UNIT II

INTRODUCTION TO DATA MINING

Data Mining

The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

Data Mining is the process of investigating hidden patterns of information to various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithm, helping decision making and other data requirement to eventually cost-cutting and generating revenue.

Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures.

Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events. Data Mining is also called Knowledge Discovery of Data (KDD).

The biggest challenge is to analyze the data to extract important information that can be used to solve a problem or for company development. There are many powerful instruments and techniques available to mine data and find better insight from it.


Business Goal
Present Data
Data Mining
Identify Data
Model & Evaluate Data
Prepare Data

Types of Data Mining

Data mining can be performed on the following types of data:

Relational Database:

A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables.

Tables convey and share information, which facilitates data searchability, reporting, and organization.

Data warehouses

A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights.

The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision- making for a business organization.

The data warehouse is designed for the analysis of data rather than transaction processing.

Data Repositories

The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.

Object-Relational Database

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc.

Transactional Database

A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately.

Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

Advantages of Data Mining

•The Data Mining technique enables organizations to obtain knowledge-based data.

•Data mining enables organizations to make lucrative modifications in operation and production.

•Compared with other statistical data applications, data mining is a cost-efficient.

•Data Mining helps the decision-making process of an organization.

•It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.

•It can be induced in the new system as well as the existing platforms.

•It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

Disadvantages of Data Mining

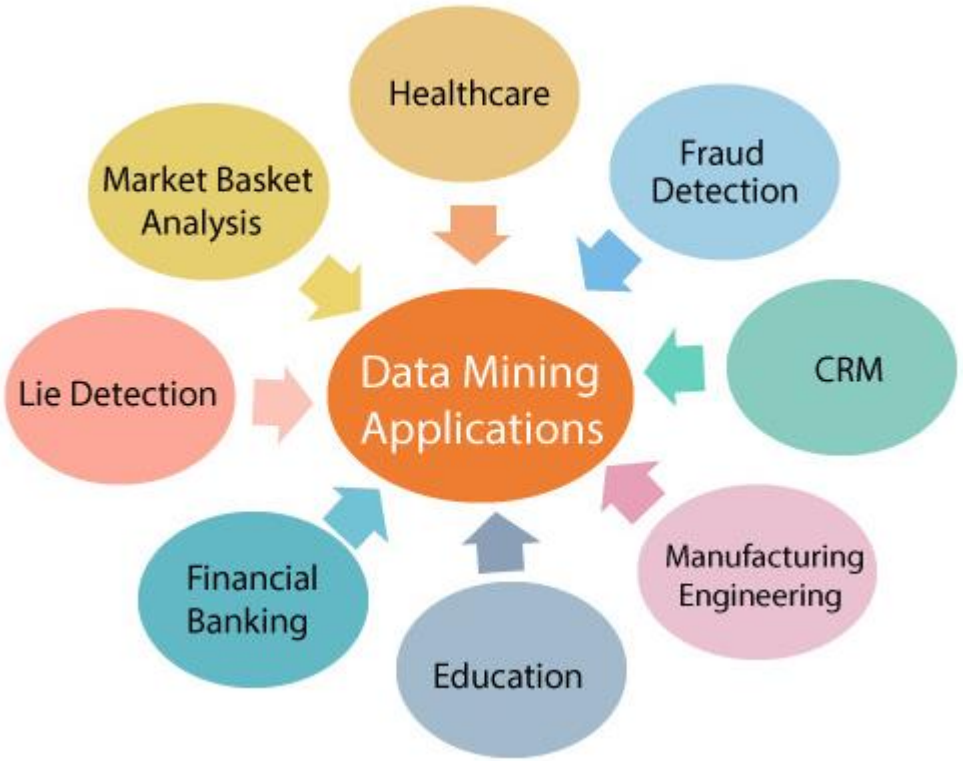
•There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.

•Many data mining analytics software is difficult to operate and needs advance training to work on.

•Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.

•The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

Data Mining Applications



What is a Data Repository?

A data repository often called a data archive or library, is a generic terminology that refers to a segmented data set used for reporting or analysis.

It’s a huge database infrastructure that gathers, manages, and stores varying data sets for analysis, distribution, and reporting.

What is a Shared Repository?

A shared repository is defined as a repository that can store revisions for multiple branches. Therefore, every branch will share one repository for its multiple revision storage.

Types of Data Repositories

Some common types of data repositories include:

Data Warehouse

- A data warehouse is a large central data repository that brings together data from several sources or business segments.
- The stored data is generally used for reporting and analysis to help users make critical business decisions.
- In a broader perspective, a data warehouse offers a consolidated view of either a physical or logical data repository gathered from numerous systems.
- The main objective of a data warehouse is to establish a connection between data from current systems. For example, product catalog data stored in one system and procurement orders for a client stored in another one.

Data Lake

A data lake is a unified data repository that allows you to store structured, semi-structured, and unstructured enterprise data at any scale.

Data can be in raw form and used for different tasks like reporting, visualizations, advanced analytics, and machine learning.

Data Mart

A data mart is a subject-oriented data repository that’s often a segregated section of a data warehouse. It holds a subset of data usually aligned with a specific business department, such as marketing, finance, or support.

Due to its smaller size, a data mart can fast-track business procedures as you can easily access relevant data within days instead of months. As it only includes the data relevant to a specific area, a data mart is an economical way to acquire actionable insights swiftly.

Metadata Repositories

Metadata incorporates information about the structures that include the actual data. Metadata repositories contain information about the data model that store and share this data.

They describe where the source of data is, how it was collected, and what it signifies. It may define the arrangement of any data or subject deposited in any format.

For businesses, metadata repositories are essential in helping people understand administrative changes, as they contain detailed information about the data.

Data Cubes

Data cubes are lists of data with multidimensions (usually 3 or more dimensions) stored as a table. They are used to describe the time sequence of an image's data and help assess gathered data from a range of standpoints.

Each dimension of a data cube signifies specific characteristics of the database such as day-to-day, monthly or annual sales.

The data contained within a data cube allows you to analyze all the information for almost any or all clients, sales representatives, products, and more.

Consequently, a data cube can help you identify trends and scrutinize business performance.

Why Do You Need A Data Repository?

A data repository can help businesses fast-track decision-making by offering a consolidated space to store data critical to your operations.

This segmentation enables easier data access and troubleshooting and streamlines reporting and analysis.

For instance, if you want to find out which of your workplaces incur the most cost, you can create an information repository for leases, energy expenses, amenities, security, and utilities, excluding employees or business function information.

Storing this data in one place can make it easier for you to come to a decision.

Clinical Data Repository

A clinical data repository (CDR) or clinical data warehouse (CDW) is defined as a real-time database that unifies data across multiple clinical sources to present a consolidated view of a patient’s details or records. Clinical data repository aids the clinic staff to access data for one patient instead of identifying a huge number of patients with similarities or common characteristics.

The common data types of clinical data repositories are as follows:

- Lab test results
- Patient information, such as demographics
- Discharge summaries
- Transfer dates
- Radiology images and reports
- Pathology reports

Challenges Associated with a Data Repository

Although an information repository offers a number of benefits, it also includes several challenges that you must manage efficiently to all possible data security risks.

Some challenges of maintaining data repositories include:

•An increase in data sets can reduce your system’s speed. To rectify this problem, make sure that the database management system can scale with data expansion.

•In case a system crashes, it can negatively impact your data. It’s best to maintain a backup of all the databases and restrict access to control the system risk.

•Unauthorized operators can access sensitive data more easily if stored in a single location than if it’s dispersed across numerous sources. On the contrary, implementing security protocols on a single data storage location is easier than on multiple ones.

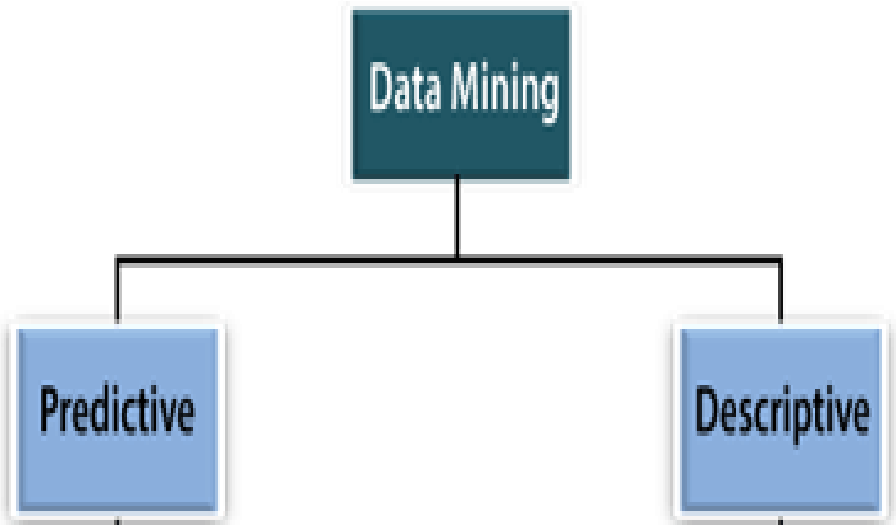
Data Mining Functionalities

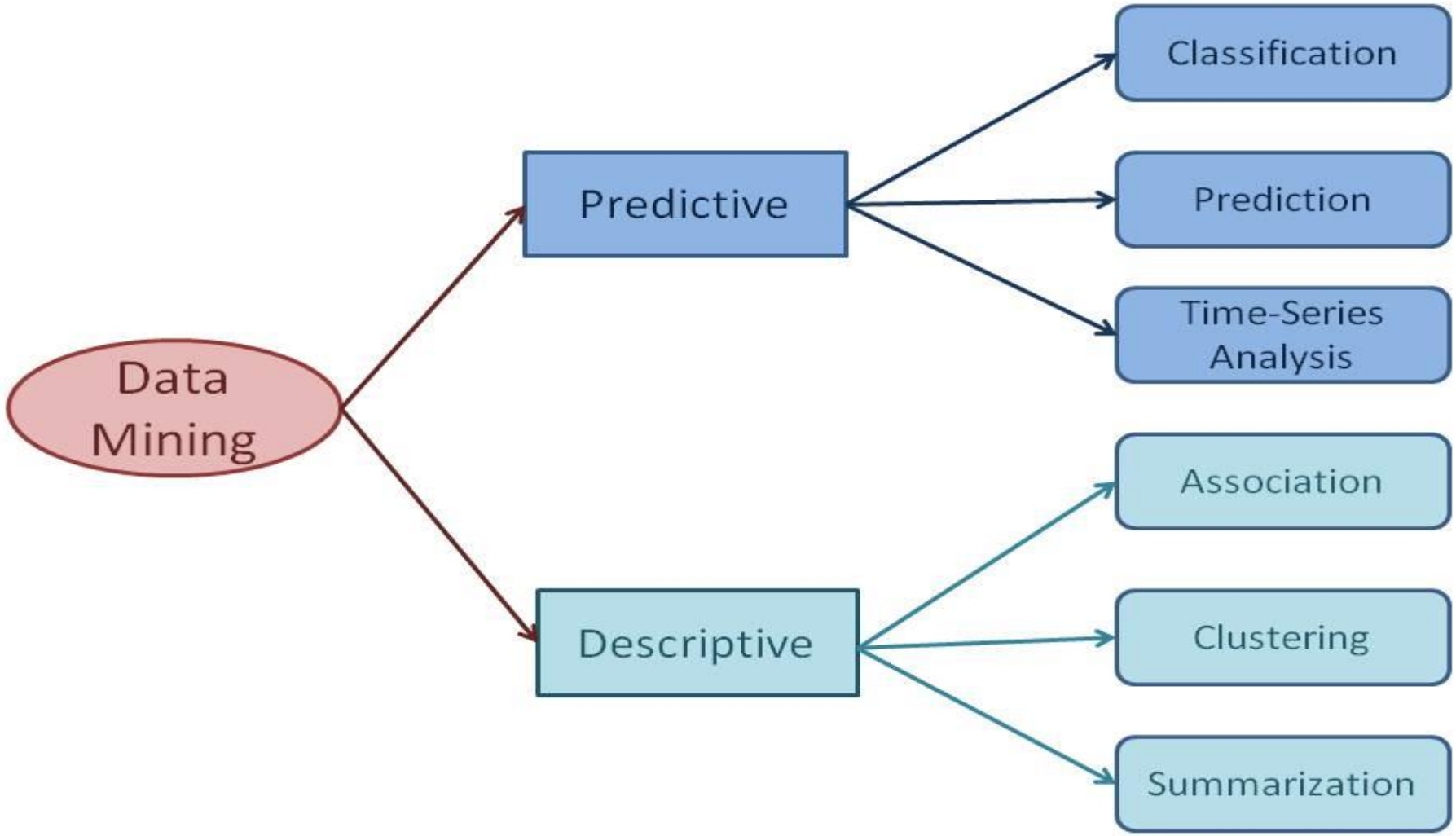
The Data Mining functionalities are basically used for specifying the different kind of patterns or trends that are usually seen in data mining tasks.

Data mining is extensively used in many areas or sectors. It is used to predict and characterize data. But the ultimate objective in **Data Mining Functionalities** is to observe the various trends in data mining.

The Data Mining tasks can be categorized into two kinds:

- Descriptive Data Mining
- Predictive Data Mining




Data Mining
Predictive
Descriptive
Classification
Prediction
Time-Series Analysis
Association
Clustering
Summarization

1.Descriptive Data Mining:

It includes certain knowledge to understand what is happening within the data without a previous idea. The common data features are highlighted in the data set.
For examples: count, average etc.

2.Predictive Data Mining:

It helps developers to provide unlabeled definitions of attributes. Based on previous tests, the software estimates the characteristics that are absent.
For example: Judging from the findings of a patient’s medical examinations that is he suffering from any particular disease.

Predictive data mining tasks come up with a model from the available data set that is helpful in predicting unknown or future values of another data set of interest.

A medical practitioner trying to diagnose a disease based on the medical test results of a patient can be considered as a predictive data mining task.

Descriptive data mining tasks usually finds data describing patterns and comes up with new, significant information from the available data set. A retailer trying to identify products that are purchased together can be considered as a descriptive data mining task.

a) Classification

Classification derives a model to determine the class of an object based on its attributes. A collection of records will be available, each record with a set of attributes. One of the attributes will be class attribute and the goal of classification task is assigning a class attribute to new set of records as accurately as possible.

Classification can be used in direct marketing, that is to reduce marketing costs by targeting a set of customers who are likely to buy a new product. Using the available data, it is possible to know which customers purchased similar products and who did not purchase in the past. Hence, {purchase, don’t purchase} decision forms the class attribute in this case. Once the class attribute is assigned, demographic and lifestyle information of customers who purchased similar products can be collected and promotion mails can be sent to them directly.

b) Prediction

Prediction task predicts the possible values of missing or future data. Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest.

For example, a model can predict the income of an employee based on education, experience and other demographic factors like place of stay, gender etc. Also prediction analysis is used in different areas including medical diagnosis, fraud detection etc.

c) Time - Series Analysis

Time series is a sequence of events where the next event is determined by one or more of the preceding events.

Time series reflects the process being measured and there are certain components that affect the behavior of a process.

Time series analysis includes methods to analyze time-series data in order to extract useful patterns, trends, rules and statistics. Stock market prediction is an important application of time-series analysis.

d) Association

Association discovers the association or connection among a set of items.

Association identifies the relationships between objects. Association analysis is used for commodity management, advertising, catalog design, direct marketing etc.

A retailer can identify the products that normally customers purchase together or even find the customers who respond to the promotion of same kind of products. If a retailer finds that beer and nappy are bought together mostly, he can put nappies on sale to promote the sale of beer.

e) Clustering

Clustering is used to identify data objects that are similar to one another. The similarity can be decided based on a number of factors like purchase behavior, responsiveness to certain actions, geographical locations and so on.

For example, an insurance company can cluster its customers based on age, residence, income etc. This group information will be helpful to understand the customers better and hence provide better customized services.

f) Summarization

Summarization is the generalization of data. A set of relevant data is summarized which result in a smaller set that gives aggregated information of the data.

For example, the shopping done by a customer can be summarized into total products, total spending, offers used, etc. Such high level summarized information can be useful for sales or customer relationship team for detailed customer and purchase behavior analysis.

Data can be summarized in different abstraction levels and from different angles.

INTERESTING PATTERN

Interestingness measures play an important role in data mining, regardless of the kind of patterns being mined.

These measures are intended 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.

MEASURES OF PATTERNS INTERESTINGNESS

There are subjective as well as objective measures of patterns interestingness as :

1] Objective Measures
- (a)Support Threshold
- (b)Confidence Threshold
- (c)Correlation Coefficient

2] Subjective Measures
- (a)User Belief and Expectations
- (b) User Need

Interestingness measures play an important role in data mining, regardless of the kind of patterns being mined.

These measures are intended 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.

MEASURES OF PATTERNS INTERESTINGNESS

There are subjective as well as objective measures of patterns interestingness as :

1] Objective Measures

- (a)Support Threshold
- (b)Confidence Threshold
- (c)Correlation Coefficient

2] Subjective Measures

- (a)User Belief and Expectations
- (b) User Need

[1] Objective Measures of Patterns Interestingness :

Objective Measures of Patterns Interestingness are based on statistics. These measures specify thresholds on statistical measures of rule interestingness, such as support, confidence and lift correlations.

(a)Support Threshold

Support represents the percentage of transactions from a database that the given rule X=>Y satisfies. This is taken to be the probability P (X U Y), where X U Y indicates that a transaction contains both the items X and Y that is the union of both X and Y. Formally support is defined by:

$$\text{support } X \Rightarrow Y = P(X \cup Y)$$

(b)Confidence Threshold

Confidence threshold assesses the degree of certainty of the discovered rule. This is taken to be the conditional probability of the rule. The probability that a transaction containing X also contains Y. It is defined as follows:

$$\text{Confidence}(X \Rightarrow Y) = P(Y/X)$$

(c)Correlation Coefficient

Coefficient of correlation is one of the most widely used statistical measures to measure the strength of relationships in two variables. Of the several mathematical methods of measuring correlation, the lift method is most widely used in practice. The coefficient of correlation is denoted by I as given below:

I(lift)= (X=> Y) =

If lift < 1, then item X and item Y appear less frequently together in the data than expected under the assumption of conditional independence. Item X and item Y are said to be negatively inter dependents.

If lift = 1, then item X and item Y appear as frequently together as expected under the assumption of conditional independence.

If lift > 1, then item X and item Y appear more frequently together in the data than expected under the assumption of conditional independence. Item X and item Y are said to positively inter dependent.

A study was carried out in which an algorithm was developed to mine frequent patterns. Objective measures were used to discover refined patterns from large sets of frequent patterns

[2] Subjective Measures of Patterns Interestingness :

Although objective interestingness measures facilitate identifying interesting patterns, they are ineffective unless combined with subjective measures that specify the need and interest of the user.

Patterns that are expected can be interesting if they confirm a hypothesis and belief that the user wished to validate.

A study was carried out in which a two step process was used. In the first step technically interesting patterns were mined and in second step business operable domain specific filtered patterns were extracted

(a)CONSTRAINTS BASED INTELLIGENT DATA MINING MECHANISM:

Constraints based Intelligent Data Mining Mechanism (IDMM) is being proposed to help the users to find relevant and valuable information.

The system consists of four modules: User Input Module, Dialog Management, Inference Engine and Data Repository.

The mechanism can be used in various applications such as e-commerce,education, farming applications etc.

In the present study the mechanism is assessed on the real world data set related to Socio-Economic conditions of Indian farmers.

Dataset is collected through questionnaire from 350 farmers located in villages near Meerut city.

The Mechanism can be used to guide various users associated with farming such as farmers, NGOs and government organization personnel working for the welfare of farmers and farming products.

The aim of the proposed mechanism is to find the most relevant information for the satisfaction of the user to improve the farmers' income and agricultural productivity.