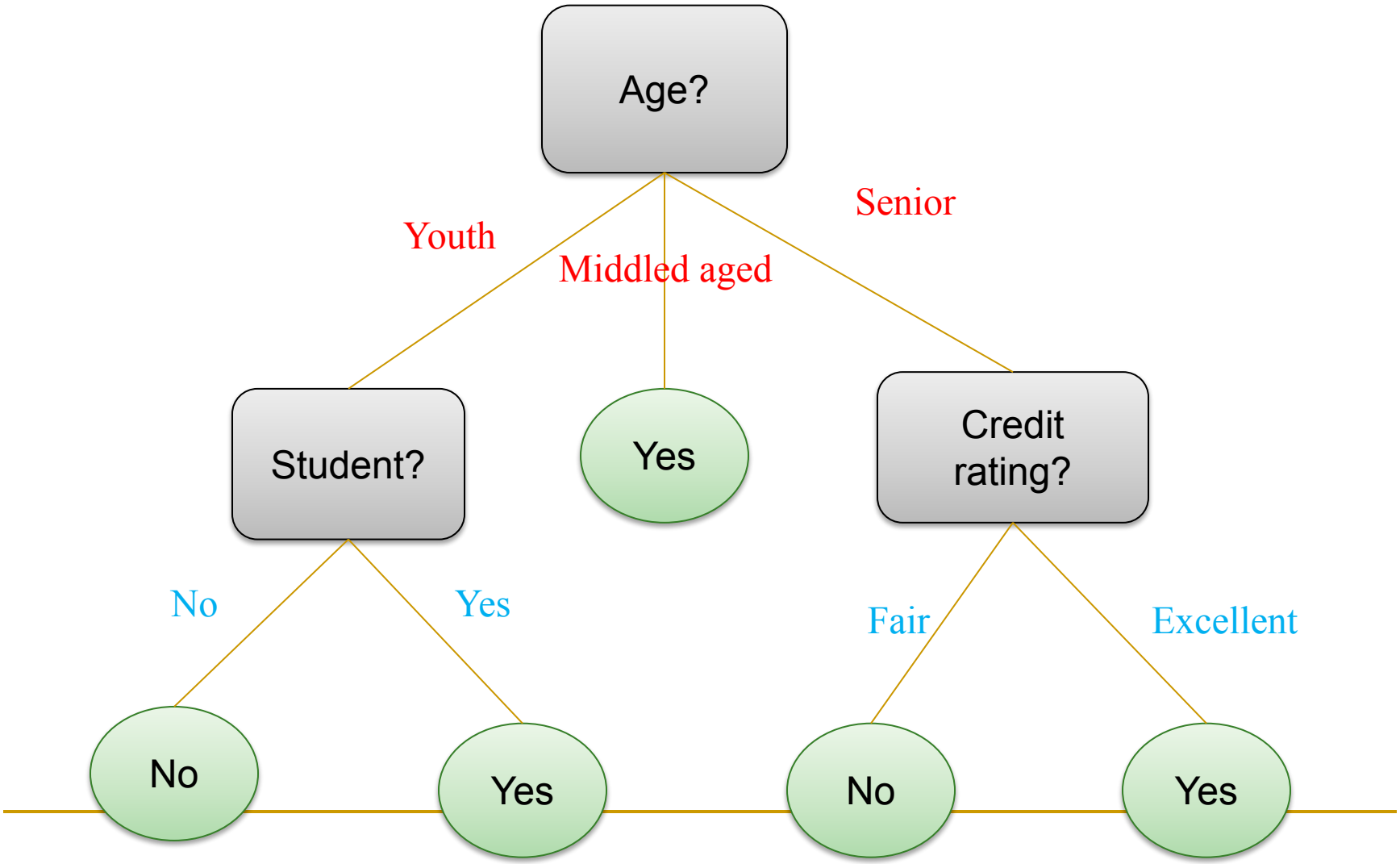
Data Preprocessing

Problem

A marketing manager at All Electronics needs data analysis to help guess whether a customer with a given profile will buy a new computer.

	|---------- X ----------|			Y
RID	Age	Income	Student	Credit rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle aged	medium	no	excellent	yes
13	middle aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Classification Tree



Understanding Data

- What data is available for the task?
- Is this data relevant?
- Is additional data available?
- How much historical data is available?
- Who is the data expert?

Classification contd…

A marketing manager at AllElectronics needs data analysis to guess set of components that a new customer with a given profile will buy from a finite set of component { computer, camera, mobile}.

	|---------- X ----------|			Y
RID	Age	Income	Student	Credit rating	Set of items buyed
1	youth	high	no	fair	Camera, Mobile
2	youth	high	no	excellent	Camera, Mobile
3	middle aged	high	no	fair	Computer, Camera
4	senior	medium	no	fair	Camera
5	senior	low	yes	fair	Camera
6	senior	low	yes	excellent	Mobile
7	middle aged	low	yes	excellent	Computer, Camera
8	youth	medium	no	fair	Camera, Mobile
9	youth	low	yes	fair	Computer
10	senior	medium	yes	fair	Camera
11	youth	medium	yes	excellent	Computer
12	middle aged	medium	no	excellent	Computer, Camera
13	middle aged	high	yes	fair	Computer, Camera
14	senior	medium	no	excellent	Mobile

Database and Data Warehouse

Database:

It is a structured collection of records or data.

Data Warehouse:

It is a logical collection of information, gathered from many different operational databases, that supports analysis activities and decision-making tasks.

A data warehouse is mostly used to facilitate reporting and analysis.

Why Data Preprocessing?

- Welcome to real data!
  - Incomplete data: missing values, containing only summarized data
    - e.g., income= --
  - Noisy: containing errors or outliers
    - e.g., weight=-10
  - Inconsistent: containing discrepancies in codes or names
    - e.g., Age= 50, Birthday=5/08/2001
    - e.g., duplicate records

Why Data Preprocessing is Important?

- No quality data, no quality result?
- Preprocessing is one of the most critical steps in data analysis process.

What is Data Mining?

Data mining is extracting the interesting (previously unknown, potentially useful) pattern or knowledge from huge amount of data.

The alternative names are

- Knowledge Discovery in Database (KDD)
- Knowledge Extraction
- Data/ Pattern Analysis
- Business Intelligence etc.

Major Tasks in Data preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers etc
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization, discretization and aggregation
- Data reduction
  - Obtains reduced representation in volume of data but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Data cleaning

Real-world data tend to be incomplete, noisy and inconsistent.

Data cleaning task:

- Fill in missing values
- Indentify outliers/smooth noisy data
- Detect and Correct inconsistent data

Missing Values

- Ignore the record or table – this is usually done when the class label is missing.

- Fill the missing values manually – In general this method is time consuming and may not be feasible given large data set with many missing values.

- Use a global constant fill in the missing values: Replace all missing attribute value by same constant such as a label like ‘’Unknown” or ∞.

Missing Values...

- Fill the missing values by the middle value of the distribution (Mean or Median)

- Fill the missing values by the most likely value (this value may be determined using regression, decision tree, most similar records etc.).

Outliers/smooth noisy data

- Noise is the random error or variance in a measured variable.

- Identify outliers using some basic statistical techniques like box plot or scatter diagram or some other data visualisation methods.

- The incorrect attributes values occur due to faulty data collection, data entry problem, data transmission problem etc.

How to Handle Noisy Data?

- Binning method

  first sort data and partition into (equal-frequency) bins

  then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

- Regression:

  smooth by fitting the data into regression functions

- Clustering:

  detect and remove outliers

Binning Method...

Example: 4, 8, 15, 21, 21, 24, 25, 28, 34.

Equal-depth (frequency) partitioning:

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

Soothing by bin means

Bin 1: 9, 9, 9

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

Inconsistent data

- Inconsistent data occurs due to data entry error, data integration (different format or codes etc.).

- It is important to data entry verification (check the format and value of data entered) and corrects with help of extra reference data (e.g. Male/0 - M, Female/1 - F).

Data Integration

Data integration: Combines data from multiple sources into a coherent store

e.g., A.cust-id ≡ B.cust-#

Integrate data from different sources

Entity identification problem:

Identify real world entities from multiple data sources

Detecting and resolving data value conflicts

For the same real world entity, attribute values from different sources are different

Possible reasons: different representations, different scales, e.g different units

Data Transformation

- *Smoothing:* This method is used to remove noise from the data (binning, regression and clustering)
- *Attributes construction:* New attributes are constructed and added from the given set of attributes.
- *Aggregation (Summarization):* Aggregation operations are applied to the data

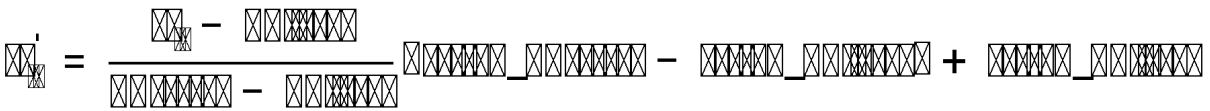
  e.g. Daily sale data may be aggregated to compute monthly and annual total amounts.
- *Discretization:* the raw value of attributed replaced by interval labels.

  e.g. Age – labeled as 0-10 years, 10-20 years like

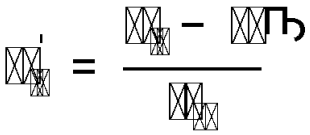
Data Transformation

Normalization

- *Min-max normalization:* Suppose minA and maxA are the minimum and maximum values of an attribute A. Min-max normalization map the value of A to in the range [new_minA, new_maxA] by computing.



- *Z-score normalization:* the value of an attribute A are normalized based on the mean and S.D of A. The transformation is



where -new value of A, -mean of A, -value of A and -S.D of A.

Data Reduction

Why data reduction?

A database/data warehouse may store data

Complex data analysis may take a very long time to run on the complete data set

Data reduction

Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results

Data reduction strategies

Data aggregation:

Dimensionality reduction — e.g., remove unimportant attributes/ Variables

Data Compression

Numerosity reduction — e.g., fit data into models

Discretization

Data Discretization

Discretization techniques can be used to reduce the number of values for a given continuous attribute, by dividing the attribute into range of interval value labels can be used to replace data values.

- Binning
- Histogram
- Clustering

Binning Method

The stored values are distributed into a number of buckets or bins and then replacing each bin value by the bin or median.

1. Equal width (distance) partitioning:

- Divide the range into N intervals of equal size: Uniform grid.
- W=(B-A)/N, B and A are the highest and lowest value.

2. Equal-depth (frequency) partitioning:

Divide the range into N intervals, each containing approximately equal number of samples.

Binning Method...

Example: 4, 8, 15, 21, 21, 24, 25, 28, 34.

1. Equal width (distance) partitioning:

Here, W=(34-4)/3=10

Bin 1: 4-14

Bin 2: 15-24

Bin 3: 25-34

Bin 1: 4, 8

Bin 2: 15, 21, 21, 24

Bin 3: 25, 28, 34

2. Equal-depth (frequency) partitioning:

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

Histogram

Histogram can also be used for discretization and partitioning rules can be applied range of values. The histogram algorithm can be applied in order to automatically generate a multilevel concept hierarchy, with the procedure terminating once a pre-specified number of concept level have been reached. A minimum interval size can be used per level to control the recursive procedure. This specify the minimum width of partition or minimum member of partition at each level.

1. Equal width
2. Equal frequency

Cluster Analysis

Cluster analysis is a popular method for data discretization. A clustering algorithm can be applied to discrete a numeric attribute A by partitioning the value of A into cluster or groups and use the following methods.

1. Top-down splitting strategy
2. Bottom-up merging strategy

Data Visualization

Data visualization is an effective technique to graphically display large volume of data by converting raw data into meaningful images for effortless human comprehension or communication. The data can be presented in many terms.

1. Summarizations: Table form
2. Pictorial representation- Graphs and Diagram

Diagram

Single variable Diagram:

- Line charts
- Bar Chart
- Histogram
- Time charts
- Pie charts
- Frequency polynomial
- frequency curve
- Steam and Leaf plots

Multivariable Diagrams:

- Scatter plot
- Contour chart
- Bubble plot

Data Mining Process

SEMMA (SAS)- Sample, Explore, Modify, Model, Asses.

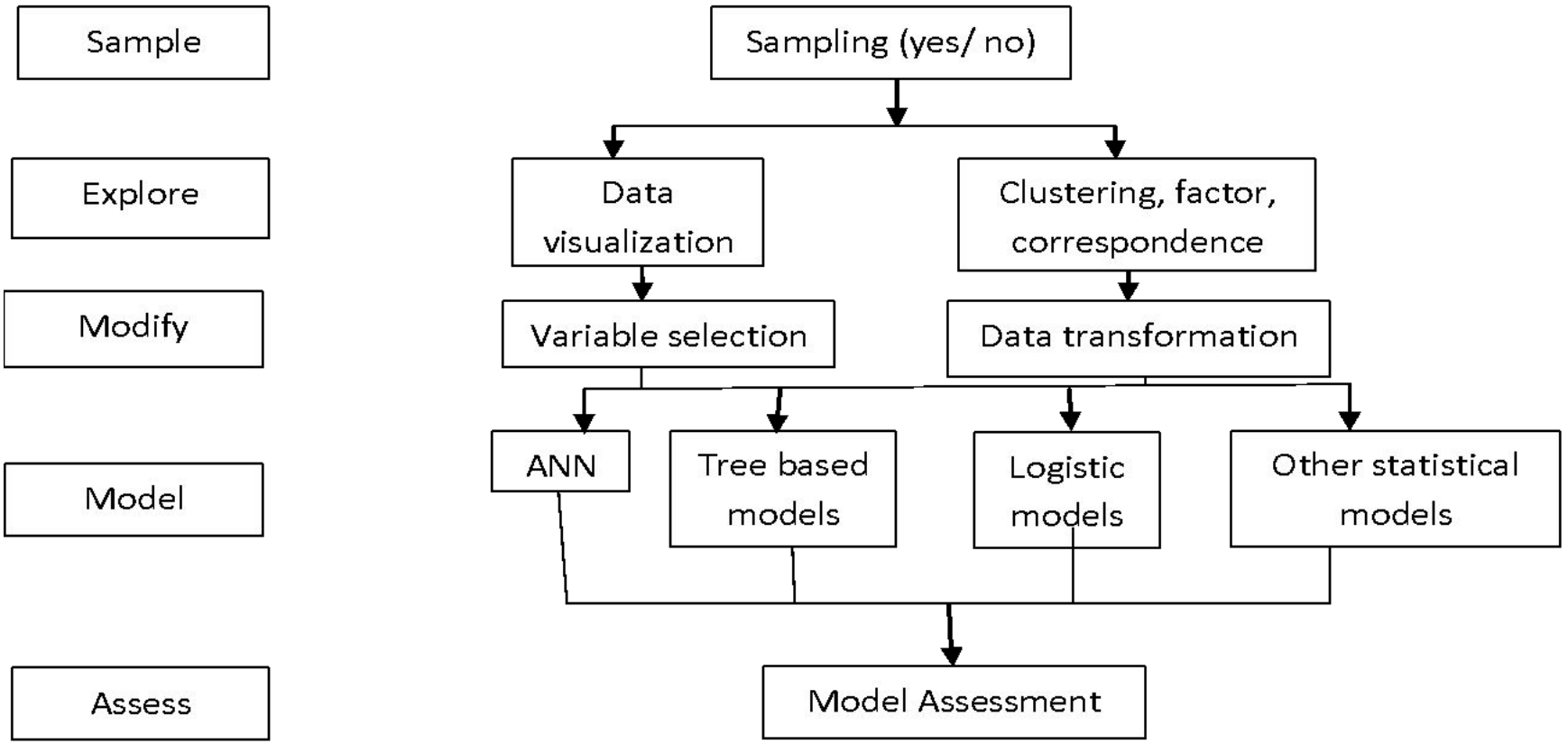
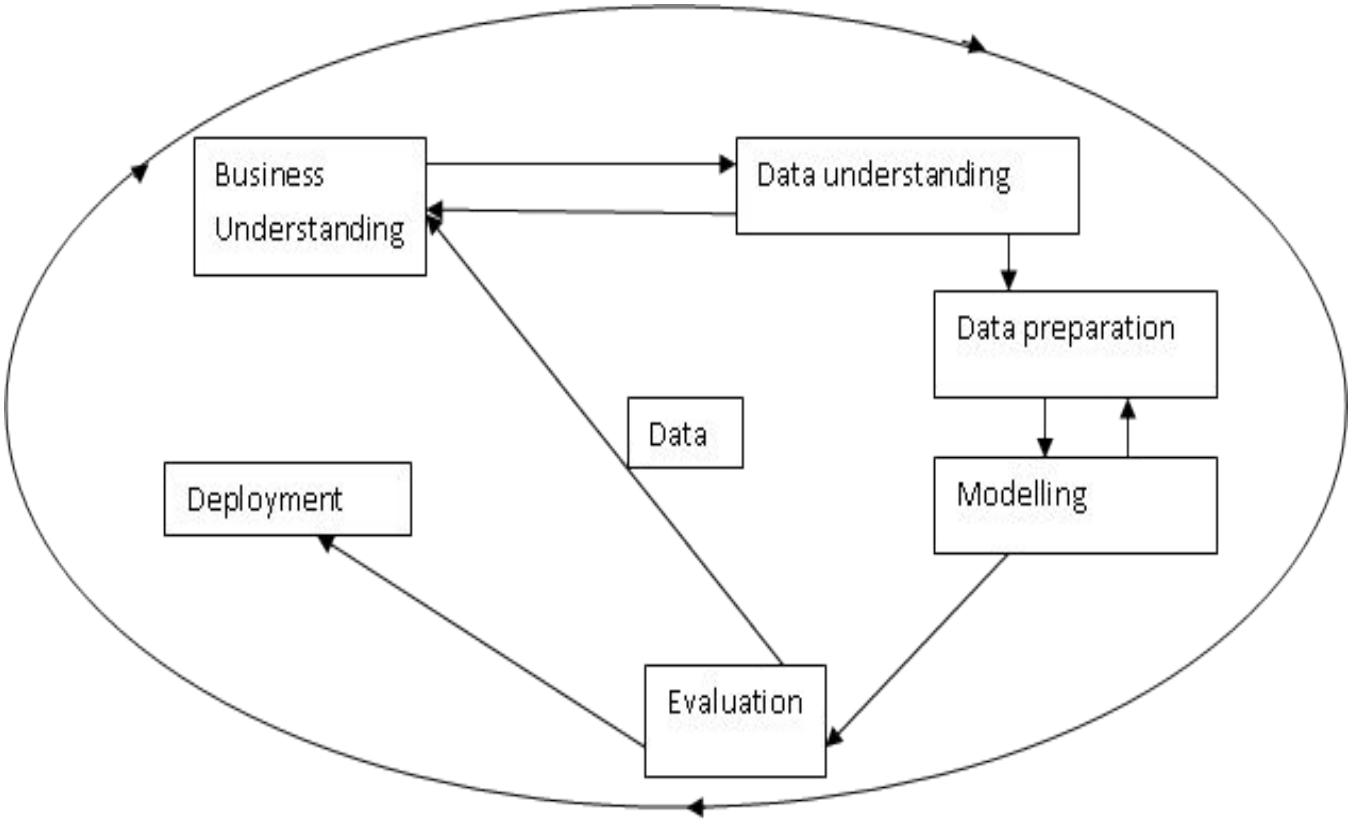


Fig. Steps in SEMMA methodology

Data Mining Process

CRISP-DM (SPSS)-

Cross Industry Standard Process for Data Mining



Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Descriptive data summarization is need for quality data preprocessing
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and variable selection
  - Discretization
- A lot a methods have been developed but data preprocessing is still an active area of research