DATA MINING

31.03.2014
(Muscat, Oman)

OUTLINE

- Definition of Data Mining
- Using Areas of Data Mining
- Data Mining Tasks
- Some examples about Data Mining Methods

Data , Information and Knowledge

(Raw facts) (Organized facts and figures) (Meaningful information, helps to make decision)

Data is anything that recorded and processed by computers. Data is raw fact, may be numeric or text format.

"**Information**" is "data" that has been processed.

Information can be converted into ***knowledge*** . We make decisions using knowledge.

Example 1:

Company 1	25 TL
Company 2	30 TL
Company3	10 TL
Company 4	40 TL

Data : 25, 30, 10, 40

Information : These are prices of the st. For these companies

Knowledge : the cheapest one is company3, deciding to buy

Problem ? :

Huge and Meaningless Data

Data is not valueable while querying with traditional methods

How can we get knowledge to compete ?

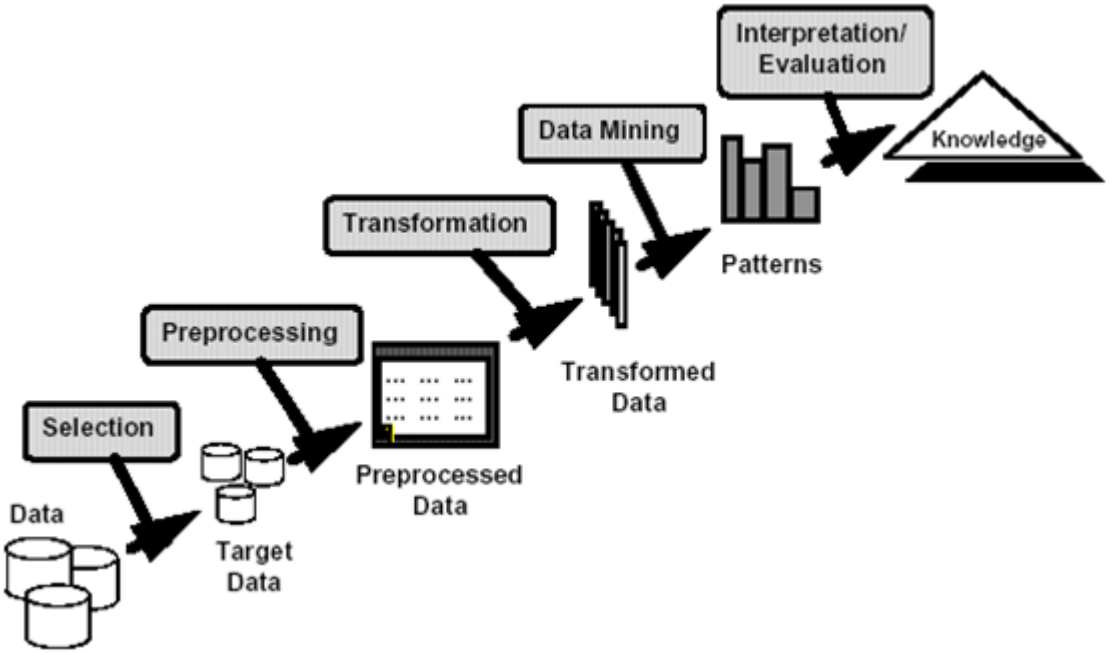
Solution :

Data Mining Process

DATA MINING OVERVIEW

Data Mining is the process of **analyzing data** from different perspectives and summarizing it into useful information

Data Mining is **knowledge discovery** from data



DATA MINING APPLICATION AREA EXAMPLES

Medical / Pharma

Computer Assisted Diagnosis (expert systems learning)
Characterization/prediction of patient's response to product dosage
Identification of successful medical therapies (successful prescription patterns).
Study of relations between dosage and potentially related adverse events

Banking / Finance

Detection of fraudulent credit card usage patterns.
Risk management related to attribution of loans using scorecards.
Find hidden correlations between different financial indicators.
Identification of stocks trading rules from historical market data.

Retail / Marketing

Discovery of buying behavior patterns
Detection of associations among customer characteristics.
Prediction of the probability that clients answer to mailing.

Some Possible Questions :

Banking :

Does he take a new credit card or manage to pay it

Stock market

Will gold prices increase

Shopping market:

Which items are frequently purchased

Does he interest new fashion shoes/hats/clothes

Hospital:

Which ilness may be..

How long does it take to recover from this illness

Answers:

Looking historical data or finding correlation with **Data Mining**

DATA MINING & DATA WAREHOUSING & OLTP

- OLTP system includes transactional data, unnecessary data for analysing
- Data Warehouses is designed for analysis
- Data warehouses are subject based, fast
- Data warehouses are designed to use in decision support systems
- Data Mining is performed on **Data Warehouses**

Data Mining Tasks

1. Defining Problem is the first task. May be it is the most important phase because, according to the purpose, using methods and rules will be changed.

The goal should be clear.

Some Example Goals:

Is there a relationship between gender and education ?

Is there a relationship between travelling expenditure and education?

What type of customer does have problem with paying their loan?

Data Mining Tasks (Cont'd)

2. Preparing data :

This phase takes most time.

While preparing data, data is taken from different sources. Then all data is combined and cleaned. Sometimes, generating new variables is needed.

Not only the size but also the quality of the data is very important.

After providing data quality, modelling task will begin.

3. Modelling task is performed using different methods.

Data Mining tools are used

After modelling finished, it will be **evaluated** and **used**.

Data Mining Methods

Prediction Methods

Some variables are used to predict unknown or future values of other variables.

Example : classification, regression, deviation detection , decision trees

Description Methods

Find human-interpretable patterns that describe the data.

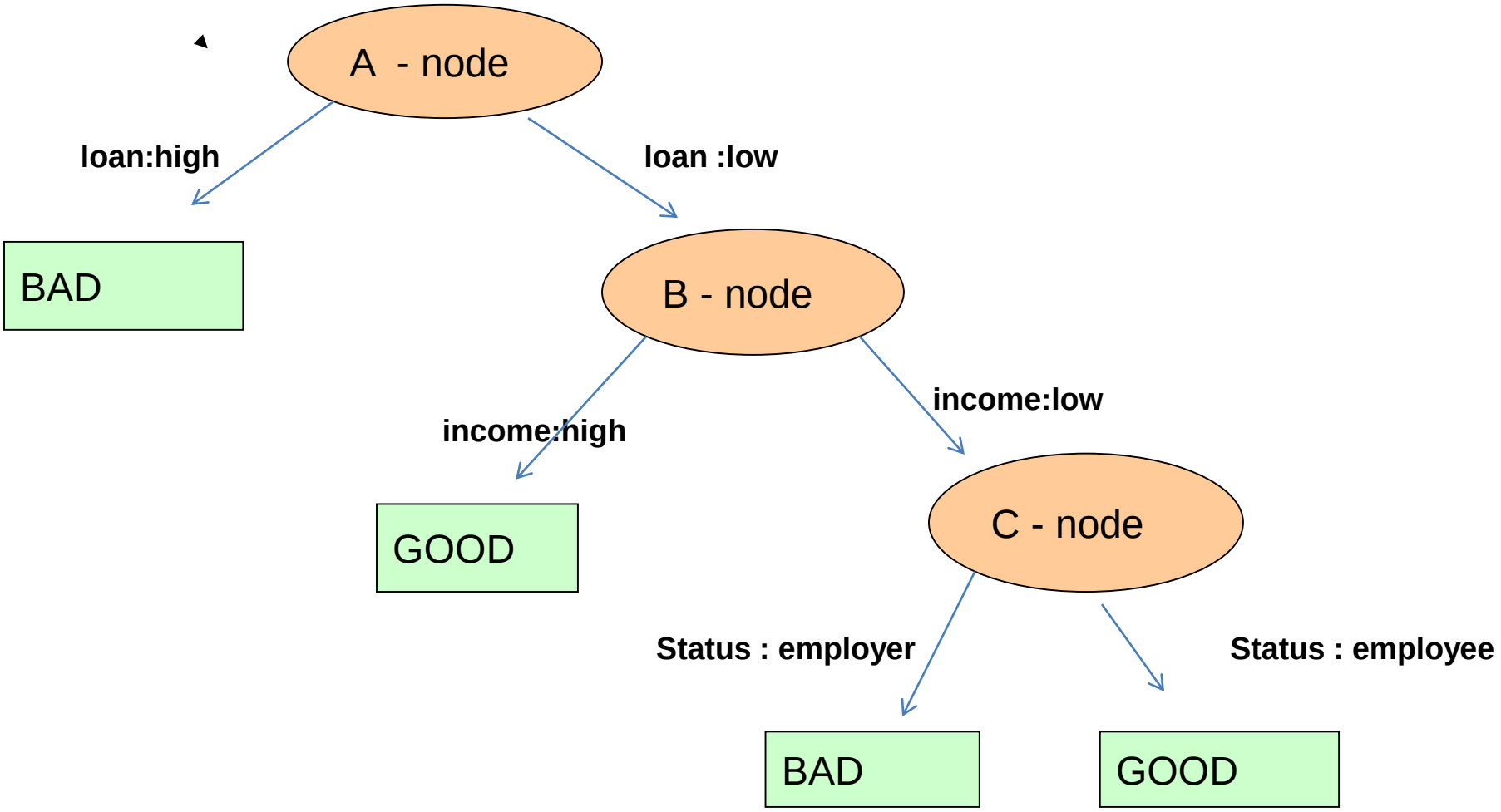
Example: clustering, assocation rule discovery,..

DECISION TREE EXAMPLE

Example : Use these data to predict the risk of a customer.

Customer	Loan	Income	Status	Risk
1	high	bad	employer	bad
2	high	high	employee	bad
3	high	low	employee	bad
4	Low	low	employee	good
5	low	low	employer	bad

DECISION TREE



REGRESSION ANALYSIS

The goal of the regression analysis is finding Relationship among variables

Regression analysis is a process of **finding releationships** between Y variable and other X1,X2,….Xn variables.

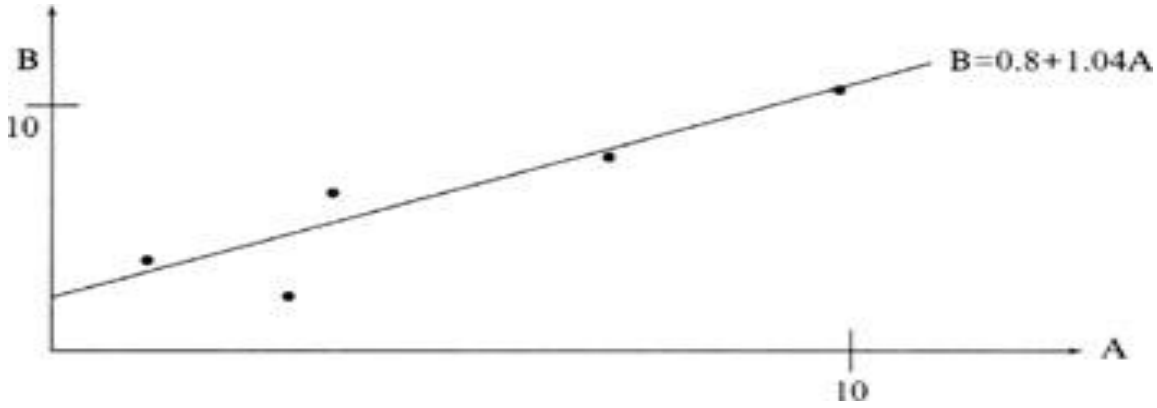
It includes many techniques for modelling and analyzing several variables

A Regression equation example is :

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \ldots + \beta_n \cdot X_n$$

Example2 : Regression Analysis

A (independent variable)	B (dependent variable)
1	3
8	9
11	11
4	5
3	2



$$B = \alpha + \beta \cdot A$$

$$\alpha = 0.8$$

$$\beta = 1.04$$

$$B = 0.8 + 1.04 \cdot A$$

Some Popular Data Mining Tools:

- SAS Enterprise Miner
- IBM SPSS Modeler
- ORACLE Data Mining
- RapidMiner (Open Source)
- Pentaho (Open Source)

Summary

- ✓ Data Mining is the process of getting information from data
- ✓ Before modelling, determining the goal is most important
- ✓ The quality of the data is also very important
- ✓ Data Mining is performed on Data Warehouses
- ✓ Data Mining has prediction and description methods
- ✓ Some Data Mining Tools are used for the Data Mining process

Summary of Business Intelligence Components:

Business Intelligence System Component	How Used in Decision-Making
ETL Tools	Used to obtain, adjust and load data from both operational databases and dispersed data sources allowing for the collection of volumes of data (Schink, 2009) which allows for: • near real-time information access • uniform data type in which to analyze
Data Warehouses	Used as repository for all data relevant to an organization to support the decision-making process (Matei, 2010) by: • gathering relevant and context aware data • providing multiple dimensions to data
OLAP Techniques	Used to analyze and report data from huge data sources (Olszak & Ziemba, 2006) by: • providing user access to data warehouses • creating data models
Data Mining	Used to identify patterns and relationships within a data warehouse and creates detailed reports (Hevner & March, 2005) allowing for: • predictions based on historical data • graphing and calculating to create formulas to analyze data

Thank you & Questions ?