Introduction to Data Mining

— UNIT-04—

Oracle Documentation

Why Data Mining?

- The Explosive Growth of Data
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets

What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

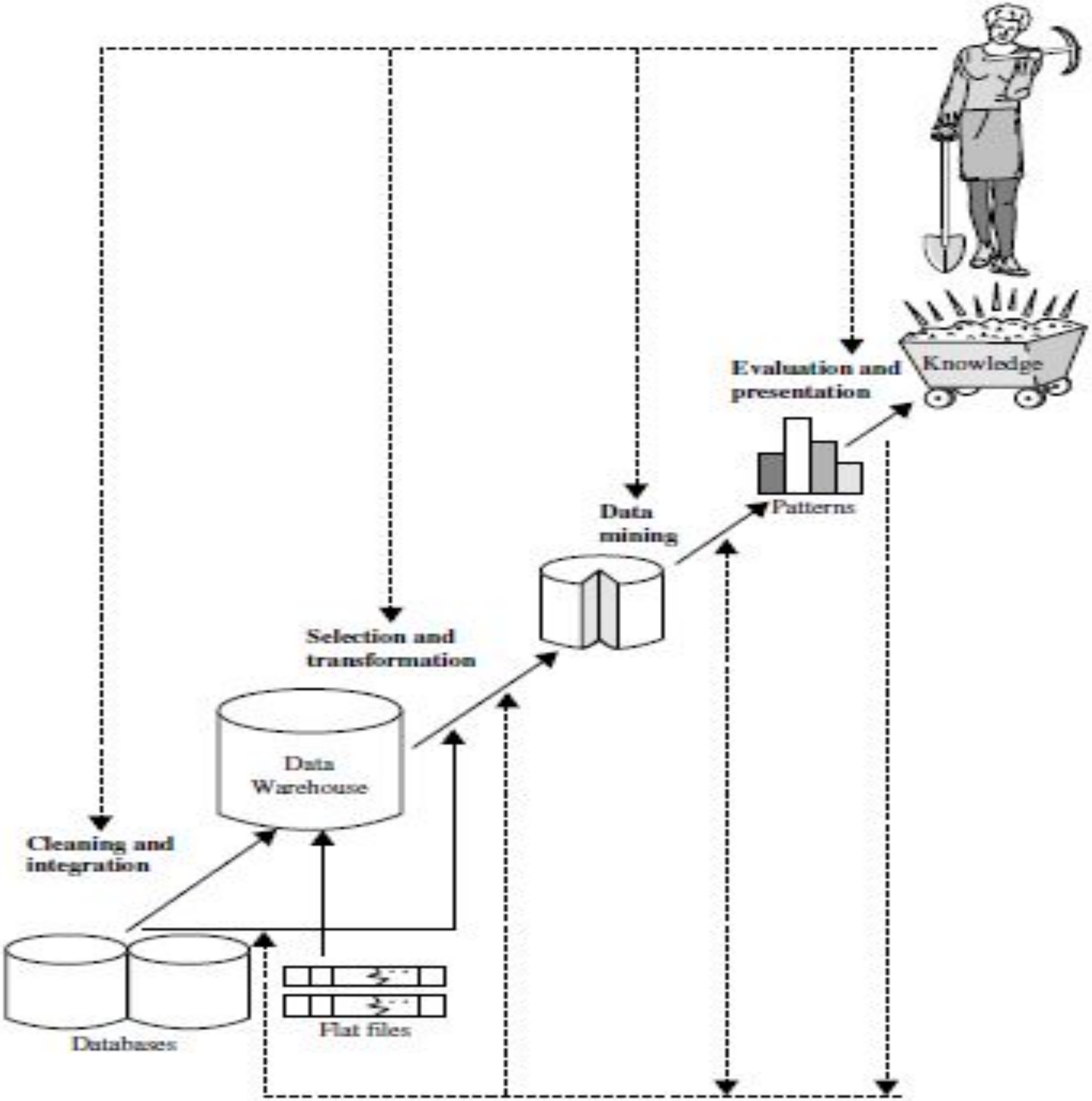


Figure 1.4 Data mining as a step in the process of knowledge discovery.

- **1. Data cleaning** (to remove noise and inconsistent data)
- **2. Data integration** (where multiple data sources may be combined)
- **3. Data selection** (where data relevant to the analysis task are retrieved from the database)
- **4. Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
- **5. Data mining** (an essential process where intelligent methods are applied to extract data patterns)
- **6. Pattern evaluation** (to identify the truly interesting patterns representing knowledge*)*
- **7. Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)

- Low-quality data will lead to low-quality mining results
- Data processing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined

What Kinds of Data Can Be Mined?

- Data mining can be applied to any kind of data as long as the data are meaningful for a target application.
- The most basic forms of data for mining applications are
  - database data
  - data warehouse data
  - transactional data
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, sequence data

- Spatial data
- Engineering design data
- Hypertext and multimedia data
- graph and networked data
- The World-Wide Web

What Kinds of Patterns Can Be Mined?

- Data mining functionalities.
  - Characterization
  - Discrimination
- Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks.
- In general, such tasks can be classified into two categories: **descriptive and predictive**.
- 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.

Data characterization

- Data characterization is a summarization of the general characteristics or features of a target class of data

Example

- A customer relationship manager at AllElectronics may order the following data mining task: Summarize the characteristics of customers who spend more than $5000 a year at AllElectronics. The result is a general profile of these customers, such as that they are 40 to 50 years old, employed, and have excellent credit ratings

Data discrimination

- Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.
- **Example**

A customer relationship manager at AllElectronics may want to compare two groups of customers—those who shop for computer products regularly (e.g., more than twice a month) and those who rarely shop for such products (e.g., less than three times a year).

The resulting description provides a general comparative profile of these customers, such as that 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education, whereas 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree.

Mining Frequent Patterns, Associations, and Correlations

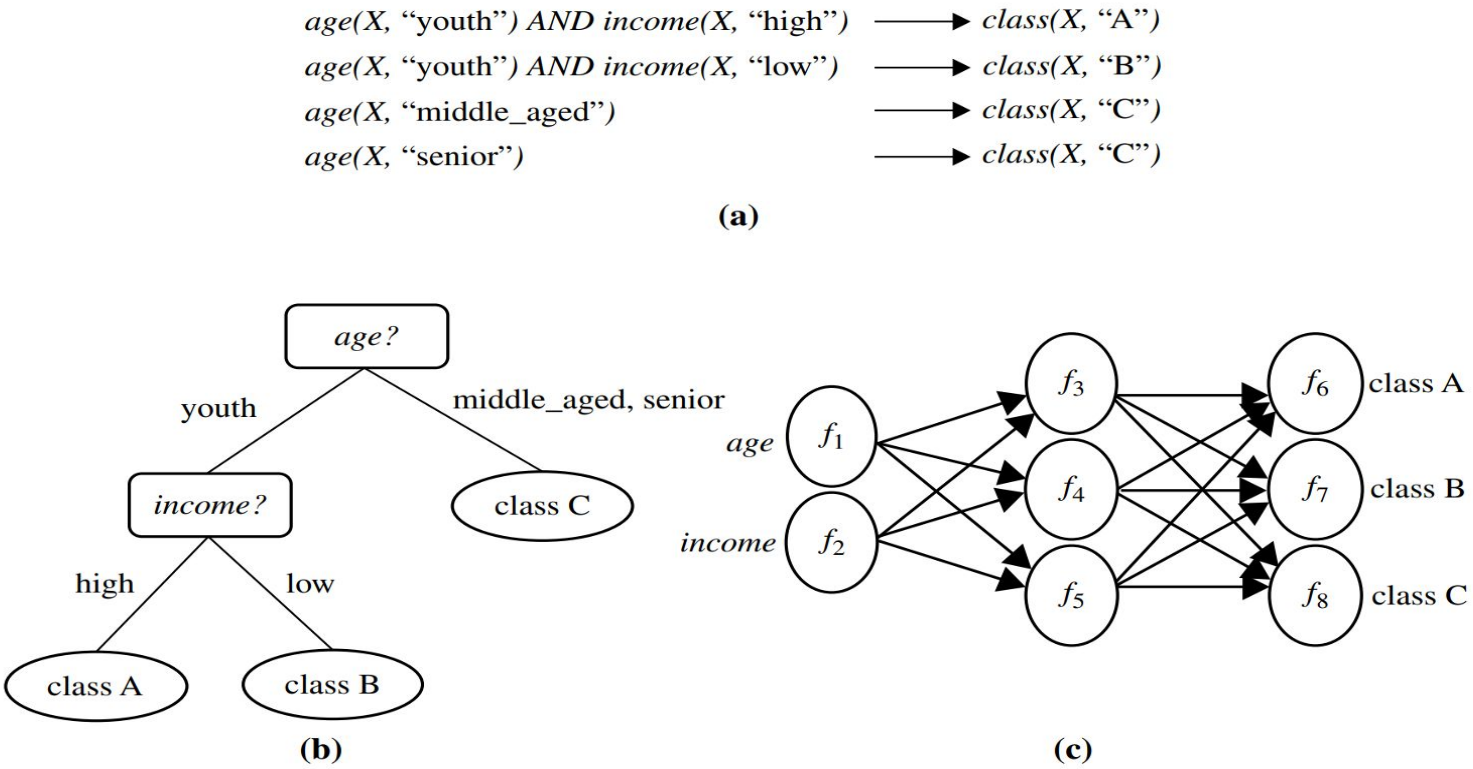
- Frequent patterns, as the name suggests, are patterns that occur frequently in data.
- A frequent itemset typically refers to a set of items that often appear together in a transactional data set—for example, milk and bread
- **Association analysis**. Suppose that, as a marketing manager at AllElectronics, you want to know which items are frequently purchased together (i.e., within the same transaction). An example of such a rule, mined from the AllElectronics transactional database, is
- **buys(X, “computer”) ⇒ buys(X, “software”) [support = 1%,confidence = 50%],**

- Association rules that contain a single predicate are referred to as single-dimensional association rules.
- Suppose, instead, that we are given the AllElectronics relational database related to purchases. A data mining system may find association rules like
- **age(X, “20..29”) $\wedge$ income(X, “40K..49K”) $\Rightarrow$ buys(X, “laptop”) [support = 2%, confidence = 60%].**
- Note that this is an association involving more than one attribute or predicate (i.e., age, income, and buys).
- Adopting the terminology used in multidimensional databases, where each attribute is referred to as a dimension, the above rule can be referred to as a multidimensional association rule

- Typically, association rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold.

Classification and Regression for Predictive Analysis

- Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts



age(*X*, “youth”) *AND income*(*X*, “high”) → *class*(*X*, “A”)
age(*X*, “youth”) *AND income*(*X*, “low”) → *class*(*X*, “B”)
age(*X*, “middle_aged”) → *class*(*X*, “C”)
age(*X*, “senior”) → *class*(*X*, “C”)

Figure 1.9 A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

Regression

- Regression analysis is a statistical methodology that is most often used for numeric prediction,
- Regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels

Cluster Analysis

- clustering analyzes data objects without consulting class labels.
- The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity

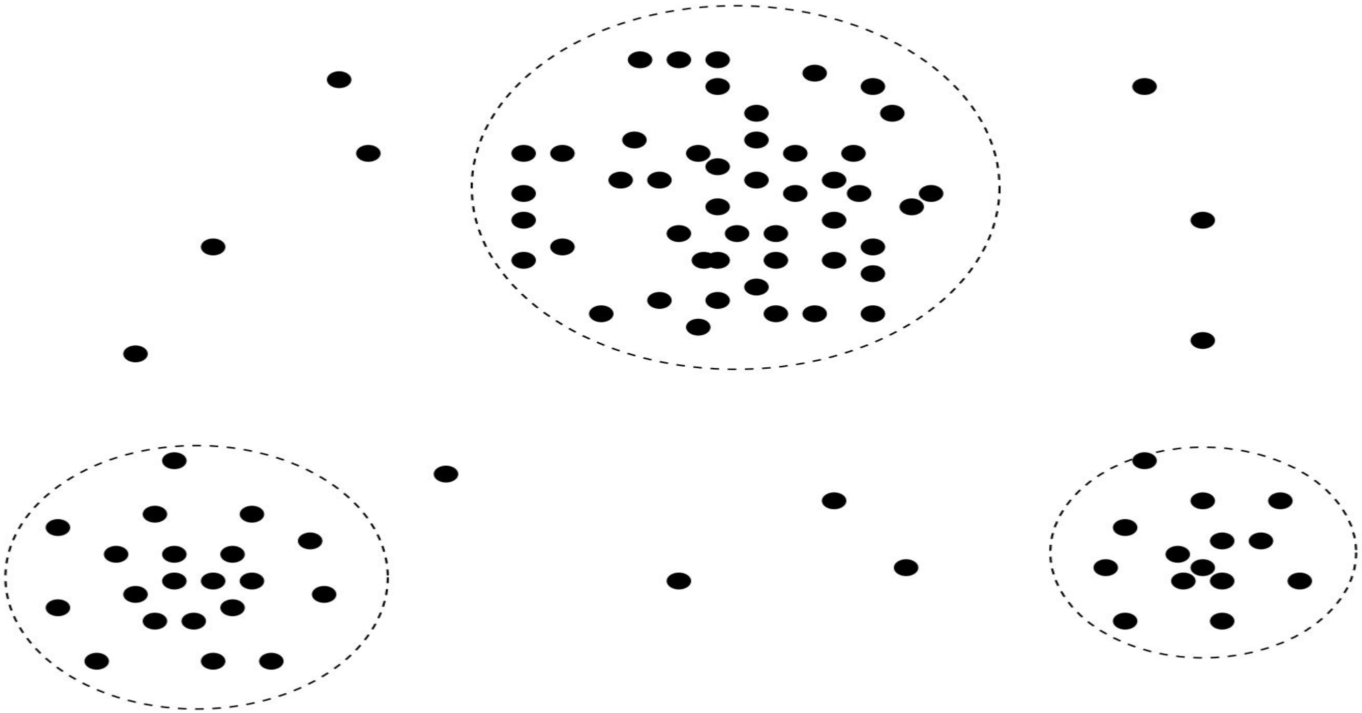


Figure 1.10 A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters

Outlier Analysis

- A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are outliers.
- The analysis of outlier data is referred to as outlier analysis or anomaly mining.