Data Mining

Part 4. Classification and Prediction

4.1 Introduction

Fall 2009

Instructor: Dr. Masoud Yaghini

Outline

- Classification vs. Prediction
- Classification Process
- Data Preparation
- Comparing Classification Methods
- References

Classification vs. Prediction

Classification vs. Prediction

- Prediction problems
  - predict future data trends
- Major types:
  - Classification
  - Numeric prediction

Classification vs. Prediction

- **Classification**
  - a model or **classifier** to predict **categorical labels** (discrete or nominal)
  - The ordering among categories has no meaning
  - e.g. such as “safe” or “risky” for the loan application data
  - guess whether a customer with a given profile will buy a new computer.

Classification vs. Prediction

- **Numeric Prediction**
  - a model or **predictor** to predict a **continuous-valued function** or ordered value
  - e.g. predicting how much a given customer will spend during a sale at *AllElectronics*
  - **Regression analysis** is a statistical methodology that is most often used for numeric prediction

Applications

- Typical applications
  - Credit approval
  - Target marketing
  - Medical diagnosis
  - Fraud detection
  - Performance prediction
  - Manufacturing

Classification

- Techniques for data classification:
  - Decision tree classifiers
  - Bayesian classifiers
  - Bayesian belief networks
  - Rule-based classifiers
  - Backpropagation (a neural network technique)
  - Support vector machines
  - K-nearest-neighbor classifiers
  - Case-based reasoning
  - Genetic algorithms

Classification Process

Classification—A Two-Step Process

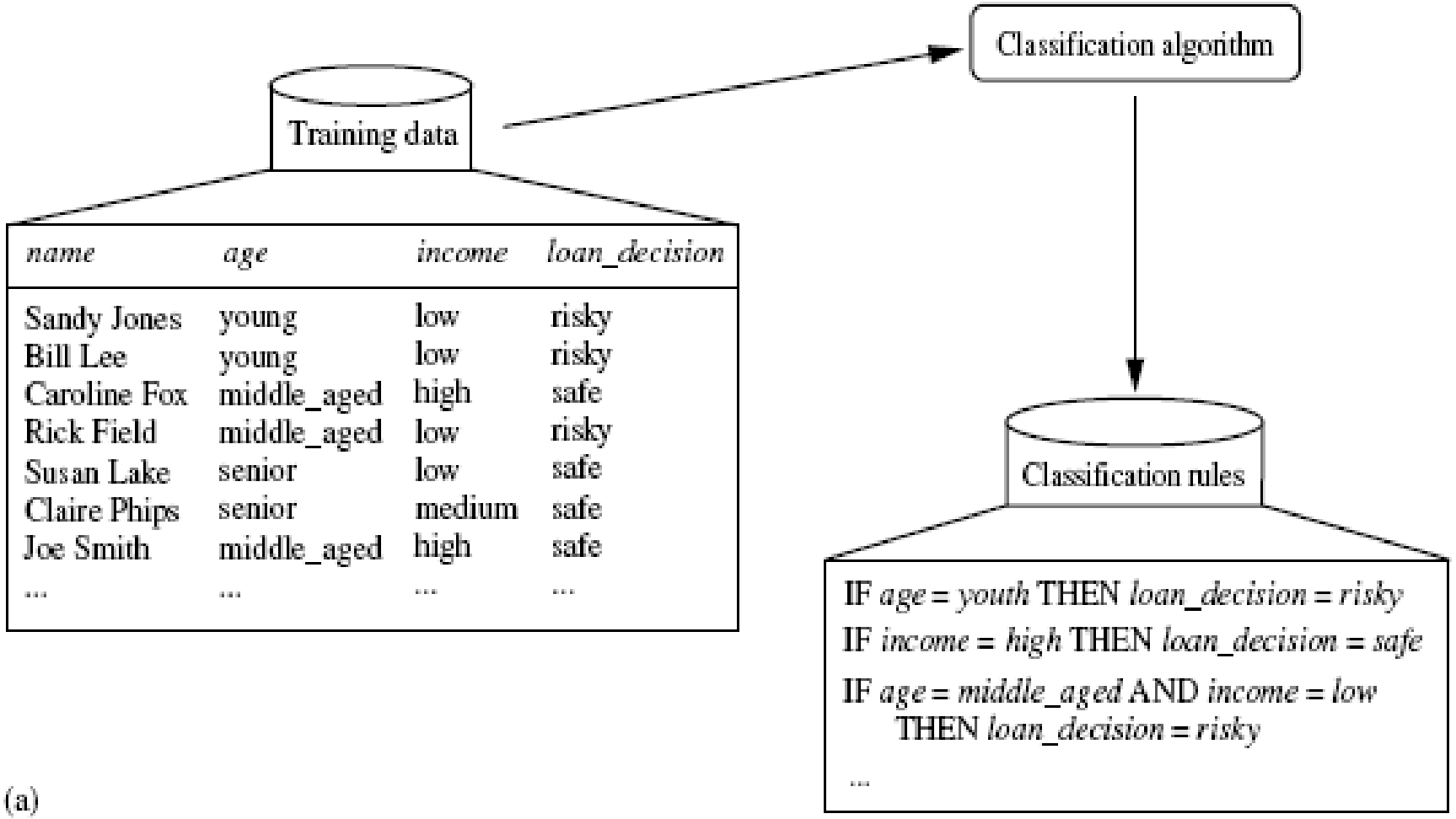
- Data classification is a two-step process:
  - **Learning step or model construction**
  - **Model usage**

Model Construction

- **Learning step (model construction)**
  - A classification algorithm builds the classifier by analyzing or “learning from” a training set
  - Each instance is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of instances used for model construction is **training set**
  - The model is represented as classification rules, decision trees, or mathematical formula
  - **Data instances** can be referred to as samples, examples, instances, data points, objects, or data tuples

Model Construction

- Example: identify loan applications as being either safe or risky

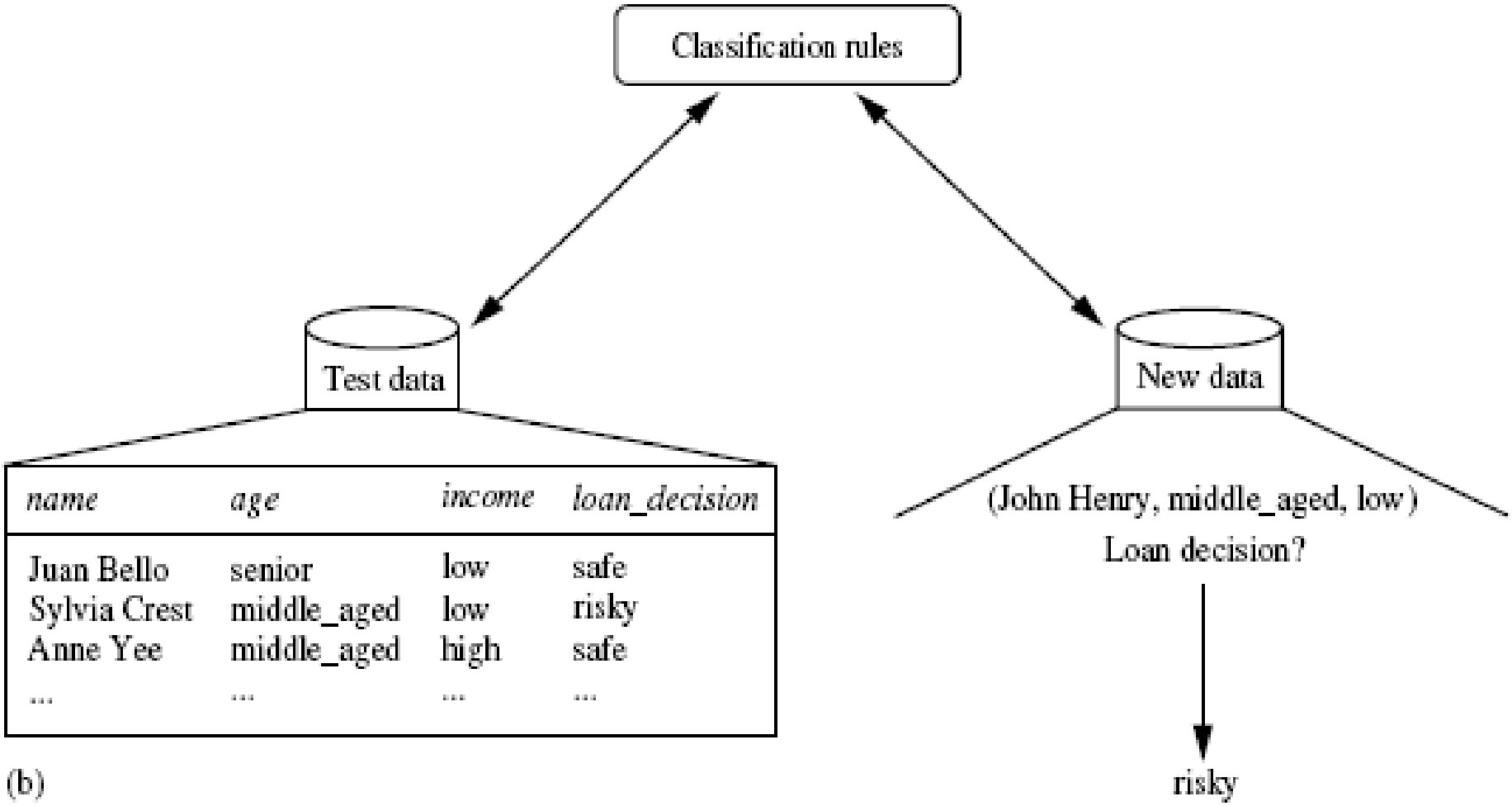


name	*age*	*income*	*loan_decision*
Sandy Jones	young	low	risky
Bill Lee	young	low	risky
Caroline Fox	middle_aged	high	safe
Rick Field	middle_aged	low	risky
Susan Lake	senior	low	safe
Claire Phips	senior	medium	safe
Joe Smith	middle_aged	high	safe
...	...	...	...

(a)

Model Usage

- **Model usage:** for classifying future or unknown objects



Estimate Accuracy

- **Estimate accuracy** of the model
  - The known label of test sample is compared with the classified result from the model
  - Accuracy rate is the percentage of **test set** samples that are correctly classified by the model
  - The test instances are randomly selected from the general data set
  - Test set is independent of training set, otherwise **over-fitting** will occur
  - If the accuracy is acceptable, use the model **to classify** new data instances whose class labels are not known

Numeric Prediction

- Numeric prediction is a two step process, similar to that of classification
- The attribute for which values are being predicted is continuous-valued (ordered) rather than categorical (discrete-valued and unordered).
  - This attribute can be referred to simply as the predicted attribute.
- Example:
  - We want to predict the amount (in dollars) that would be “safe” for the bank to loan an applicant.
  - We use the continuous-valued *loan_amount* as the predicted attribute, and build a predictor for our task.

Supervised vs. Unsupervised Learning

- **Supervised learning**
  - The class label of each training instance is known
  - Learning step is also known as **supervised learning**
  - i.e., the learning of the classifier is “supervised” in that it is told to which class each training instance belongs
- **Unsupervised learning (clustering)**
  - The class label of each training instance is not known
  - The number or set of classes to be learned may not be known in advance
  - the aim of establishing the existence of classes or clusters in the data

Data Preparation

Data Preparation

- The preprocessing may be applied to the data to help improve the accuracy, efficiency, and scalability of the classification process.
- The preprocessing steps:
  - Data cleaning
  - Relevance analysis
  - Data transformation

Data Cleaning

- **Data cleaning**
  - to remove or reduce *noisy values*
  - the treatment of *missing values*

- Although most classification algorithms have some mechanisms for handling noisy or missing data, this step can help reduce confusion during learning.

Relevance Analysis (feature selection)

- **Relevance analysis:**
  - **Correlation analysis**
    - to detect redundant attributes
    - Correlation analysis can be used to identify whether any two given attributes are statistically related.
    - For example, a strong correlation between attributes *A1* and *A2* would suggest that one of the two could be removed from further analysis.
  - **Attribute subset selection**
    - to remove irrelevant attributes
    - to find a reduced set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.

Data Transformation

- Data transformation
  - Normalization
  - Discretization
  - Generalization

Data Transformation

- **Normalization**
  - The normalization is used particularly when methods involving distance measurements are used in the learning step.
  - The values a given attribute fall within a small specified range, such as -1.0 to 1.0, or 0.0 to 1.0.
  - Normalization would prevent attributes with initially large ranges (like income) from outweighing attributes with initially smaller ranges (such as binary attributes).

Data Transformation

- **Discretization**
  - For example, numeric values for the attribute income can be generalized to discrete ranges, such as *low, medium,* and *high. Similarly, categorical* attributes
- **Generalization**
  - The data can also be transformed to higher-level concepts.
  - Example: like street, can be generalized to higher-level concepts, like city.
  - Because generalization compresses the original training data, fewer input/output operations may be involved during learning.

Comparing Classification and Prediction Methods

Comparing Methods

- Classification and prediction methods can be compared and evaluated according to the following criteria:
- **Accuracy**
  - the ability of a given classifier to correctly predict the class label of new data
- **Speed**
  - time to construct the model (training time)
  - time to use the model (classification/prediction time)
- **Robustness**
  - the ability of the classifier to make correct predictions given noisy data or data with missing values.

Comparing Methods

- **Scalability**
  - The ability to construct the classifier or predictor efficiently given large amounts of data.
- **Interpretability**
  - the level of understanding and insight that is provided by the classifier.

References

References

- J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Elsevier Inc. (2006). (Chapter 6)

The end