
RGPVNOTES.IN

Program : **B.E**

Subject Name: **Data Mining**

Subject Code: **CS-8003**

Semester: **8th**

UNIT-III Introduction to Data Mining, Knowledge Discovery, Data Mining Functionalities, Data Mining System categorization and its Issues. Data Processing :- Data Cleaning, Data Integration and Transformation. Data Reduction, Data Mining Statistics. Guidelines for Successful Data Mining.

**Introduction to Data Mining**

Data Mining(DM) is processing data to identify patterns and establish reltionships. DM is the process of analyzing data from different perspectives and summearizing it into useful information. This information can be used in decision making. DM is the extraction of hidden predictive information from large amounts of data stored in the data warehouse for useful information, using technology with great potential to help companies focus on the most important information.

Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.



“Data mining is the identification or extraction of relationships and patterns from data using computational algorithms to reduce, nodel, understand, or analyze data.”

**Knowledge Discovery**

KDD refers to the overall process of discovering useful knowledge from data also called knowledge discovery process (KDP). KD concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concern support for learing and analyzing the application domain. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.

Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

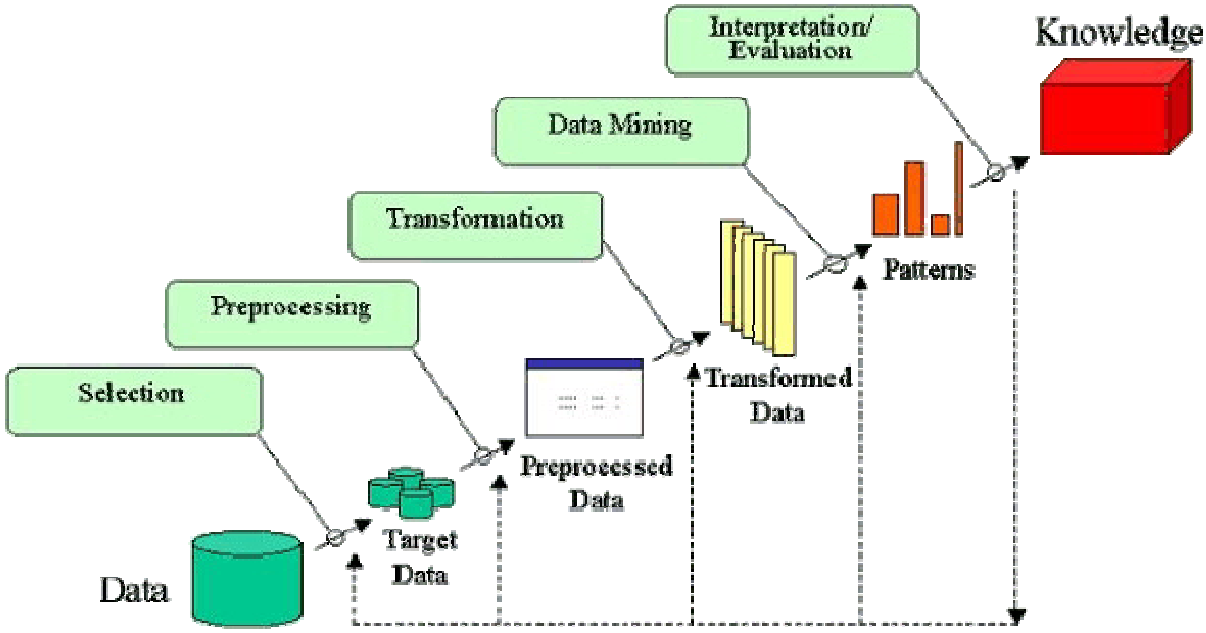


**Figure 3.1**: KDD Process

Here is the list of steps involved in the knowledge discovery process −

- Data Cleaning − In this step, the noise and inconsistent data is removed.
- Data Integration − In this step, multiple data sources are combined.
- Data Selection − In this step, data relevant to the analysis task are retrieved from the database.
- Data Transformation − In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- Data Mining − In this step, intelligent methods are applied in order to extract data patterns.
- Pattern Evaluation − In this step, data patterns are evaluated.
- Knowledge Presentation − In this step, knowledge is represented.

**Data Mining Functionalities**

Different kind of patterns can be discovered depending on the data mining task in use. There are mainly two types of data mining tasks :

1. Descriptive Data Mining Tasks
2. Predictive Data Mining Tasks

Descriptive mining tasks characterize the common properties of the existing data. Predictive mining tasks perform inference on the existing data in order to make predictions.

- **Concept/Class Description: Characterization and Discrimination**

Data can be associated with classes or concepts. For example, A grocery store manager may want to characterize the customer products whose sale increased by 15%in the last month. We can collect such data using SQL query in database

Discrimination: Data discrimination produces discrimination rules; this is comparison of common features of object between two classes referred as target class and contrasting class.

Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a database query the output of data characterization can be presented in various forms. Examples include pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables, including crosstabs.

Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. The target and contrasting classes can be specified by the user, and the corresponding data objects retrieved through database queries.

- **Mining Frequent Patterns, Associations, and Correlations**



Frequent patterns, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including itemsets, subsequences, and substructures.

A frequent itemset typically refers to a set of items that frequently appear together in a transactional data set, such as Computer and Software. A frequently occurring subsequence, such as the pattern that customers tend to purchase first a PC, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern.

Example: Suppose, as a marketing manager of All Electronics, you would like to determine which items are frequently purchased together within the same transactions.

An example of such a rule, mined from the All Electronics transactional database, is

buys(X;—computer) buys(X; —software) [support = 1%, confidence = 50%]

where X is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% support means that 1% of all of the transactions under analysis showed that computer and software were

 

purchased together. This association rule involves a single attribute or predicate (i.e., buys) that repeats. Association rules that contain a single predicate are referred to as single-dimensional association rules. Dropping the predicate notation, the above rule can be written simply as ―compute software [1%, 50%]

- **Classification and Prediction**

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

"How is the derived model presented?" The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks

A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units. There are many other methods for constructing classification models, such as naïve



Bayesian classification, support vector machines, and k-nearest neighbor classification. Whereas classification predicts categorical (discrete, unordered) labels, prediction models Continuous-valued functions. That is, it is used to predict missing or unavailable numerical data values rather than class labels. Although the term prediction may refer to both numeric prediction and class label prediction,

- **Cluster Analysis**

Clustering is the data mining techniques used to place data elements into related groups without advance knowledge of the group definitions.

- **Outlier Analysis**

Data object that does not comply with the general behavior of the data. It is useful in fraud detection rare event analysis. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. Based on functionality it can be classified into two categories:

i) Descriptive Mining
ii) Predictive Mining

 

- **Evolution Analysis**

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of time related data, distinct features of such an analysis include time-series data analysis, Sequence or periodicity pattern matching, and similarity-based data analysis.

**Data Mining System categorization and its Issues.**

There is a large variety of data mining systems available. Data mining systems may integrate techniques from the following −

- Spatial Data Analysis
- Information Retrieval
- Pattern Recognition
- Image Analysis
- Signal Processing
- Computer Graphics
- Web Technology
- Business
- Bioinformatics



**Data Mining System Classification**

A data mining system can be classified according to the following criteria −

- Database Technology
- Statistics
- Machine Learning
- Information Science
- Visualization
- Other Disciplines

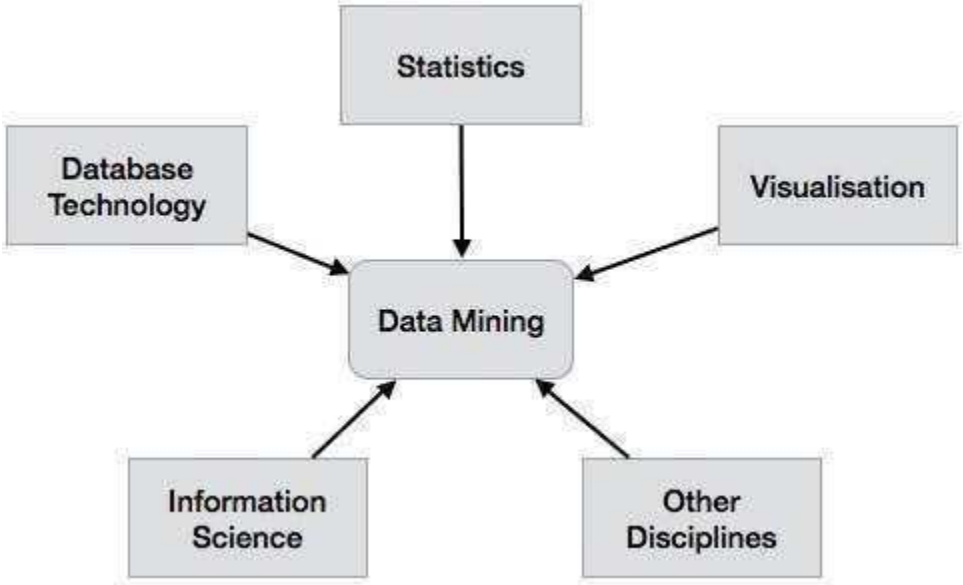


**Figure 3.2:** Data Mining System Classification

Data mining systems depend on databases to supply the raw input and this raises problems, such as the databases tend to be dynamic, incomplete, dynamic, noisy and large. Other problems arise as a result of the inadequacy and irrelevance of the information stored. The difficulties in data mining can be categorized as



a) Limited information
b) Noise or missing data
c) User interaction and prior knowledge
d) Uncertainty
e) Size, updates and irrelevant fields

**Data Processing**

1. Real world data are generally

Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data

Noisy: containing errors or outliers

Inconsistent: containing discrepancies in codes or names

2. Tasks in data preprocessing

 

Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

Data integration: using multiple databases, data cubes, or files.

Data transformation: normalization and aggregation.

Data reduction: reducing the volume but producing the same or similar analytical results.

Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

**Data Cleaning**

Data cleaning is a technique deal with detecting and removing inconsistencies and error from the data in-order to get better quality data. Data cleaning is performed as a data preprocessing step while preparing the data for a data warehouse. good quality data requires passing a set of quality criteria. Those criteria include: Accuracy, Integrity, Completeness, validity, consistency, uniformity, Density and uniqueness.

**Data Integration**



Data Integration is a data preprocessing technique that takes data from one or more sources and mapping it, field by field onto a new data structure. Idea is to merge the data from multiple sources into a coherent data store. Data may be distributed over different databases or data warehouses. There may be necessity of enhancement of data with additional (external) data. Issues like entity identification problem.

**Data Transformation**

In data transformation data are consolidated into appropriate form to make suitable for mining, by performing summary or aggregation operations. Data transformation involves following

- Data Smoothing
- Data aggregation
- Data Generalization
- Normalization
- Attribute Construction

**Data Reduction:**

If the data set is quite huge then the task of data mining and analysis can take much longer time, making the whole exercise of analysis useless and infeasible. Data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts.

The data reduction strategies include:

- Data cube aggregation
- Dimensionality reduction
- Data discretization and concept hierarchy generation
- Attribute Subset Selection

**Data Mining Statistics**

Statistics is only about quantifying data. While it uses tools to find relevant properties of data, it is a lot like math. It provides the tools necessary for data mining. For most of the data preprocessing tasks, we would like to learn about data characteristics regarding both central tendency and dispersion of the data. Measures of central tendency include mean, median mode and midrange, while measures of data dispersion include quartiles, interquartile range (IQR) and variance.

Data mining is thus a confluence of various other frontiers or fields like statistics, artificial intelligence, machine learning, database management, pattern recognition, and data visualization.

**Guidelines for Successful Data Mining**

- Anomaly or Outlier Detection
- Association Rule Learning
- Clustering Analysis
- Classification Analysis
- Regression Analysis
- Choice Modeling
- Rule Induction
- Neural Networks


RGPVNOTES.IN

We hope you find these notes useful.

You can get previous year question papers at https://qp.rgpvnotes.in .

If you have any queries or you want to submit your study notes please write us at rgpvnotes.in@gmail.com

LIKE & FOLLOW US ON FACEBOOK
facebook.com/rgpvnotes.in