B.C.A. (Sem – VI)

B.C.A. - 603

Data Warehousing & Data Mining

Purushottam Singh

Unit:- 3

Definition:-

Data mining is also called **knowledge discovery and data base (KDD),** is the process of automatically searching large volumes of data for patterns.

"Data mining is the process of discovering interesting knowledge, such as patterns, associations, changes, and structure from large amount of data stored in database, data warehouse or other information repositories".

Data mining is the exploration and analysis of large quantities of data, inorder to discover to valid, novel, potentially useful and ultimately understandable patterns in data.

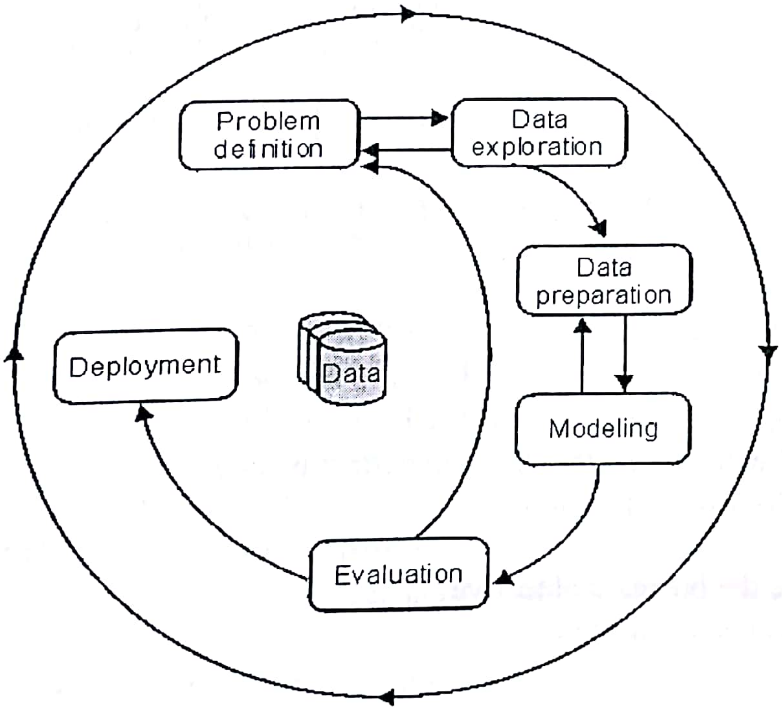
Valid-: The pattern holds in General.

Novel-: we did not know the pattern beforehand.

Useful:-we can device action from the patterns.

Understandable:-we can interpret and comprehend the patterns.

Data mining as process:-



1. Problem definition:-

- A data mining project starts with the understanding of the business problem.
- Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective.
- The project objective is then translated into a data mining problem definition.
- In the problem definition phase, data mining tools are not yet required.

2. Data exploration:-

- Domain experts understand the meaning of the metadata.
- They collect, describe, and explore the data. They also identify quality problems of the data.
- A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital.
- In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

3. Data preparation:-

- Domain experts build the data model for the modeling process.
- They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format.
- They also create new derived attributes, for example, an average value.
- Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

4. Modeling:-

- Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem.
- Some of the mining functions require specific data types. The data mining experts must assess each model.
- In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required.
- When the final modeling phase is completed, a model of high quality has been built.

5. Evaluation:-

- Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved.
- When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions:

  - Does the model achieve the business objective?
  - Have all business issues been considered?

6. Deployment:-

- Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets.

Data mining Functionality:-

- Descriptive
- Classification and Prediction

Descriptive

The descriptive function deals with 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

Class/Concept Description:-

- ➢ Class/Concepts refer the data to be associated with classes or concepts.
- ➢ For example, in a company classes of items for sale 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 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 mapping or classification of a class with some predefined group or class.

Mining of Frequent Patterns:-

- ➢ Frequent patterns are those patterns that occur frequently in transactional data. Here is the list of kind of frequent patterns:
- **Frequent Item Set** - It refers to 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 itemsets or subsequences.

Mining of Association:-

- ➢ Associations are used in retail sales to identify patterns that are frequently purchased together.
- ➢ This process refers to process of uncovering the relationship among data and determining association rules.
- ➢ For example A retailer generates association rule that show that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

Mining of Correlations

- ➢ It is 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.

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.

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 analysis of set of training data. The derived model can be presented in the following forms:
- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

Here is the list of functions involved in this:

- **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 analysis set of training data i.e the data object whose class label is well known.
- **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.
- **Outlier Analysis** - The Outliers may be defined as the data objects that do not comply with general behavior or model of the data available.
- **Evolution Analysis** - Evolution Analysis refers to description and model regularities or trends for objects whose behavior changes over time.

Difference bent KDD V/S Data Mining:

- o KDD is the whole process of discovering.
- o Data mining is the intelligent technical step in that process.
- o In KDD is a small Database.
- o In Data mining is a large Database.

- KDD process which deals with identifying patterns in data.
- Data mining is only the application of a specific algorithm based on the overall goal.
- KDD is overall process of extracting knowledge from data.
- Data mining is a step inside the KDD process.
- It is automated discovery of patterns and relationship.
- It is search the discovery of patterns and relationship.

Data mining Techniques:-

1. Classification:-

- Classification is a process that seeks to group together objects with similar attributes.
- When classifying a new object, its attributes are compared with the attributes of objects in existing groups, and the object is assigned to the group whose attributes are most similar.
- Data mining techniques that support classification are decision trees and nearest neighbor techniques.

Two types of Classification:-

1. **Supervised Classification**

   The set of possible classes is known in advance.

2. **Unsupervised Classification**

   Set of possible classes is not known. After classification we can try to assign name to that class. Unsupervised classification is called clustering.

2. Association Detection:-

- Association Discovery is the task of determining which items in a set belong together based upon the frequency of their occurrence.
- Reveals the degree to which variables are related and the nature and frequency of this relationship in the information.

  - **Market basket analysis:-** analyzes such items as web sites and checkout scanner information to detect customers buying behavior and predict future behavior by identifying affinities among customers choice of products and services.
  - **Statistical analysis:-**perform such function as information correlations. Distributions, calculations, and variance analysis.

3. Cluster Detection:-

- Clustering Analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.

- Data points in separate clusters are less similar to one another.

- **Document Clustering:**
  - **Goal:** To find groups of documents that are similar to each other based on the important terms appearing in them.
  - **Approach:** To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - **Gain:** Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

4. Automatic Cluster Detection:-

- Automatic cluster detection is described as a tool for undirected knowledge discovery.
- Automatic cluster detection is an undirected data mining technique that can be used to learn about the structure of complex databases.
- The automatic cluster detection algorithms themselves are simply finding structure that exists in the data without regard to any particular target variable.
- Automatic cluster detection is a data mining technique that is rarely used in isolation because finding clusters is not often an end in itself.
- In clustering, there is no preclassified data and no distinction between independent and dependent variables.
- One of the most popular algorithms for automatic cluster detection is K-means. The K-means algorithm is an iterative approach to finding K clusters based on distance.

5. Sequential Pattern Detection:-

- Sequence detection is similar to association discover but attempts to identify commonly occurring sequences of events over time.
- Finds patterns between events such that the presence of one set of items is followed by another set of items in a database of events over a period of time.
  - e.g. Used to understand long term customer buying behaviour.

- Finds links between two sets of data that are time-dependent, and is based on the degree of similarity between the patterns that both time series demonstrate.
  - E.g. within three months of buying property, new home owners will purchase goods such as cookers, freezers, and washing machines.

Model Selection:-

- A sequence is a list of item sets that tends to occur in a predictable order.
- Sequence modeling detects frequent sequences and generates a model that can be used to make predictions.

Testing Model Design:-

- The sequence detection model for this research is based upon the CARMA Algorithm proposed by Christian Hidber in 1999.

Build Model:-

- Building the model involves determining model parameters and optimizing the model performance or results.

6. Similar Time Sequence Detection:-

- Time Series Database introduces new aspects and challenges to the tasks of data mining and knowledge discovery.
- These new challenges include: finding the most efficient representation of time series data, measuring similarity of time series, detecting change points in time series, and time series classification and clustering.
- Some of these problems have been treated in the past by experts in time series analysis.
- statistical methods of time series analysis are focused on sequences of values representing a single numeric variable (e.g., price of a specific stock).
- In a real-world database, a time-stamped record may include several numerical and nominal attributes, which may depend not only on the time dimension but also on each other.
- Efficient and effective representation of time series is a key to successful discovery of time-related patterns.

Data Mining Applications:-

Here is the list of areas where data mining is widely used:

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates the systematic data analysis and data mining. Here are the few typical cases:

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of increasing ease, availability and popularity of web.

The Data Mining in Retail Industry helps in identifying customer buying patterns and trends. That leads to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in retail industry:

- Design and Construction of data warehouses based on benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

Telecommunication Industry

Today the Telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, Internet messenger, images, e-mail, web data transmission etc.Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Data Mining in Telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list examples for which data mining improve telecommunication services:

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

Biological Data Analysis

Now a days we see that there is vast growth in field of biology such as genomics, proteomics, functional Genomics and biomedical research.Biological data mining is very important part of Bioinformatics. Following are the aspects in which Data mining contribute for biological data analysis:

- Semantic integration of heterogeneous , distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy etc. There is large amount of data sets being generated because of the fast numerical simulations in various fields such as climate, and ecosystem modeling, chemical engineering, fluid dynamics etc. Following are the applications of data mining in field of Scientific Applications:

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

Intrusion Detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or availability of network resources. In this world of connectivity security has become the major issue. With increased usage of internet and availability of tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection:

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

Difference betn Discovery v/s Verification Mode:-

- Discovery can be very efficient.
- Verification mode can be very inefficient.
- Discovery it's a more information created in this process.
- Very little information is created in this process.
- Discovery cost is effective.
- Verification mode cost is in effective.
- The Discovery is divided into two mode Descriptive & Predictive.
- Verification modc involves Queries, Multidimensional analysis, Visualization.
- The large number of useful about data in Shortest amount of time.
- The large number of useful about data in largest amount of time.
- In Discovery increase modeling Accuracy.
- Verification Mode increases modeling Inaccuracy.