Introduction to Data Mining

By Mrs.Chitra Nagarkar
Head, Computer Science Department
Abasaheb Garware College, Pune

Introduction – Why?

- Data is growing
- Requirements of users are becoming more sophisticated
- SQL is not adequate to support these demands

Introduction – What?

- Definition of data mining
  - Finding hidden information from database
  - Exploratory data analysis
  - Data driven discovery
  - Deductive learning

Introduction

- Access to data in Mining environment differs from access to data in DB environment
  - Query – it might not be well formed in mining environment
  - Data – Data accessed is in a different form than original operational DB
  - Output – Output of the data mining query is not a subset of the DB, but it is analysis of contents of DB

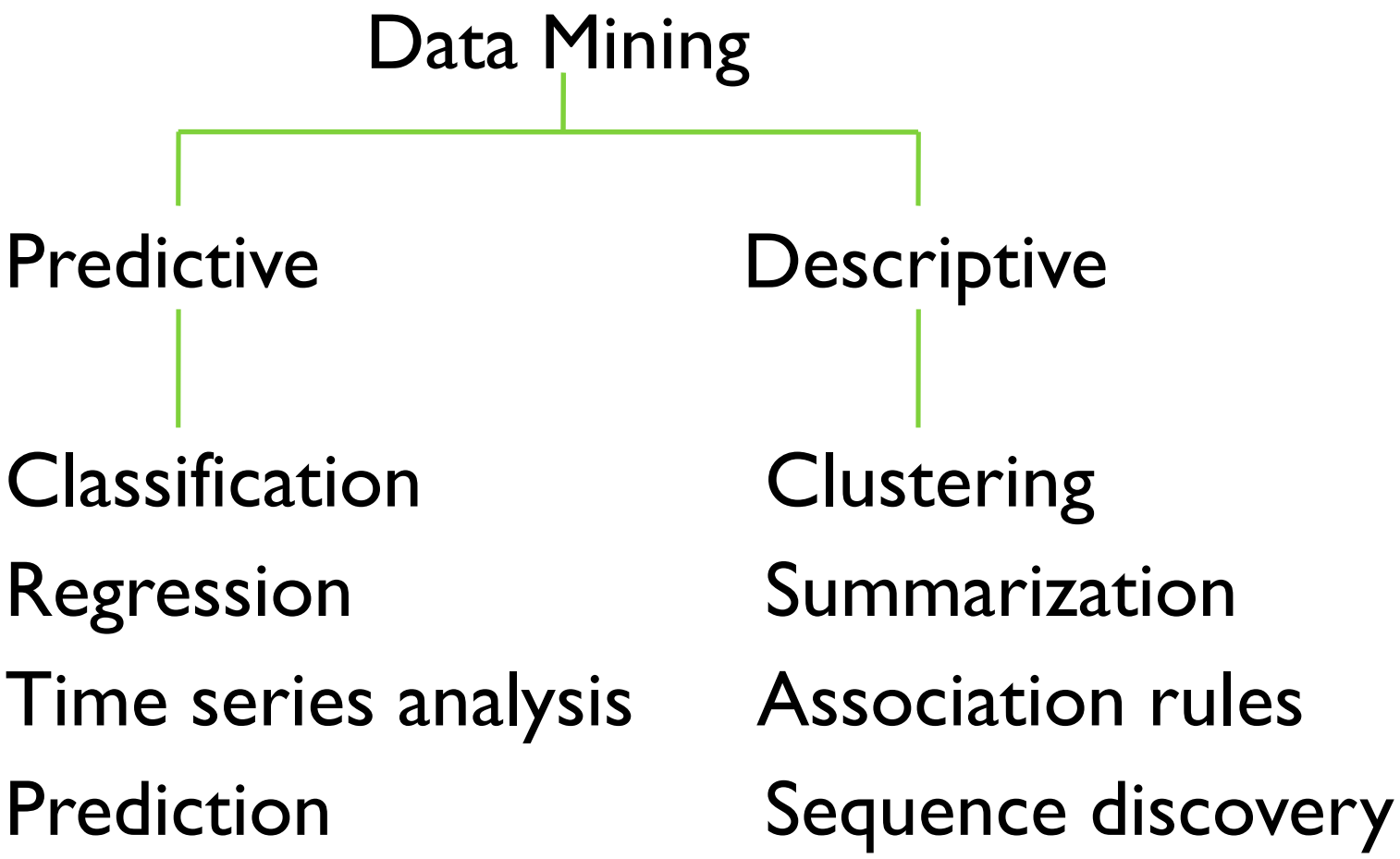
Introduction – How?

- Different algorithms are used to accomplish Data mining tasks. These algorithms examine the data and determines a model that is closest to the data being examined. These algorithms consist of three parts
  - Model – purpose of algorithm is to fit a model to data
  - Preference – criteria used to fit one model over other
  - Search – algorithms require some technique to search data

Introduction

- Predictive model makes predictions regarding data values using the results found from available data. Thus it makes use of historical data to make predictions
- Descriptive model identifies patterns or relationships in data. It finds out the properties of existing data and does not predict the new properties.

Introduction



Basic DM tasks

- Classification – maps data onto predefined groups or classes. This is called as supervised learning as classes are decided before examining the data. Classes are decided based on characteristic of data already belonging to the class
- Pattern recognition is a type of classification, where a given pattern is classified into one of several classes based on its similarity with predefined patterns.

Basic DM tasks

- Regression – maps a data item to real valued prediction variable. This function assumes that target data fits into some known function and tries to find out best function that models the given data.

  Error analysis is used to determine which function is the best.

Basic DM tasks

- Time series analysis – Value of an attribute is examined as it varies over time. Values are obtained at equal time intervals. This function can be used in 3 ways
  - Distance measure is used to determine similarity between different time series
  - Structure of line is used to determine its behavior
  - Historical time series plot can be used to predict future values

Basic DM tasks

- Prediction – In many DM applications future data is predicted based on current or past data.
- Examples are
  - prediction of flooding
  - Speech recognition
  - Machine learning
  - Pattern recognition

Basic DM tasks

- Clustering – It is similar to classification except that the classes are not predefined, but they are defined by data.
- This is also referred as unsupervised learning.
- The similarities among the data based on predefined attributes are used for clustering.
- Since clusters are not predefined, interpretation of clusters is required.
- Segmentation, which partitions DB into disjoint groups of similar tuples, is special type of clustering

Basic DM tasks

- Summarization – maps data onto subsets associated with simple description.
- It is also referred as characterization or generalization.
- This task can be performed by retrieving portion of data or summarizing data

Basic DM tasks

- Association Rules – Tries to find out relationship between data.
- Also called as link analysis or affinity analysis
- Best application of this task is association rules, which is a model identifying specific type of data associations.

Basic DM tasks

- Sequence Discovery – determines sequential patterns in data. Patterns are based on time sequence of actions.
- Patterns are similar to the association in the data, but the relationship is based on time.
- This is also called as sequential analysis or sequence discovery

DM v/s KDD

- Knowledge Discovery in DBs is a process of finding useful information and patterns in data.
- Data Mining is the use of algorithms to extract the information and patterns by KDD process
- KDD involves 5 steps
  - Selection
  - Preprocessing
  - Transformation
  - Data Mining
  - Interpretation / Evaluation

DM v/s KDD

- Visualization is visual representation of data.
- It helps user to summarize, extract and grasp complex results easily than mathematical or textual description.
- Visualization techniques are
  - Graphical
  - Geometric
  - Icon-based
  - Pixel-based
  - Hierarchical
  - Hybrid

Development of Data Mining Techniques

- Data Mining techniques are evolved from many disciplines, which include
  - Databases
  - Information Retrieval
  - Statistics
  - Algorithms
  - Machine Learning
  - Multimedia and Graphics

Development of Data Mining Techniques

- Development in all these areas have created the current view of Data Mining. They have given different views to data mining functions
  - Induction
  - Compression
  - Querying
  - Approximation
  - Search

Data Mining Issues

- Human Interaction
- Over fitting
- Outliers
- Interpretation of Results
- Visualization of Results
- Large Datasets
- High Dimensionality

Data Mining Issues

- Multimedia Data
- Missing Data
- Irrelevant Data
- Noisy Data
- Changing Data
- Integration
- Application