Data Mining

Eng. Mahmoud Ouf

Day 01

Data Mining Definition

There are several definitions for Data Mining:

- Mining is a term characterizing the process that finds a small set of important knowledge from a great deal of raw material.
- Knowledge mining from data
- Knowledge extraction
- Data/Pattern analysis.

Data mining is an essential step in the process of knowledge discovery from Data (KDD).

Knowledge Discovery from Data (KDD)

The knowledge discovery (KDD) process is an iterative sequence of the following steps:

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining
6. Pattern evaluation
7. Knowledge presentation

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining.

What Kinds of Data Can Be Mined?

- Database Data:

searching for trends or data patterns.

detect deviations

- Data Warehouses

Although data warehouse tools help support data analysis, additional tools for data mining are often needed for in-depth analysis

- Transactional Data

Market basket Data Analysis

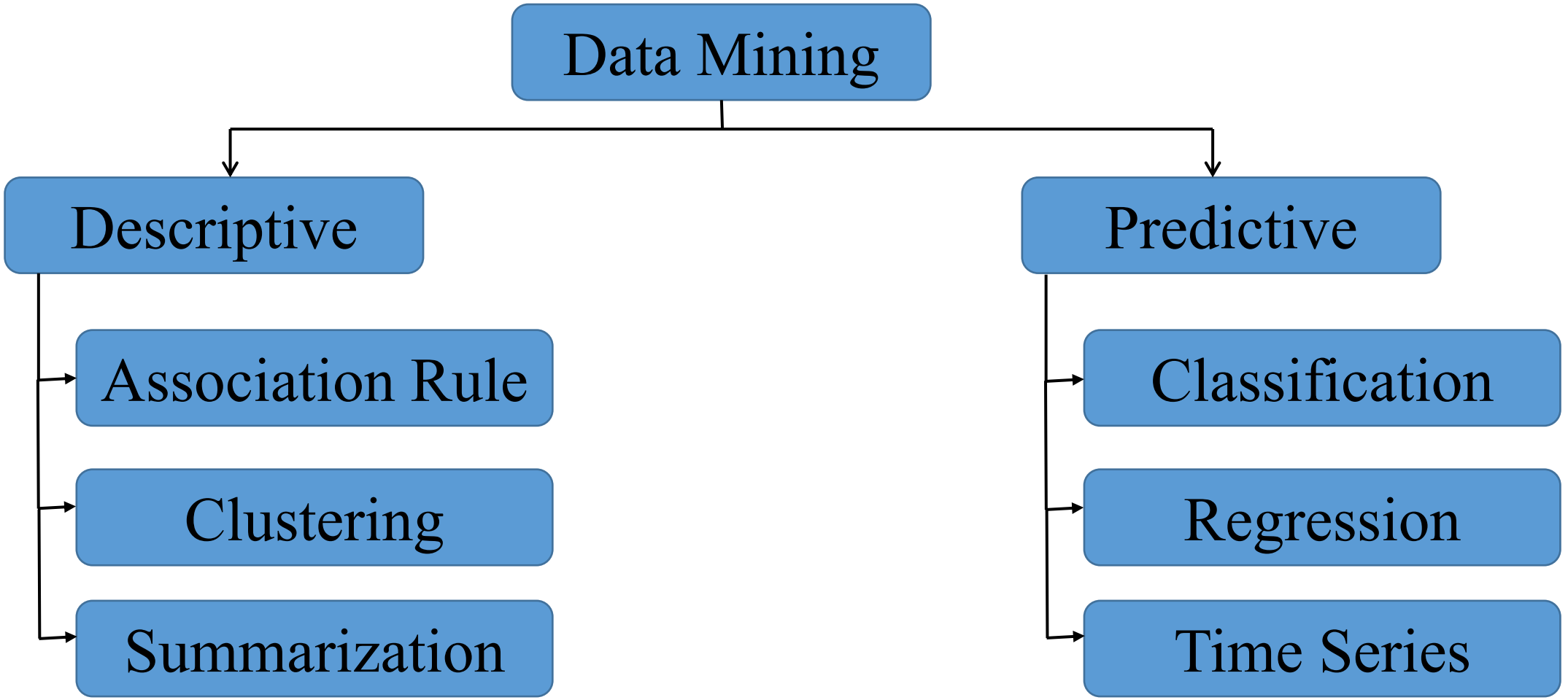
What Kinds of Patterns Can Be Mined?

- There are a number of data mining functionalities includes:
- Characterization and discrimination
- The mining of frequent patterns, Associations, and correlations,
- Classification and regression,
- Clustering analysis, and outlier analysis.

Data mining functionality can be classified into two categories:

- Descriptive mining tasks characterize properties of the data in a target data set.
- Predictive mining tasks perform induction on the current data in order to make predictions.

What Kinds of Patterns Can Be Mined?



Descriptive Data mining

- This is used to generate correlation, frequency, cross tabulation.
- It can be used to discover regularities in the data and to uncover patterns.
- It is also used to find subgroups in the bulk of data.

Association Rules:

What is Association rule?

Association rule is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases

To select interesting rules, constraints on various measures of significance are used.

The best-known constraints are minimum thresholds on support and confidence.

Association Rules

Example:

Transaction_ID	milk	Bread	Butter	Bear	Diapers
1	1	1	0	0	0
2	0	1	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Assume X = {Bread, Butter}

Assume Y = {Milk}

Association Rules

Support:

The support value of X with respect to T is defined as the proportion of transactions in the database which contains the item-set X.

$$\text{Supp(X)} = \frac{Number\ of\ Transactions\ Contains\ Item\ of\ X\ \{Bread,Butter\}}{Total\ Number\ of\ Transactions}$$

Transaction Contains Item of X: Transaction 2 and Transaction 4

Total number of transaction = 5

Supp(X) = 2/5 = 0.4

This Means 40% of all transaction contains itemSet X

Association Rules

Confidence:

The confidence value of a rule, X => Y , with respect to a set of transactions T, is the proportion of the transactions that contains X which also contains Y.

$$\text{Conf}(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)}$$

$X \cup Y$ = {Bread, Butter, Milk}

Supp($X \cup Y$) = 1 / 5 = 0.2

Conf(X=>Y) = 0.2 / 0.4 = 0.5

This means 50% the transactions containing butter and bread contains Milk.

Association Rules:

Mining one level Association (Apriori)

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases.

It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database.

This has applications in domains such as market basket analysis.

Association Rules:

Mining one level Association (Apriori)

Example:

Assume the following Database transaction:

Transaction	Items
T1	Milk, Bread, Cookies, Juice
T2	Milk, Juice
T3	Milk, Egg
T4	Bread, Cookies, Coffee

With minimum support = 0.5 (2)

Association Rules:

Mining one level Association (Apriori)

Solution:

Step1: Create 1st Level Item set

Item	Support
Milk	3
Bread	2
Cookies	2
Juice	2
Egg	1
Coffee	1

Rejected as they are Below the minimum support

Association Rules:

Mining one level Association (Apriori)

Solution:

Step2: Create 2nd Level Item set

Items	Support
Milk, Bread	1
Milk, Cookies	1
Milk, Juice	2
Bread, Cookies	2
Bread, Juice	1
Cookies, Juice	1

Rejected as they are Below the minimum support

Rejected as they are Below the minimum support

Association Rules:

Mining one level Association (Apriori)

Solution:

Step3: Create 3rd Level Item set

Items	Support
Milk, Juice, Bread	1
Milk, Juice, Cookies	1
Milk, Bread, Cookies	1
Juice, Bread, Cookies	1

Rejected as they are Below the minimum support

There is no association at the 3rd level item set

Association Rules:

Mining one level Association (Apriori)

Solution:

We stop the combination of itemset in one of two cases:

- All the last level items are neglected as they are less than the min support
- Reach Level Item set contains all element

Last Step: Association Rules

Milk=>Juice [support = 0.5, confidence = 0.67]

Juice=>Milk [support = 0.5, confidence = 1]

Bread=>Cookies [support = 0.5, confidence = 1]

Cookies=>Bread [support = 0.5, confidence = 1]