Unit 1: Introduction

1.1 Motivation, Importance, Definition of Data Mining

Motivation and Improtance

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and imminent need for turning such data into information and knowledge. The information and. knowledge gained, can be used for applications ranging, from market analysis, fraud detection, and customer retention, to production control and science exploration.

- Data Mining is defined as the procedure of extracting information from huge sets of data.
- Data mining is mining knowledge from data.
- The terminologies involved in data mining and then gradually moves on to cover topics such as knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, and how to mine the Web.
- Here is a huge amount of data available in the Information Industry.
- This data is of no use until it is converted into useful information.
- It is necessary to analyse this huge amount of data and extract useful information from it.
- Extraction of information is not the only process we need to perform.
- Data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation.

Definition

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Major Sources of data: -

Business –Web, E-commerce, Transactions, Stocks - Science – Remote Sensing, Bio informatics, Scientific Simulation - Society and Everyone – News, Digital Cameras, You Tube * Need for turning data into knowledge – Drowning in data, but starving for knowledge.

1.2 Kind of Data, Data Mining Functionalities, Kinds of Patterns, Classification of Data Mining Systems, Data Mining Task Primitives

Various Types of Data

- **Structured Data**: This refers to data that is organized and formatted in a predefined manner, typically stored in relational databases. Structured data is represented in tables with rows and columns, where each column has a specific data type. Examples of structured data include sales records, customer information, financial data, and transaction logs.
- **Unstructured Data**: Unstructured data does not have a predefined format or organization. It includes textual data, such as emails, social media posts, customer reviews, and documents like PDFs or Word files. Other forms of unstructured data include multimedia files (images, videos, audio recordings), sensor data, and web pages.
- **Semi-Structured Data**: Semi-structured data lies between structured and unstructured data. It possesses some organizational structure but does not fit neatly into a traditional relational database. Examples of semi-structured data include XML files, JSON data, and log files.
- **Time-Series Data:** Time-series data is collected and recorded over regular time intervals. It represents data points or measurements taken at consecutive time points. Time-series data is commonly found in fields like finance (stock prices), weather (temperature recordings), energy consumption, and IoT (Internet of Things) sensor data.
- **Spatial Data:** Spatial data refers to data that has a geographic or spatial component. It includes coordinates, maps, satellite images, and GIS (Geographic Information System) data. Spatial data mining techniques are used to extract patterns and relationships related to geographical locations.
- **Text Data:** Textual data encompasses any form of written or typed text, including emails, documents, articles, social media posts, and online reviews. Text mining techniques are employed to extract valuable information from text data, such as sentiment analysis, topic modeling, and text classification.
- **Graph Data:** Graph data represents entities (nodes) and their relationships (edges) in a network. Examples of graph data include social networks, citation networks, web graphs, and knowledge graphs. Graph mining techniques help analyze the structure and properties of networks, identify key nodes, and uncover patterns within the graph.

Data mining Functionalities

Data mining consists of various functionalities and techniques that are used to extract valuable insights and knowledge from data. Some of the key functionalities of data mining include:

1. **Classification**: Classification is the process of categorizing data into predefined classes or categories based on the patterns and relationships discovered in the data. It involves building predictive models that can assign new, unseen data instances to the appropriate class. Classification is widely used for tasks such as customer segmentation, spam detection, disease diagnosis, and credit risk assessment.
2. **Clustering**: Clustering involves grouping similar data instances together based on their intrinsic similarities or patterns. It aims to discover natural groupings or clusters within the data without prior knowledge of the class labels. Clustering is useful for market segmentation, image recognition, anomaly detection, and recommendation systems.
3. **Regression**: Regression analysis is used to establish relationships between variables and predict numerical values. It helps in understanding the dependencies and correlations between different attributes in the data. Regression is commonly used for sales forecasting, price prediction, demand estimation, and trend analysis.
4. **Association Rule Mining:** Association rule mining discovers interesting relationships or associations between items in a dataset. It identifies patterns such as "if X, then Y," indicating that certain items or events tend to co-occur. Association rule mining is widely used in market basket analysis, recommendation systems, cross-selling, and product bundling.

?

5. **Anomaly Detection**: Anomaly detection aims to identify rare or unusual patterns in the data that deviate significantly from the norm. It helps in detecting outliers, fraud, network intrusions, and unusual behavior in various domains such as cybersecurity, finance, and healthcare.
6. **Text Mining:** Text mining involves extracting meaningful information from unstructured textual data. It includes tasks such as sentiment analysis, topic modeling, text classification, and entity recognition. Text mining is used for analyzing customer feedback, social media sentiment, document clustering, and content recommendation.
7. **Time Series Analysis:** Time series analysis focuses on analyzing data points collected over regular time intervals to identify patterns, trends, and seasonality. It helps in forecasting future values, detecting anomalies, and understanding temporal dependencies in the data. Time series analysis is commonly used in finance, weather forecasting, demand forecasting, and resource planning.
8. **Visualization and Exploration**: Data mining techniques are often coupled with data visualization and exploration tools to facilitate the interactive exploration and presentation of patterns and insights. Visualization techniques help in gaining a better understanding of the data, discovering hidden relationships, and communicating findings effectively.

These functionalities represent a range of techniques and methods that are employed in data mining to uncover patterns, relationships, and insights within data. Depending on the specific problem and data characteristics, different combinations of these functionalities are utilized to extract actionable knowledge from diverse datasets.

Note * Data mining functionalities—what kinds of patterns can be mined?

Kind of Pattern

In data mining, various types of patterns can be discovered depending on the nature of the data and the specific mining task. Here are some common types of patterns that can be found:

1. **Association Patterns:** Association patterns identify relationships or associations between items or events in a dataset. They indicate which items tend to co-occur or are frequently seen together. For example, in a retail setting, association patterns can reveal that customers who buy bread often purchase butter as well.
2. **Sequential Patterns:** Sequential patterns capture temporal relationships in sequential data, where the order of events or items matters. They uncover patterns of events that frequently occur in a particular order. This type of pattern is commonly used in analyzing sequences of customer transactions, web clickstreams, or sensor data.
3. **Classification Patterns:** Classification patterns are derived from supervised learning algorithms and involve categorizing data instances into predefined classes or categories. These patterns establish rules or models that can predict the class of new, unseen data instances based on their attributes. For example, classifying emails as spam or non-spam based on their content and characteristics.
4. **Clustering Patterns:** Clustering patterns group similar data instances together based on their intrinsic similarities. These patterns identify natural clusters or groups within the data without any predefined class labels. Clustering can reveal segments or clusters of customers, documents, or other entities that share similar characteristics.
5. **Regression Patterns:** Regression patterns involve predicting numerical values or establishing relationships between variables in the data. They help in understanding the dependencies and correlations between different attributes and can be used for forecasting or estimating continuous values. For example, predicting house prices based on various features like location, size, and number of rooms.
6. **Deviation Patterns:** Deviation patterns detect anomalies or unusual instances in the data that deviate significantly from the norm. These patterns highlight data points that are rare, unexpected, or abnormal compared to the majority of the data. Anomaly detection is useful in fraud detection, network intrusion detection, or identifying outliers in healthcare data.
7. **Text Patterns:** Text patterns refer to patterns discovered in textual data, such as documents, articles, or social media posts. Text mining techniques can uncover patterns like sentiment analysis (identifying positive or negative sentiment), topic modeling (extracting key themes or topics), or named entity recognition (identifying named entities like people, organizations, or locations).

Classification of Data Mining Systems

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science. Moreover, depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or highperformance computing.

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

- **Classification according to the kinds of databases mined:** A data mining system can be classified according to the kinds of databases mined. Database systems can be classified according to different criteria (such as data models, or the types of data or applications involved), each of which may require its own data mining technique.
- **Classification according to the kinds of knowledge mined:** Data mining systems can be categorized according to the kinds of knowledge they mine, that is, based on data mining functionalities, such as characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis.
- **Classification according to the kinds of techniques utilized:** Data mining systems can be categorized according to the underlying data mining techniques employed. These techniques can be described according to the degree of user interaction involved (e.g., autonomous systems, interactive exploratory systems, query-driven systems) or the methods of data analysis employed (e.g., database-oriented or data warehouse– oriented techniques, machine learning, statistics, visualization, pattern recognition, neural networks, and so on).
- **Classification according to the applications adapted:** Data mining systems can also be categorized according to the applications they adapt. For example, data mining systems may be tailored specifically for finance, telecommunications, DNA, stock markets, e-mail, and so on. Different applications often require the integration of application-specific methods.

Data mining task primitives:

A data mining query is defined in terms of the following primitives:

Task-relevant data: This is the database portion to be investigated. For example, suppose that you are a manager of All Electronics in charge of sales in the United States and Canada. In particular, you would like to study the buying trends of customers in Canada. Rather than mining on the entire database. These are referred to as relevant attributes.

The kinds of knowledge to be mined: This specifies the data mining functions to be performed, such as characterization, discrimination, association, classification, clustering, or evolution analysis. For instance, if studying the buying habits of customers in Canada.

Background knowledge: Users can specify background knowledge, or knowledge about the domain to be mined. This knowledge is useful for guiding the knowledge discovery process, and for evaluating the patterns found. There are several kinds of background knowledge.

Interestingness measures: These functions are used to separate uninteresting patterns from knowledge. 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.

Presentation and visualization of discovered patterns: This refers to the form in which discovered patterns are to be displayed. Users can choose from different forms for knowledge presentation, such as rules, tables, charts, graphs, decision trees, and cubes.

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

Integration of a data mining system with a database or data warehouse system is essential for efficient and effective data analysis. By integrating these systems, you can leverage the power of data mining algorithms to discover hidden patterns, extract valuable insights, and make data-driven decisions. Here's an **overview of the integration process:**

- **Understand the data mining system:** Familiarize yourself with the data mining system you plan to integrate. Determine its capabilities, supported algorithms, data preprocessing requirements, and output formats. This knowledge will help you design the integration effectively.
- **Define integration objectives:** Clearly define your objectives for integrating the data mining system. Understand the specific tasks you want to accomplish, such as predictive modeling, clustering, or association rule mining. This will guide you in selecting the appropriate algorithms and integration approach.
- **Identify data sources:** Identify the data sources you want to analyze using the data mining system. These sources can include databases, data warehouses, or other data repositories. Determine the required data attributes, formats, and quality for effective data mining.
- **Data preprocessing:** Data preprocessing is often a crucial step before data mining. It involves cleaning, transforming, and aggregating the data to prepare it for analysis. Depending on the data mining system, you may need to perform preprocessing tasks within the data mining system itself or use external tools to preprocess the data before integration.
- **Extract data from the database/data warehouse:** Develop processes or scripts to extract the required data from the database or data warehouse. Depending on the integration approach, you can use database query languages (e.g., SQL) or specialized tools to retrieve the data. Consider factors like data volume, extraction frequency, and security requirements.
- **Data transformation and integration:** Transform the extracted data into a format suitable for the data mining system. This may involve converting data types, handling missing values, and ensuring compatibility with the system's input requirements. Use integration techniques such as data mapping, merging, or concatenation to combine data from multiple sources, if needed.
- **Load data into the data mining system:** Load the transformed data into the data mining system for analysis. The process may vary depending on the system. Some systems have built-in data loading capabilities, while others may require importing data in specific file formats or through API integrations.
- **Perform data mining tasks:** Utilize the data mining algorithms and functionalities offered by the system to analyze the loaded data. Apply techniques like classification, regression, clustering, or association rule mining to extract meaningful patterns and insights from the data.
- **Interpret and visualize results:** Analyze the results generated by the data mining system and interpret the findings. Use visualization techniques such as charts, graphs, or dashboards to present the results in a clear and understandable manner. This step is crucial for effective communication and decision-making based on the insights obtained.
- **Iterative process:** Data mining is an iterative process, so it's important to refine your approach based on the results and insights gained. Make adjustments to the data extraction, preprocessing, or analysis steps as necessary to improve the accuracy and relevance of your results.

1.4 Major Issues in Data Mining, Types of Data Sets and, Attribute Values, Basic Statistical Descriptions of Data, Data Visualization, Measuring Data Similarity

Major issues in data mining:

Mining different kinds of knowledge in databases. - The need of different users is not the same. And Different user may be in interested in different kind of knowledge. Therefore it is necessary for data mining to cover 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 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 level 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, visual representations. This representations should be easily understandable by the users.

Handling noisy or incomplete data. - The data cleaning methods are required that can handle the noise, incomplete objects while mining the data regularities. If data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

Pattern evaluation. - It refers to interestingness of the problem. The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

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 algorithm divide the data into partitions which is further processed parallel. Then the results from the partitions is merged. The incremental algorithms, updates databases without having mine the data again from scratch.

Types of Data Sets:

- **Cross-Sectional Data:** This type of data is collected from a single point in time, capturing information about multiple individuals, objects, or subjects at that specific moment. For example, a survey conducted on a specific day or a snapshot of a database.
- **Time Series Data:** Time series data consists of observations taken at successive points in time, typically at regular intervals. This data type is commonly used in analyzing trends, seasonality, and forecasting future values. Examples include stock prices, weather data, or website traffic over time.
- **Longitudinal Data:** Longitudinal data involves repeated observations of the same individual, subject, or entity over an extended period. It allows the analysis of changes and trends over time. Examples include medical patient records or longitudinal studies on educational performance.
- **Spatial Data:** Spatial data represents information associated with specific geographic locations. This type of data is commonly used in geographic information systems (GIS) and can include maps, satellite images, or location-based data like GPS coordinates.
- **Textual Data:** Textual data includes unstructured or semi-structured text information, such as documents, emails, social media posts, and web pages. Analyzing textual data involves natural language processing (NLP) techniques to extract meaningful insights.
- **Categorical Data:** Categorical data represents qualitative characteristics with distinct categories or labels. Examples include gender (male/female), product types, or educational levels.
- **Numerical Data:** Numerical data represents quantitative values and can be further categorized into discrete and continuous data. Discrete data consists of distinct, separate values (e.g., counting the number of students in a class), while continuous data represents measurements that can take any value within a range (e.g., temperature, height, weight).

Attribute Values:

Attribute values are the specific values associated with each attribute or variable in a data set. These values depend on the data type of the attribute:

- For categorical data, attribute values will be the distinct categories or labels (e.g., "Male" or "Female").
- For numerical data, attribute values will be the specific measurements or numbers (e.g., "25.5" for temperature).

Basic Statistical Descriptions of Data: Basic statistical descriptions of data provide a summary of the main characteristics of a dataset. Common statistical measures include:

- **Mean**: The average value of a set of numerical data points.
- **Median**: The middle value in a sorted list of numerical data points. It is less affected by extreme values compared to the mean.
- **Mode**: The value that appears most frequently in a dataset.
- **Standard Deviation**: A measure of the dispersion or spread of data points around the mean.
- **Range**: The difference between the maximum and minimum values in a dataset.
- **Quartiles**: Values that divide the dataset into four equal parts, representing 25th, 50th (median), and 75th percentiles.

Data Visualization:

Data visualization is the graphical representation of data to better understand patterns, trends, and relationships within the dataset. Common data visualization techniques include:

- **Bar charts:** Used to compare categorical data by displaying bars of different heights.
- **Line charts**: Suitable for displaying trends in time series data.
- **Scatter plots:** Used to visualize relationships between two numerical variables.
- **Histograms**: Represent the distribution of numerical data by dividing it into bins.
- **Pie charts:** Used to show the proportion of each category in a dataset.
- **Heatmaps**: Visualize data in a grid format using color to represent values.

Measuring Data Similarity for Data Mining:

Data similarity measures are used to quantify the similarity or dissimilarity between data points or objects. Common similarity measures include:

- **Euclidean Distance:** Calculates the straight-line distance between two data points in a multi-dimensional space. It is commonly used for numerical data.
- **Manhattan Distance:** Also known as the city block distance or L1 norm, it measures the sum of absolute differences between corresponding attribute values of two data points. It is useful for numerical data and categorical data with ordinal relationships.
- **Cosine Similarity:** Measures the cosine of the angle between two vectors, representing data points. It is often used for text analysis or high-dimensional data where the magnitude of the vectors is important.
- **Jaccard Similarity:** Used for comparing sets of binary or categorical data. It calculates the ratio of the size of the intersection to the size of the union between two sets.
- **Hamming Distance:** Measures the number of positions at which two strings of equal length differ. It is commonly used for binary or categorical data.
- **Pearson Correlation Coefficient:** Measures the linear correlation between two numerical variables. It ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation).
- **Edit Distance:** Calculates the minimum number of operations (insertions, deletions, or substitutions) required to transform one string into another. It is used for text analysis or comparing sequences.
- **Minkowski Distance:** A generalized distance measure that includes both Euclidean distance and Manhattan distance as special cases. It can be adjusted using a parameter, typically denoted as p, to control the distance calculation.

1.5 PREPROCESSING: Data Quality, Major Tasks in Data Preprocessing, Data Reduction, Data Transformation, Data Discretization, Data Cleaning, and Data Integration.

Data preprocessing is a crucial step in data mining that involves preparing and cleaning raw data to ensure its quality, consistency, and suitability for analysis. Here are the major tasks involved in data preprocessing:

- **Data Quality Assessment**: Assessing data quality involves identifying and handling issues such as missing values, outliers, inconsistent formatting, and erroneous data entries. It is essential to ensure that the data is accurate and reliable.
- **Data Cleaning:** Data cleaning involves removing or correcting errors, inconsistencies, or noise in the dataset. This includes handling missing data (e.g., imputation techniques), removing duplicate records, and resolving inconsistencies in attribute values.
- **Data Integration:** Data integration combines data from multiple sources into a unified dataset. It involves resolving schema and attribute conflicts, harmonizing data formats, and combining relevant information from different sources.
- **Data Transformation:** Data transformation involves converting the data into a suitable format for analysis. This includes normalization to bring data within a specific range, logarithmic or power transformations to handle skewed distributions, and standardization to give variables equal weight.
- **Data Reduction:** Data reduction techniques aim to reduce the size or dimensionality of the dataset while preserving its important characteristics. This helps to improve computational efficiency and reduce noise. Techniques include feature selection (choosing relevant attributes), feature extraction (creating new features from existing ones), and instance sampling (selecting a representative subset of the data).
- **Data Discretization:** Data discretization involves converting continuous attributes into discrete intervals or categories. This is useful for handling numerical attributes or reducing the complexity of the data. Discretization methods include equal-width/binning, equal-frequency/binning, and clustering-based discretization.
- **Data Normalization:** Data normalization transforms the data to a common scale, enabling fair comparisons between different attributes. Common normalization techniques include min-max scaling, z-score normalization, and decimal scaling.
- **Data Integration:** Data integration combines data from different sources or databases into a single, consistent dataset. It involves resolving conflicts, standardizing attribute formats, and ensuring data integrity.

These tasks in data preprocessing help to ensure that the data used for analysis is of high quality, consistent, and appropriately prepared for subsequent data mining tasks. Proper preprocessing can significantly impact the accuracy and effectiveness of the data mining process.