Khai Phá Dữ Liệu

Nguyễn Nhật Quang

quangnn-fit@mail.hut.edu.vn

Viện Công nghệ Thông tin và Truyền thông

Trường Đại học Bách Khoa Hà Nội

Năm học 2010-2011

Nội dung môn học:

- Giới thiệu về Khai phá dữ liệu
- **Giới thiệu về công cụ WEKA**
- Tiền xử lý dữ liệu
- Phát hiện các luật kết hợp
- Các kỹ thuật phân lớp và dự đoán
- Các kỹ thuật phân nhóm

WEKA – Giới thiệu

- WEKA là một công cụ phần mềm viết bằng Java, phục vụ lĩnh vực học máy và khai phá dữ liệu
- Các tính năng chính
  - Một tập các công cụ tiền xử lý dữ liệu, các giải thuật học máy, khai phá dữ liệu, và các phương pháp thí nghiệm đánh giá
  - Giao diện đồ họa (gồm cả tính năng hiển thị hóa dữ liệu)
  - Môi trường cho phép so sánh các giải thuật học máy và khai phá dữ liệu
- Có thể tải về từ địa chỉ:

  http://www.cs.waikato.ac.nz/ml/weka/

WEKA – Các môi trường chính

- Simple CLI

  Giao diện đơn giản kiểu dòng lệnh (như MS-DOS)

- **Explorer** (chúng ta sẽ chủ yếu sử dụng môi trường này!)

  Môi trường cho phép sử dụng tất cả các khả năng của WEKA để khám phá dữ liệu

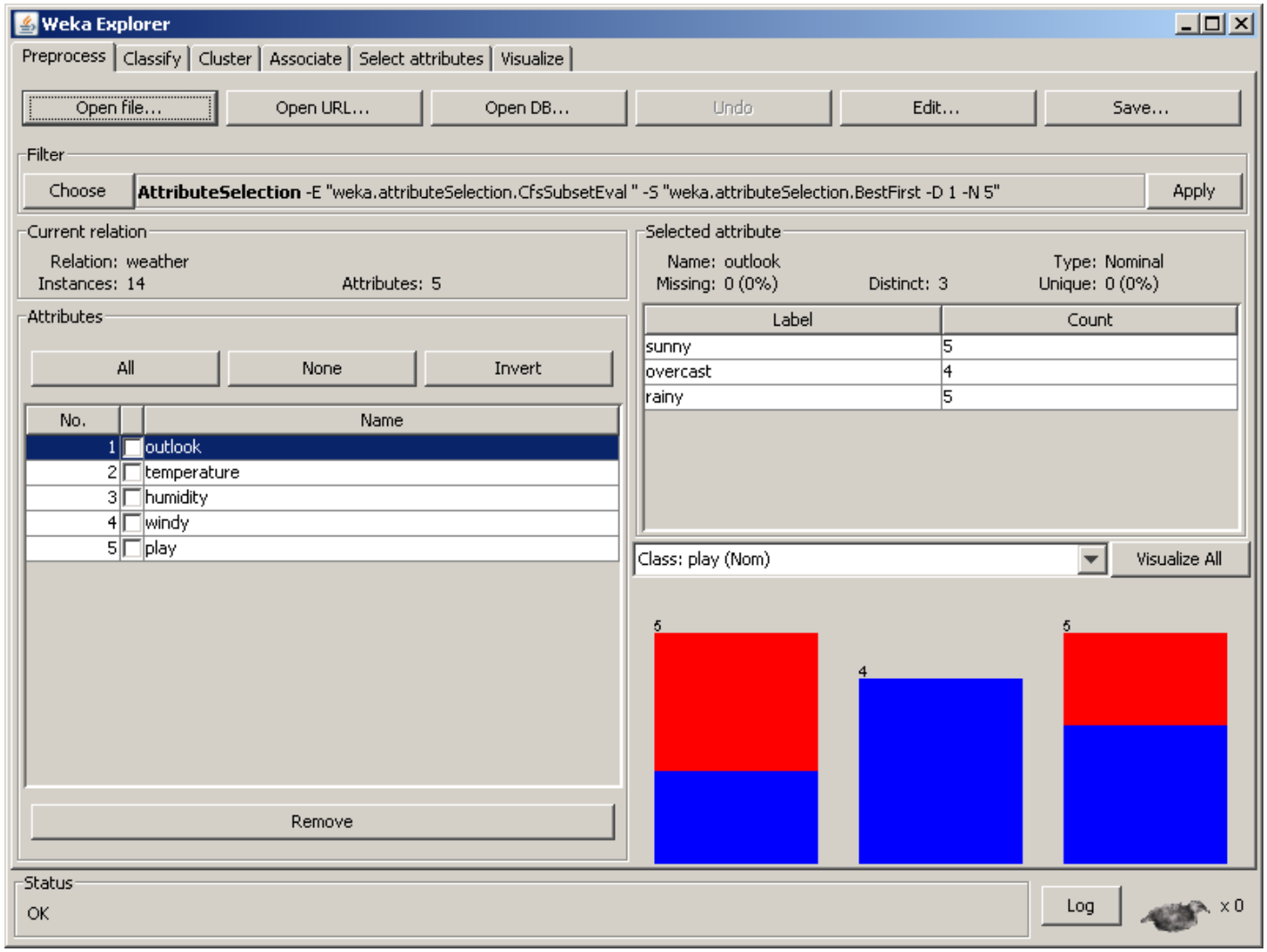
- Experimenter

  Môi trường cho phép tiến hành các thí nghiệm và thực hiện các kiểm tra thống kê (statistical tests) giữa các mô hình học máy

- KnowledgeFlow

  Môi trường cho phép bạn tương tác đồ họa kiểu kéo/thả để thiết kế các bước (các thành phần) của một thí nghiệm

WEKA – Môi trường Explorer



WEKA – Môi trường Explorer

- **`Preprocess`**

  Để chọn và thay đổi (xử lý) dữ liệu làm việc
- **`Classify`**

  Để huấn luyện và kiểm tra các mô hình học máy (phân loại, hoặc hồi quy/dự đoán)
- **`Cluster`**

  Để học các nhóm từ dữ liệu (phân cụm)
- **`Associate`**

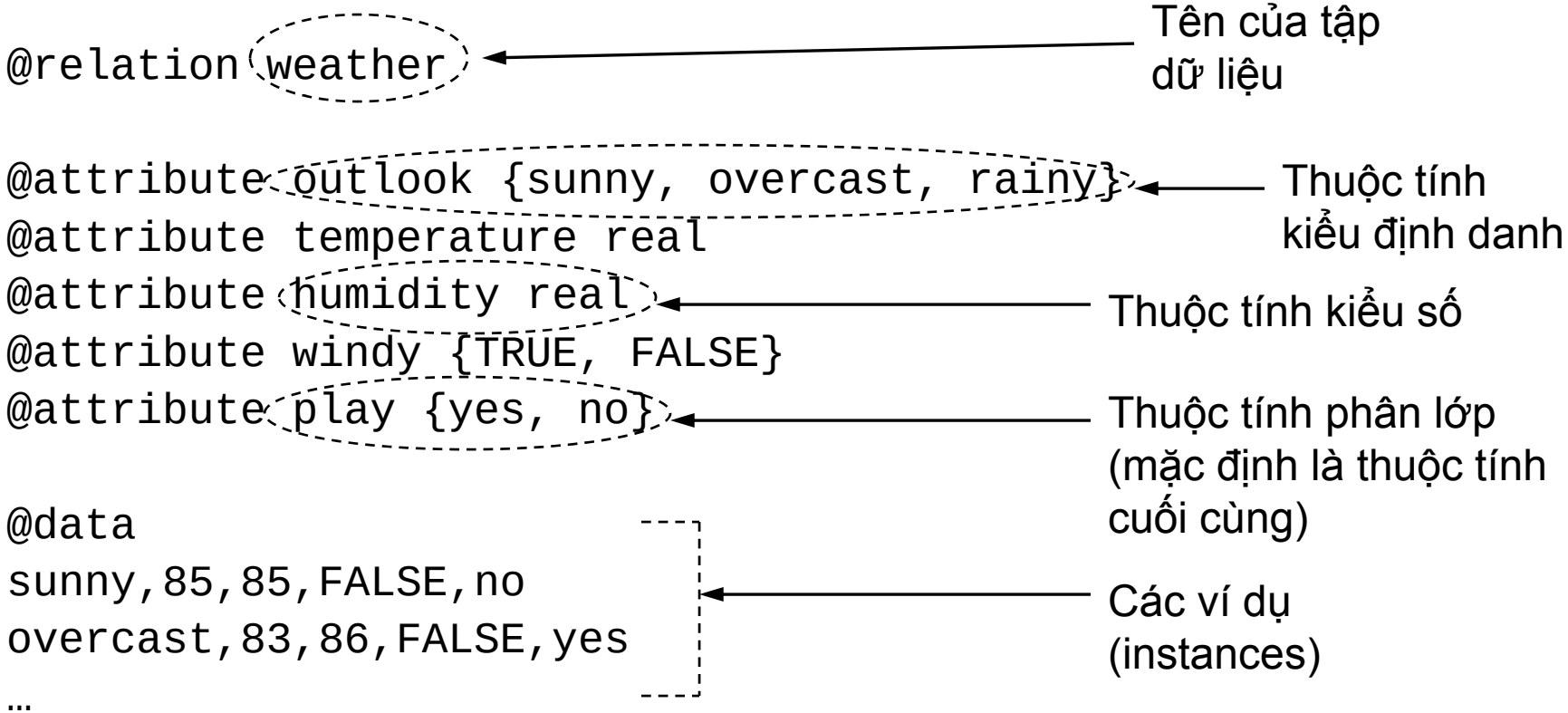
  Để khám phá các luật kết hợp từ dữ liệu
- **`Select attributes`**

  Để xác định và lựa chọn các thuộc tính liên quan (quan trọng) nhất của dữ liệu
- **`Visualize`**

  Để xem (hiển thị) biểu đồ tương tác 2 chiều đối với dữ liệu

WEKA – Khuôn dạng của tập dữ liệu

- WEKA chỉ làm việc với các tập tin văn bản (text) có khuôn dạng *ARFF*
- Ví dụ của một tập dữ liệu



WEKA Explorer: Tiền xử lý dữ liệu

- Dữ liệu có thể được nhập vào (imported) từ một tập tin có khuôn dạng: ARFF, CSV
- Dữ liệu cũng có thể được đọc vào từ một địa chỉ URL, hoặc từ một cơ sở dữ liệu thông qua JDBC
- Các công cụ tiền xử lý dữ liệu của WEKA được gọi là ***`filters`***
  - Rời rạc hóa (Discretization)
  - Chuẩn hóa (Normalization)
  - Lấy mẫu (Re-sampling)
  - Lựa chọn thuộc tính (Attribute selection)
  - Chuyển đổi (Transforming) và kết hợp (Combining) các thuộc tính
  - …

→*Hãy xem giao diện của WEKA Explorer…*

WEKA Explorer: Các bộ phân lớp (1)

- Các bộ phân lớp (Classifiers) của WEKA tương ứng với các mô hình dự đoán các đại lượng kiểu định danh (phân lớp) hoặc các đại lượng kiểu số (hồi quy/dự đoán)
- Các kỹ thuật phân lớp được hỗ trợ bởi WEKA
  - Naïve Bayes classifier and Bayesian networks
  - Decision trees
  - Instance-based classifiers
  - Support vector machines
  - Neural networks
  - …

→ *Hãy xem giao diện của WEKA Explorer…*

WEKA Explorer: Các bộ phân lớp (2)

- Lựa chọn một bộ phân lớp (classifier)
- Lựa chọn các tùy chọn cho việc kiểm tra (test options)
  - **`Use training set`**. Bộ phân loại học được sẽ được đánh giá trên tập học
  - **`Supplied test set`**. Sử dụng một tập dữ liệu khác (với tập học) để cho việc đánh giá
  - **`Cross-validation`**. Tập dữ liệu sẽ được chia đều thành k tập (folds) có kích thước xấp xỉ nhau, và bộ phân loại học được sẽ được đánh giá bởi phương pháp *cross-validation*
  - **`Percentage split`**. Chỉ định tỷ lệ phân chia tập dữ liệu đối với việc đánh giá

WEKA Explorer: Các bộ phân lớp (3)

- **More options…**
  - **Output model**. Hiển thị bộ phân lớp học được
  - **Output per-class stats**. Hiển thị các thông tin thống kê về precision/recall đối với mỗi lớp
  - **Output entropy evaluation measures**. Hiển thị đánh giá độ hỗn tạp (entropy) của tập dữ liệu
  - **Output confusion matrix**. Hiển thị thông tin về ma trận lỗi phân lớp (confusion matrix) đối với phân lớp học được
  - **Store predictions for visualization**. Các dự đoán của bộ phân lớp được lưu lại trong bộ nhớ, để có thể được hiển thị sau đó
  - **Output predictions**. Hiển thị chi tiết các dự đoán đối với tập kiểm tra
  - **Cost-sensitive evaluation**. Các lỗi (của bộ phân lớp) được xác định dựa trên ma trận chi phí (cost matrix) chỉ định
  - **Random seed for XVal / % Split**. Chỉ định giá trị *random seed* được sử dụng cho quá trình lựa chọn ngẫu nhiên các ví dụ cho tập kiểm tra

WEKA Explorer: Các bộ phân lớp (4)

- **`Classifier output`** hiển thị các thông tin quan trọng
  - **Run information**. Các tùy chọn đối với mô hình học, tên của tập dữ liệu, số lượng các ví dụ, các thuộc tính, và f.f. thí nghiệm
  - **Classifier model (full training set)**. Biểu diễn (dạng text) của bộ phân lớp học được
  - **Predictions on test data**. Thông tin chi tiết về các dự đoán của bộ phân lớp đối với tập kiểm tra
  - **Summary**. Các thống kê về mức độ chính xác của bộ phân lớp, đối với f.f. thí nghiệm đã chọn
  - **Detailed Accuracy By Class**. Thông tin chi tiết về mức độ chính xác của bộ phân lớp đối với mỗi lớp
  - **Confusion Matrix**. Các thành phần của ma trận này thể hiện số lượng các ví dụ kiểm tra (test instances) được phân lớp đúng và bị phân lớp sai

WEKA Explorer: Các bộ phân lớp (5)

- **`Result list`** cung cấp một số chức năng hữu ích
  - **Save model**. Lưu lại mô hình tương ứng với bộ phân lớp học được vào trong một tập tin nhị phân (binary file)
  - **Load model**. Đọc lại một mô hình đã được học trước đó từ một tập tin nhị phân
  - **Re-evaluate model on current test set**. Đánh giá một mô hình (bộ phân lớp) học được trước đó đối với tập kiểm tra (test set) hiện tại
  - **Visualize classifier errors**. Hiển thị cửa sổ biểu đồ thể hiện các kết quả của việc phân lớp

    Các ví dụ được phân lớp chính xác sẽ được biểu diễn bằng ký hiệu bởi dấu chéo (x), còn các ví dụ bị phân lớp sai sẽ được biểu diễn bằng ký hiệu ô vuông (□)
  - …

WEKA Explorer: Các bộ phân cụm (1)

- Các bộ phân cụm (Cluster builders) của WEKA tương ứng với các mô hình tìm các nhóm của các ví dụ tương tự đối với một tập dữ liệu
- Các kỹ thuật phân cụm được hỗ trợ bởi WEKA
  - Expectation maximization (EM)
  - k-Means
  - ...
- Các bộ phân cụm có thể được hiển thị kết quả và so sánh với các cụm (lớp) thực tế

→*Hãy xem giao diện của WEKA Explorer …*

WEKA Explorer: Các bộ phân cụm (2)

- Lựa chọn một bộ phân cụm (cluster builder)
- Lựa chọn chế độ phân cụm (cluster mode)
  - **Use training set**. Các cụm học được sẽ được kiểm tra đối với tập học
  - **Supplied test set**. Sử dụng một tập dữ liệu khác để kiểm tra các cụm học được
  - **Percentage split**. Chỉ định tỷ lệ phân chia tập dữ liệu ban đầu cho việc xây dựng tập kiểm tra
  - **Classes to clusters evaluation**. So sánh độ chính xác của các cụm học được đối với các lớp được chỉ định
- Store clusters for visualization

  → Lưu lại các bộ phân lớp trong bộ nhớ, để có thể hiện thị sau đó
- Ignore attributes

  → Lựa chọn các thuộc tính sẽ không tham gia vào quá trình học các cụm

WEKA Explorer: Luật kết hợp

- Lựa chọn một mô hình (giải thuật) phát hiện luật kết hợp
- **`Associator output`** hiển thị các thông tin quan trọng
  - **Run information**. Các tùy chọn đối với mô hình phát hiện luật kết hợp, tên của tập dữ liệu, số lượng các ví dụ, các thuộc tính
  - **Associator model (full training set)**. Biểu diễn (dạng text) của tập các luật kết hợp phát hiện được
    - Độ hỗ trợ tối thiểu (minimum support)
    - Độ tin cậy tối thiểu (minimum confidence)
    - Kích thước của các tập mục thường xuyên (large/frequent itemsets)
    - Liệt kê các luật kết hợp tìm được

→ *Hãy xem giao diện của WEKA Explorer…*

WEKA Explorer: Lựa chọn thuộc tính

- Để xác định những thuộc tính nào là quan trọng nhất
- Trong WEKA, một phương pháp lựa chọn thuộc tính (attribute selection) bao gồm 2 phần:
  - *Attribute Evaluator*. Để xác định một phương pháp đánh giá mức độ phù hợp của các thuộc tính

    Vd: correlation-based, wrapper, information gain, chi-squared,…
  - *Search Method*. Để xác định một phương pháp (thứ tự) xét các thuộc tính

    Vd: best-first, random, exhaustive, ranking,…

→ *Hãy xem giao diện của WEKA Explorer…*

WEKA Explorer: Hiển thị dữ liệu

- Hiển thị dữ liệu rất cần thiết trong thực tế

  Giúp để xác định mức độ khó khăn của bài toán học
- WEKA có thể hiển thị
  - Mỗi thuộc tính riêng lẻ (1-D visualization)
  - Một cặp thuộc tính (2-D visualization)
- Các giá trị (các nhãn) lớp khác nhau sẽ được hiển thị bằng các màu khác nhau
- Thanh trượt **`Jitter`** hỗ trợ việc hiển thị rõ ràng hơn, khi có quá nhiều ví dụ (điểm) tập trung xung quanh một vị trí trên biểu đồ
- Tính năng phóng to/thu nhỏ (bằng cách tăng/giảm giá trị của **`PlotSize`** và **`PointSize`**)

→*Hãy xem giao diện của WEKA Explorer…*