Data Mining

3.1 Classification and Prediction

Fall 2008

Instructor: Dr. Masoud Yaghini

Outline

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

Classification vs. Prediction

Classification vs. Prediction

- **Classification**
  - a model or **classifier** is constructed to predict **categorical labels** (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- **Prediction**
  - A model or **predictor** is constructed to predict a continuous-valued function or ordered value, i.e., predicts unknown or missing values

Classification vs. Prediction

- 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
  - Rough sets
  - Fuzzy logic techniques

Classification Process

Classification—A Two-Step Process

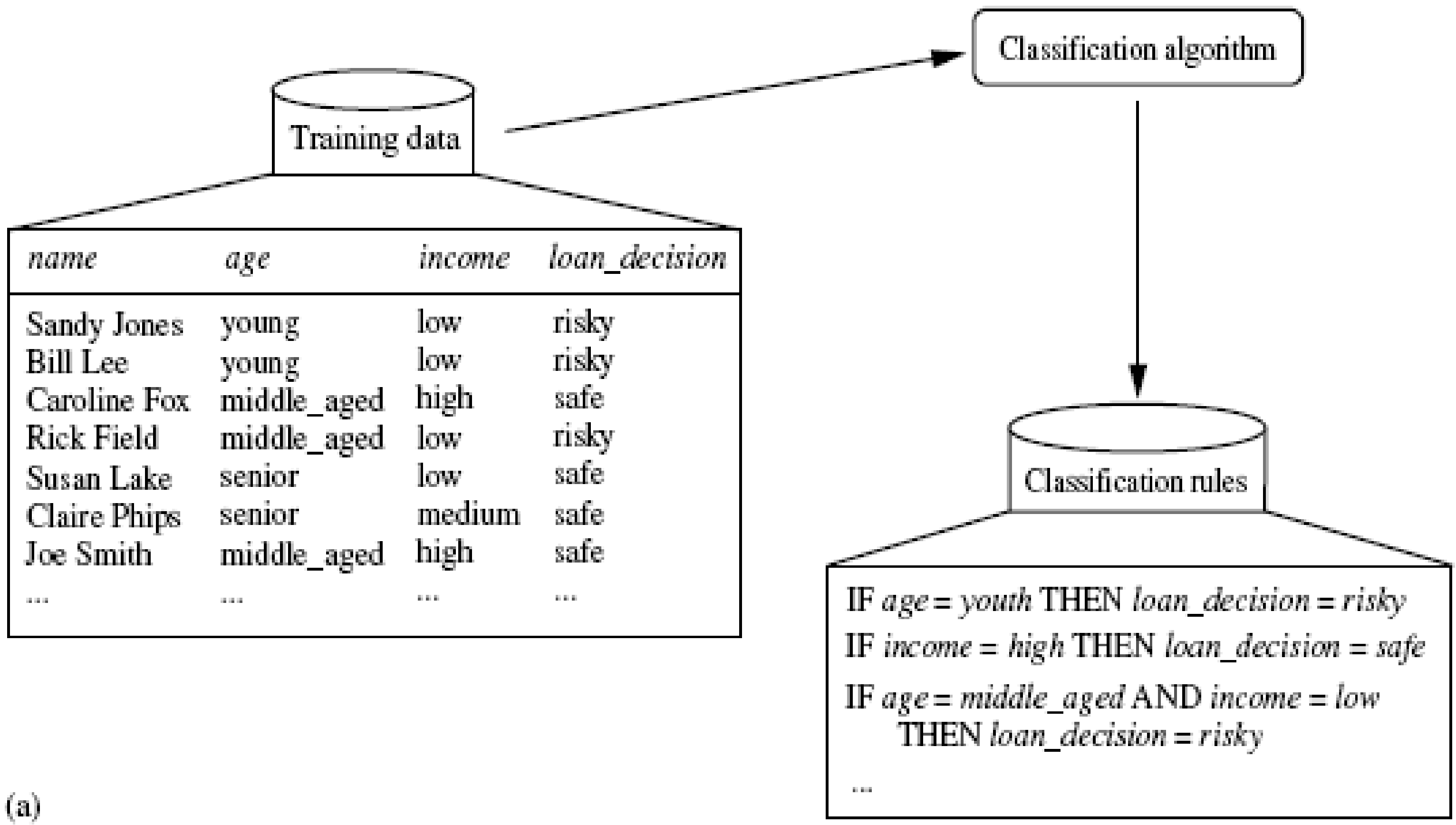
- Data classification is a two-step process:
  - **Model construction or learning step**
  - **Model usage**

Classification—Model construction

- **Model construction (learning step)**: describing a set of predetermined classes
  – Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  – The set of tuples used for model construction is **training set**
  – The model is represented as classification rules, decision trees, or mathematical formulae
  – **Data tuples** can be referred to as samples, examples, instances, data points, or objects (in Han's book)

Classification—Model construction

- The loan application example

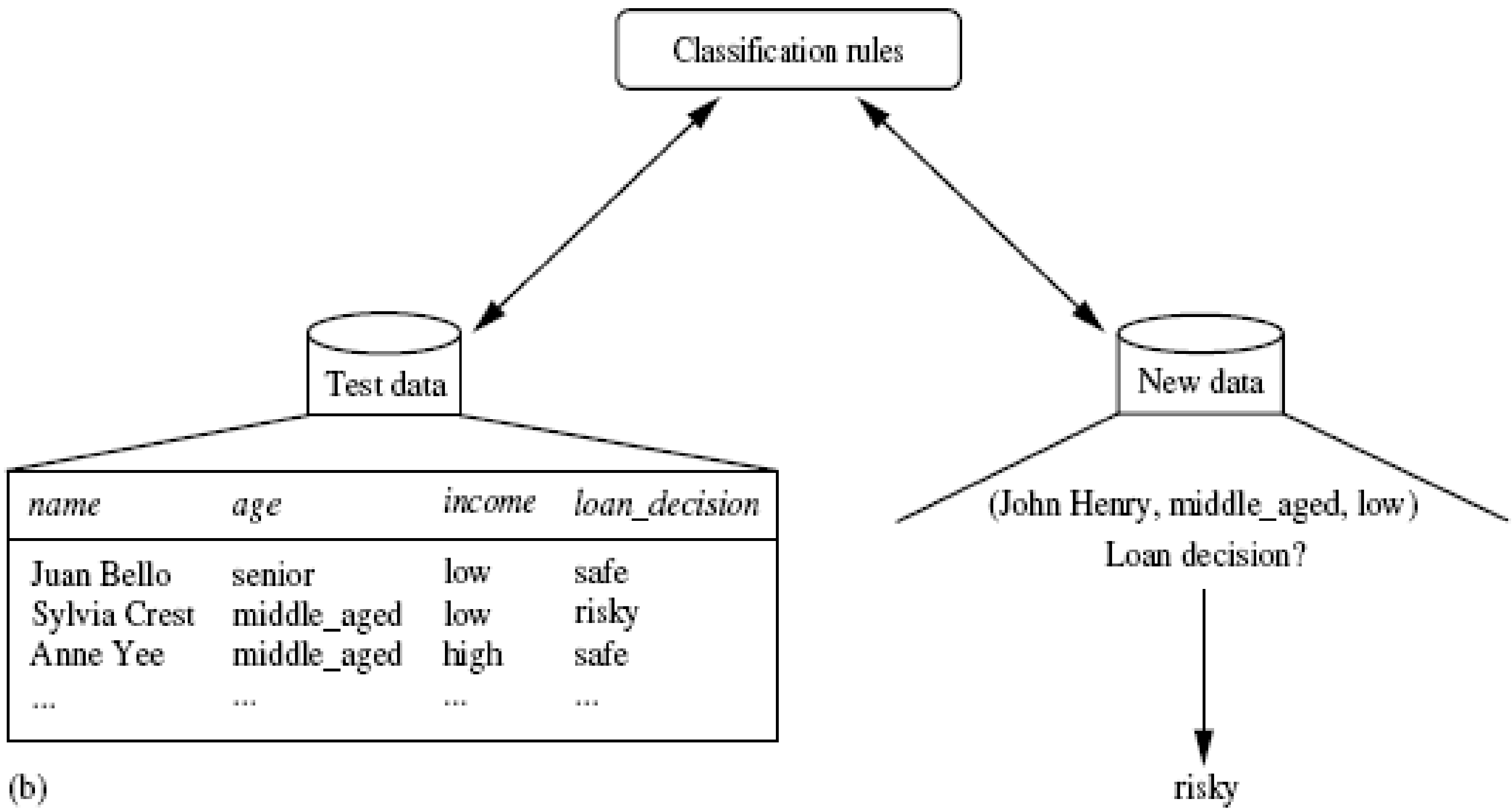


Classification—Model usage

- **Model usage**: for classifying future or unknown objects
  - **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 tuples 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** data tuples whose class labels are not known

Classification—Model usage

- The loan application example



Supervised vs. Unsupervised Learning

- **Supervised learning (classification)**
  - The class label of each training tuple is known
  - New data is classified based on the training set
- **Unsupervised learning (clustering)**
  - The class labels of training tuple is unknown
  - The number or set of classes to be learned may not be known in advance
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Data Preparation

Data Preparation

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

Data Cleaning

- Preprocess data in order to remove or to reduce noise
- Handle missing values
  - e.g., by replacing a missing value with the most commonly occurring value for that attribute, or with the most probable value based on statistics
- 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 can be used to detect attributes that do not contribute to the classification
- Including such attributes may otherwise slow down, and possibly mislead, the learning step.
- Relevance analysis (feature selection) includes:
  - **Correlation analysis:** to remove redundant attributes
  - **Attribute subset selection:** to remove irrelevant attributes

Relevance Analysis (feature selection)

- **Correlation analysis:**
  - Many of the attributes in the data may be **redundant**.
  - 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.

Relevance Analysis (feature selection)

- **Attribute subset selection**:
  - A database may also contain **irrelevant attributes***.*
  - Attribute subset selection can be used 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
  - Generalization

Data Transformation

- **Normalization**:
  - The data may be transformed by normalization, particularly when methods involving distance measurements are used in the learning step.
  - **Normalization** involves scaling all values for a given attribute so that they fall within a small specified range, such as -1.0 to 1.0, or 0.0 to 1.0.
  - this would prevent attributes with initially large ranges (like, say, income) from outweighing attributes with initially smaller ranges (such as binary attributes).

Data Transformation

- **Generalization:**
  - The data can also be transformed by generalizing it to higher-level concepts*.*
  - This is particularly useful for continuous-valued attributes.
  - For example,
    - numeric values for the attribute income can be generalized to discrete ranges, such as *low, medium,* and *high. Similarly, categorical* attributes,
    - 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 Methods

Comparing Classification 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 Classification 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