked by a particular path. Multi-centre-defined clusters identify clusters of arbitrary shape. Clusters of arbitrary shape can also be defined mathematically.

15.9.4 Clustering Algorithms Based on Grid-Based Methods

Grid based methods quantize the object space into a finite number of cells (hyper-rectangles) and then perform the required operations on the quantized space. The main advantage of Grid based method is its fast processing time which depends on number of cells in each dimension in quantized space. Some grid based methods are CLIQUE (CLustering In QUEst), STING (STatistical INformation Grid), Wave clustering etc.

STING (STatistical INformation Grid based method): This is a grid based multi resolution clustering technique in which the spatial area is divided into rectangular cells(using latitude and longitude) and employs a hierarchical structure. Corresponding to different resolution, different levels of rectangular cells are arranged to form a hierarchical structure. Each cell at a higher level is partitioned into a number of cells at its immediate lower level and so on. Statistical information associated with each cell is calculated and is used to answer queries.

Wave Clustering: Wave Cluster is a multi-resolution clustering algorithm. It is used to find clusters in very large spatial databases. Given a set of spatial objects O_i, 1 � i � N, the goal of the algorithm is to detect clusters. It first summarizes the data by imposing a multi-dimensional grid structure on to the data space. The main idea is to transform the original feature by applying wavelet transform and then find the dense regions in the new space. A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub bands.

CLIQUE (CLustering In QUEst): The CLIQUE algorithm integrates density based and grid based clustering unlike other clustering algorithms described earlier, CLIQUE is able to discover clusters in the subspace of the data. It is useful for clustering high dimensional data which are usually very sparse and do not form clusters in the full dimensional space.

In CLIQUE, the data space is partitioned into non-overlapping rectangular units by equal space partition along each dimension. A unit is dense if the fraction of total data points contained in it exceeds an input model parameter.

CLIQUE performs multidimensional clustering by moving from the lower dimensional space to higher. When it searches for dense units at the k-dimensional space, CLIQUE make use of information that is obtained from clustering at the (k–1) dimensional space to prune off unnecessary search.

CLIQUE automatically finds subspace of highest dimensionality such that high density clusters exist in those subspaces. It is intensive to the order of input tuples and does not presume any canonical distribution. It scales linearly with the size of input and has good scalability as the number of dimensions in the data is increased. However, the accuracy of clustering results may be degraded at the expense of the simplicity of the method.

The spatial clustering algorithms are categorized based on four factors like time complexity, input, handling of higher dimensional data, capability to detect irregularly shaped clusters. These factors were found to be necessary for effective clustering of large spatial datasets with high dimensionality. The previous nine algorithms are addressed by how they matched these four requirements. It can be seen that although several of the algorithms meet some of the requirements, and some meet most, none of the algorithms as originally proposed meet all the requirements. The hierarchical clustering methods are similar in performance but consume more time as compared to the other. The performance of partitional clustering methods like K-means and K-medoid algorithms are not well in handling irregularly shaped clusters. The density based methods and grid based methods are more suitable for handling spatial data, but when considering time complexity, grid based methods are more preferable.

CHECK YOUR PROGRESS

Q.4: Which of the following is a density based clustering algorithm

a) CLIQUE b) DBSCAN

c) K-Means d) None of the above

Q.5: Which of the following is not a partitional clustering algorithm?

a) K-means b) K-Medoids

c) AGNES d)

Q.6: STING is a _____.

a) Hierarchical Clustering Algorithm

b) Partitional Clustering Algorithm

c) Density Based Clustering Algorithm

d) Grid-Based Clustering Algorithm

Q.7: Which of the following is not a spatial data mining task?

a) Data summarization b) Generalization

c) Classification d) Association Rule Mining

15.10 LET US SUM UP

- Temporal Data Mining is an important extension to the conventional data mining techniques.
- In temporal data sets data are ordered with respect to time.
- Temporal data mining can be defined as the search for interesting correlations or patterns in large sets of temporal data.
- Data used in temporal data mining can be categorized as– Sequences, Time Stamped, Time Series, Fully Temporal data.

- Temporal data mining tasks can be categorized as Association rule mining, Prediction, Classification, Clustering, Characterization, Search & retrieval, Pattern discovery, Trend Analysis and Sequence Analysis.
- A temporal association rule is of the form $X \Rightarrow (T\ Y)$, which states that if X occurs then Y will occur within time T.
- A cyclic rule is one that occurs at regular time intervals
- If the temporal association rules with time intervals which follow some user-given calendar schema then it is called calendric association rules.
- Sequence mining is a data mining task specialized for analysing sequential data, to discover sequential patterns.
- Traditionally, sequence mining is being used to find sub-sequences that appear often in a sequence database.
- Spatial data mining is the application of data mining to spatial models.
- In spatial data mining, analysts use geographical or spatial information to produce business intelligence or other results.
- Spatial data mining is a non-trivial process to extract knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases.
- Spatial data mining tasks are the extension of conventional data mining tasks where spatial data and criteria are combined.
- Spatial data mining tasks are can be categorized as- Data Summarization, Generalisation, Classification, Clustering, Spatial Data Dependencies,
- Spatial clustering is a process of grouping a set of spatial objects into groups called clusters.
- Based on the technique adopted to define clusters, the clustering algorithms can be divided into four broad categories:
  - Hierarchical clustering methods (AGNES, BIRCH etc.),
  - Partitional clustering algorithms (K-means, K-medoids etc.),
  - Density-based clustering algorithms (DBSCAN DENCLUE etc.),
  - Grid based clustering algorithms (STING etc.).

15.11 FURTHER READING

1) Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.

2) Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

3) Tan, P. N. (2018). Introduction to Data Mining. Pearson Education India.

15.12 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: a) True; b) False; c) True; d) True

Ans. to Q. No. 2: (d) **Ans. to Q. No. 3:** (c)

Ans. to Q. No. 4: (b) **Ans. to Q. No. 5:** (c)

Ans. to Q. No. 6: (a) **Ans. to Q. No. 7:** (d)

15.13 MODEL QUESTIONS

Q.1: What is temporal data mining? Explain the different types of temporal data.

Q.2: Explain the different temporal data mining tasks.

Q.3: What is temporal association rule mining? What are the different types of temporal association rules? Explain.

Q.4: What is sequence mining? Explain with example.

Q.5: Explain the different spatial data mining task.

Q.6: What is spatial data mining? Explain briefly.

Q.7: What is spatial data clustering? What are the different types of Spatial data clustering? Explain in detail.

*** ***** ***

REFERENCES

1) Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.
2) Berson, A., & Smith, S. J. (1997). *Data Warehousing, Data Mining and OLAP*. McGraw-Hill, Inc.
3) Han, J., Pei, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
4) Inmon, W. H. (2005). Building the Data Warehouse. John Wiley & Sons.
5) Ponniah, P. (2004). *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals*. John Wiley & Sons.
6) Pudi, V., Krishna R. P. (2008) *Data Mining*. Oxford University Press.
7) Pujari, A. K. (2001). *Data Mining Techniques.* Universities Press.
8) Saxena, A., Saxena, K. Saxena, S. (2015). *Data Mining and Warehousing.* BPB Publications.
9) Tan, P. N. (2018). *Introduction to Data Mining*. Pearson Education India.
10) Tan, P. N., Steinbach, M. & Kumar, V. (2013). *Data Mining Cluster Analysis: Basic Concepts and Algorithms. Introduction to Data Mining*.
11) https://www.tutorialspoint.com/data_mining/dm_quick_guide.htm

BAOU
Education
for All

स्वाध्यायः परम तपः

યુનિવર્સિટી ગીત

સ્વાધ્યાયઃ પરમં તપઃ
સ્વાધ્યાયઃ પરમં તપઃ
સ્વાધ્યાયઃ પરમં તપઃ

શિક્ષણ, સંસ્કૃતિ, સદ્‌ભાવ, દિવ્યબોધનું ધામ
ડૉ. બાબાસાહેબ આંબેડકર ઓપન યુનિવર્સિટી નામ;
સૌને સૌની પાંખ મળે, ને સૌને સૌનું આભ,
દશે દિશામાં સ્મિત વહે હો દશે દિશે શુભ-લાભ.

અભણ રહી અજ્ઞાનના શાને, અંધકારને પીવો ?
કહે બુદ્ધ આંબેડકર કહે, તું થા તારો દીવો;
શારદીય અજવાળા પહોંચ્યાં ગુર્જર ગામે ગામ
ધ્રુવ તારકની જેમ ઝળહળે એકલવ્યની શાન.

સરસ્વતીના મયૂર તમારે ફળિયે આવી ગહેકે
અંધકારને હડસેલીને ઉજાસના ફૂલ મહેંકે;
બંધન નહીં કો સ્થાન સમયના જવું ન ઘરથી દૂર
ઘર આવી મા હરે શારદા દૈન્ય તિમિરના પૂર.

સંસ્કારોની સુગંધ મહેંકે, મન મંદિરને ધામે
સુખની ટપાલ પહોંચે સૌને પોતાને સરનામે;
સમાજ કેરે દરિયે હાંકી શિક્ષણ કેરું વહાણ,
આવો કરીયે આપણ સૌ
ભવ્ય રાષ્ટ્ર નિર્માણ...
દિવ્ય રાષ્ટ્ર નિર્માણ...
ભવ્ય રાષ્ટ્ર નિર્માણ

●

DR. BABASAHEB AMBEDKAR OPEN UNIVERSITY
(Established by Government of Gujarat)
'Jyotirmay' Parisar,
Sarkhej-Gandhinagar Highway, Chharodi, Ahmedabad-382 481
Website : www.baou.edu.in | Email : office.scs@baou.edu.in