Data mining

What is Data Mining?

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.

The data sources can include

- databases,
- data warehouses,
- the Web,
- other

What is NOT Data Mining?

Simple Search & Querying

- The query takes a decision according to the given condition in SQL.
- For example, a database query "SELECT * FROM table" is just a database query and it displays information from the table but actually, this is not hidden information.
- So it is a simple query and not data mining.

Expert systems (in artificial intelligence)

- The expert system takes a decision on the experience of designed algorithms

Steps in Data Mining Process

1. Data Cleaning
2. Data Integration
3. Data Reduction
4. Data Transformation
5. Data Mining
6. Pattern Evaluation
7. Knowledge Representation

Steps in Data Mining Process

1. Data Cleaning
2. Data Integration
3. Data Reduction
4. Data Transformation
5. Data Mining
6. Pattern Evaluation
7. Knowledge Representation

Steps in Data Mining Process

1. Data Cleaning
2. Data Integration
3. Data Reduction
4. Data Transformation

} Data Preprocessing

5. Data Mining
6. Pattern Evaluation
7. Knowledge Representation

Data Pre-Processing
Or
Pre-Processing of Data

“How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results?

How can the data be preprocessed so as to improve the efficiency and ease of the mining process?”

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing

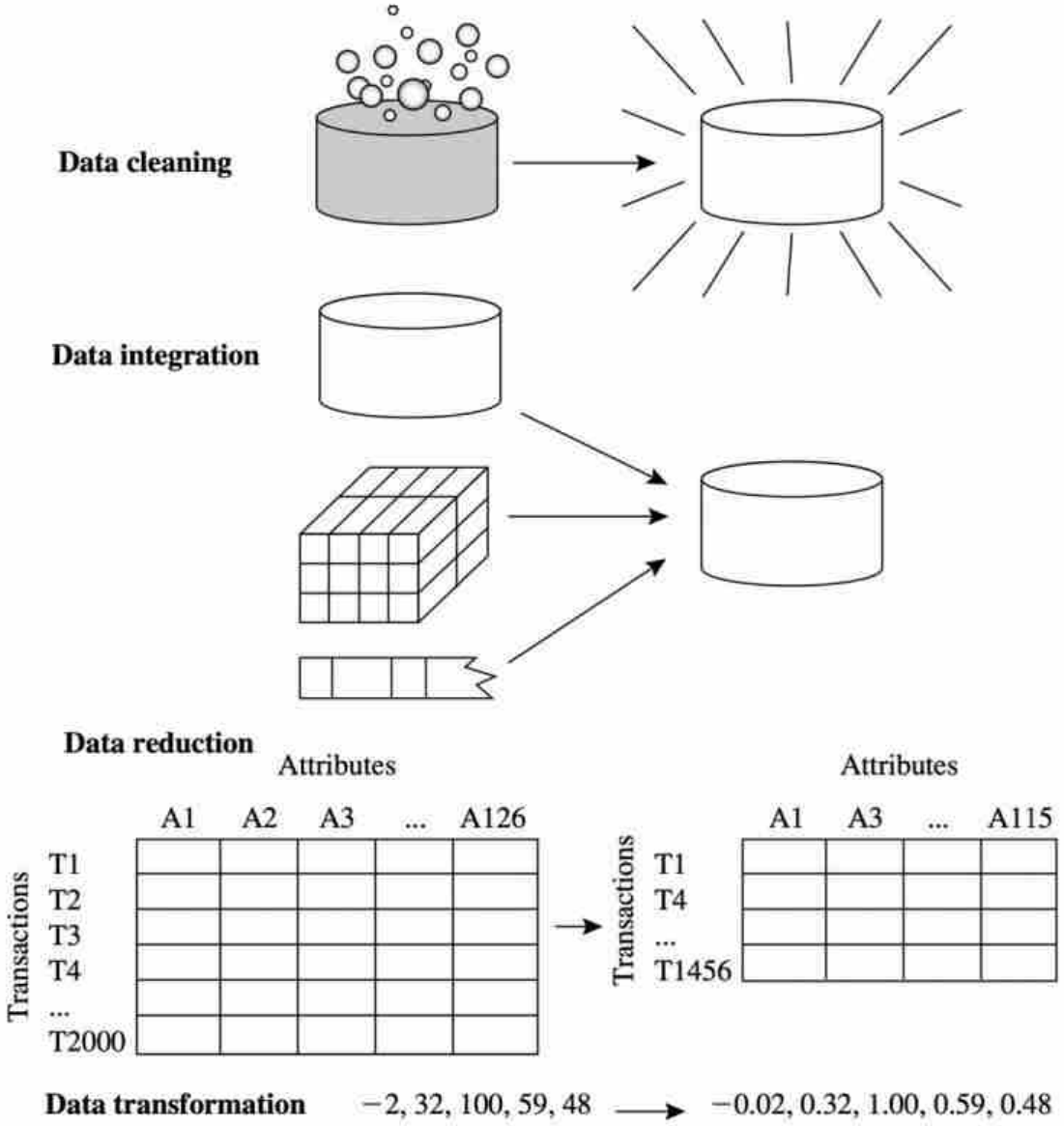
- Data cleaning
- Data integration
- Data reduction
- Data transformation and data discretization

Data Cleaning
Data cleaning is a process to clean the data in such a way that data can be easily integrated.

Data Integration
Data integration is a process to integrate/combine all the data.

Data Reduction
Data reduction is a process to reduce the large data into smaller once in such a way that data can be easily transformed further.

Data Transformation
Data transformation is a process to transform the data into a reliable shape.



Data Cleaning - *Incomplete data* or *missing data*

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

Dirty Data	Example (missing data)
Incomplete or missing data	Salary = “ ”
	Occupation = “ ”
	Marks = “ ”

Data Cleaning - *Noisy data*

- containing noise, errors, or outliers

Dirty Data	Example
an error	Salary = “-5000”
	Salary = “-10”
	Name = “123”

Data Cleaning - *inconsistent*

- containing discrepancies in codes or names,

Dirty Data	Example
	Age="42", Birthday="03/07/2010"
	Previous rating "1, 2, 3", now rating "A, B, C"
	discrepancy between duplicate records *student names are different in differnet records*

Data Cleaning - *Intentional*

- Intentional error (e.g., disguised missing data)

Dirty Data	Example
Intentional error	Jan. 1 as everyone's birthday?
	gender="male" e.g some application put gender value as male by default. *(e.g. google form or survey)*
	Sometimes applications alot auto value to attribute.

How to Handle Missing Data?

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with a measure of central tendency (mean or median)
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
- the most probable value: inference-based such as Bayesian formula or decision tree

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value
- Fill in it automatically wi
  - a global constant : e
  - the attribute mean
  - the attribute mean f
- the most probable value
  tree

What is tuple?

Tuples are used to store multiple items in a single variable

(1.0, 9.9, 10) — a tuple containing three numeric objects

('Casey', 'Darin', 'Bella', 'Mehdi') — a tuple containing four string objects

('Casey', ' ', 'Bella', 'Mehdi') — a tuple with missing value

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value
- Fill in it automatically w
  - a global constant : e
  - the attribute mean
  - the attribute mean f
- the most probable value tree

What is tuple?

Tuples are used to store multiple items in a single variable

(1.0, 9.9, 10) — a tuple containing three numeric objects.

('Casey', 'Darin', 'Bella', 'Mehdi') — a tuple containing four string objects.

('Casey', ' ', 'Bella', 'Mehdi') — a tuple with missing value.

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with a measure of central tendency (mean or median)
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
- the most probable value: inference-based such as Bayesian formula or decision tree

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with a measure of central tendency (mean or median)
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
- the most probable value: inference-based such as Bayesian formula or decision tree

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with a measure of central tendency (mean or median)
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
- the most probable value: inference-based such as Bayesian formula or decision tree

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with a measure of central tendency (mean or median)
  - a global constant : e.g., "unknown", or "$-\infty$", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
- the most probable value: inference-based such as Bayesian formula or decision tree

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with a measure of central tendency (mean or median)
  - a global constant : e.g., "unknown", or "$-\infty$", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter

Explain What you Learned today to Your Friend

Thank You!