LAB01 – Preprocessing

Mục tiêu của bài tập

- Làm quen với các thao tác cơ bản trong tác vụ tiền xử lý dữ liệu thông qua việc áp dụng các công cụ hỗ trợ được cung cấp bởi phần mềm mã nguồn mở WEKA
- Phát huy kỹ năng lập trình để tự cài đặt các thủ tục tiền xử lý dữ liệu đơn giản.

Quy định

- Thời gian thực hiện: **2 tuần** (xem ngày cụ thể trên Moodle)
- Tổ chức thư mục bài làm: thư mục có tên là ⟨**tên nhóm**⟩ (nếu tên nhóm có dấu và khoảng trắng thì bỏ dấu và viết dính liền), bao gồm các tài liệu sau
    - Báo cáo viết trả lời các câu hỏi tự luận và hướng dẫn sử dụng chương trình tự cài đặt, định dạng **pdf**. Trang đầu tiên ghi rõ thông tin nhóm, tỉ lệ thực hiện bài tập của mỗi thành viên (nếu không ghi, mặc định tỉ lệ tương đương), những phần chưa thực hiện được (để giáo viên tránh chấm sót).
    - Dữ liệu thu được trong quá trình thực nghiệm (nếu có)
    - Mã nguồn của chương trình tự cài đặt. Ngôn ngữ: **Python**. Các ngôn ngữ khác tối đa được 80% điểm.
- Nén thư mục bài làm theo định dạng **zip** hoặc **rar** và nộp theo link Moodle
- **Bài làm cần tuân thủ nghiêm ngặt đặc tả của đề bài, mọi sự khác biệt sẽ không được xét tính điểm.**
- Bài làm được đánh giá trên thang điểm 50 rồi quy đổi về tỉ lệ 20% điểm thực hành.

A – Nội dung thực hiện báo cáo viết (25 điểm)

Dữ liệu thực nghiệm

Bài tập này làm việc trên tập dữ liệu bệnh tim **heart-h.arff** và **heart-c.arff** (tải dữ liệu tại http://prdownloads.sourceforge.net/weka/datasets-UCI.jar , 1.1MB)

Mục tiêu của việc khai thác dữ liệu từ các tập dữ liệu này là để hiểu rõ hơn các nhân tố nguy hiểm cho bệnh tim, cụ thể là ở thuộc tính thứ 14 – num (<50: không có bệnh, từ 50-1 đến 50-4 cho biết các mức tăng của bệnh).

Câu hỏi đặt ra là có thể dự đoán tình trạng bệnh tim của một bệnh nhân từ những dữ kiện sức khỏe khác của người này hay không? Tác vụ khai thác dữ liệu được chọn để trả lời câu hỏi này là phân lớp/dự đoán, và một vài giải thuật khác nhau sẽ được sử dụng để tìm ra phương án cho kết quả dự đoán tốt nhất.

Yêu cầu thực hiện

1. Tích hợp dữ liệu (integration) [1] (5 điểm)

Bước này thực hiện các thao tác cần thiết cho việc hợp nhất hai tập dữ liệu **heart-h.arff** và **heart-c.arff** thành một tập dữ liệu chung.

a. (1đ) Định nghĩa thế nào là tích hợp dữ liệu?
b. (1đ) Vấn đề nhận diện thực thể (*entity identification*) có xảy ra trong hai tập dữ liệu hay không? Nếu có, bạn sẽ giải quyết vấn đề này như thế nào?
c. (1đ) Vấn đề dữ liệu dư thừa (*data redundancy*) có xảy ra trong hai tập dữ liệu hay không? Nếu có, bạn sẽ giải quyết vấn đề này như thế nào?
d. (1đ) Vấn đề mâu thuẫn giá trị dữ liệu (*data value conflicts*) có xảy ra trong hai tập dữ liệu hay không? Nếu có, bạn sẽ giải quyết vấn đề này như thế nào?
e. (1đ) Tích hợp hai tập dữ liệu đã cho thành tập dữ liệu mới có tên là **heart-integration.arff**. Sử dụng WEKA để đọc tập dữ liệu tích hợp. Chụp màn hình cửa sổ Explorer, đánh dấu các vùng trong cửa sổ có thể cho biết số mẫu và số thuộc tính của dữ liệu.

Cần nộp lại tập tin **heart-integration.arff**. Tập tin sai định dạng hoặc thiếu dòng dữ liệu (-1đ). Không nộp tập tin (-2đ).

2. Tóm tắt dữ liệu mô tả (descriptive data summarization) [2] (8.0 điểm)

Một bước quan trọng trong tiền xử lý dữ liệu là làm quen với dữ liệu thông qua các giá trị thống kê, chúng cho ta biết những đặc tính phổ thông của dữ liệu.

[1] Xem [4], phần 2.4.1 – Data Integration
[2] Xem [4], phần 2.2 – Descriptive Data Summarization

Mở tab **Preprocess** của cửa sổ WEKA Explorer, đọc tập dữ liệu **heart-integration.arff**.

a. (1đ) Cho biết giá trị trung bình, độ lệch chuẩn, giá trị nhỏ nhất và giá trị lớn nhất của thuộc tính **age**. Chụp màn hình cửa sổ Explorer, đánh dấu các vùng trong cửa sổ cho biết những thông tin này.
b. (1đ) Xác định **five-number summary** của thuộc tính **age**. Chụp màn hình cửa số Explorer hiển thị thông tin này nếu WEKA có cung cấp. Nếu không, bạn căn cứ vào những giá trị nào khác có trong WEKA để tính?
c. (1đ) Cho biết thuộc tính nào có kiểu thuộc tính số (*numeric*), kiểu rời rạc không có thứ tự (*categorical/nominal*), hoặc kiểu rời rạc có thứ tự (*ordernal*).
d. (1đ) Giải thích ý nghĩa của đồ thị ở góc dưới bên phải của cửa sổ. Bạn gọi tên đồ thị này là gì? Đồ thị biểu diễn điều gì về tập dữ liệu? Màu xanh và màu đỏ có ý nghĩa gì (chú ý các pop-up hiện lên khi chuột di chuyển vào vùng đồ thị)?
e. (1đ) Lần lượt xem xét các thuộc tính khác ngoài thuộc tính age. Chụp màn hình cửa sổ Explorer tương ứng với từng thuộc tính.
f. (1đ) Bạn có nhận xét gì từ những đồ thị trong câu e.?

Bây giờ chúng ta sẽ chuyển sang tab **Visualize**

g. (1đ) Các đồ thị này được gọi tên bằng thuật ngữ gì trong textbook [4]? Chọn **jitter** tối đa, chú ý cột **num** (cột cuối cùng). Bạn cho rằng thuộc tính (Y) nào có khả năng dự đoán tốt nhất về bệnh tim như là một hàm của num (X)? Chụp hình đồ thị của cặp thuộc tính (X) – (Y) này.
h. (1đ) Còn có những cặp thuộc tính nào khác có vẻ tương quan với nhau không?

3. Chọn lọc dữ liệu (selection) [3] (3.0 điểm)

Các tập dữ liệu **heart-h.arff** và **heart-c.arff** khi công bố đã được xử lý bằng cách chọn ra tập hợp các thuộc tính liên quan đến mục tiêu khai thác dữ liệu.

a. (1đ) Dựa vào phần mô tả ở đầu tập tin arff, cho biết có bao nhiêu thuộc tính trong các tập dữ liệu heart-h và heart-c trước khi xử lý?

Sử dụng tab **Select attributes** của cửa số WEKA Explorer

b. (1đ) Giải thích ngắn gọn từng phương pháp chọn lọc thuộc tính trong WEKA.
c. (1đ) So sánh với các phương pháp chọn lọc dữ liệu trong textbook. Phương pháp nào có trong textbook nhưng không có trong WEKA? Phương pháp nào có trong WEKA nhưng không có trong textbook?

[3] Xem [3], mục thay đổi tùy phiên bản – Selecting Attributes

4. Làm sạch dữ liệu (cleaning) [4] (5.0 điểm)

Dữ liệu thực tế thường phải đối diện với tình trạng không đầy đủ, nhiễu hoặc không nhất quán. Chúng ta quay lại tab **Preprocess**, đọc tập dữ liệu tích hợp **heart-integration.arff** và sử dụng các bộ lọc để làm sạch tập dữ liệu này

a. (1đ) Liệt kê các phương pháp đã học trong bài giảng để xử lý vấn đề thiếu giá trị (*missing values*). WEKA hỗ trợ những phương pháp nào cho vấn đề này?
b. (1đ) Liệt kê các phương pháp đã học để loại bỏ dữ liệu nhiễu (*noisy data*). WEKA hỗ trợ những phương pháp nào cho vấn đề này?
c. (1đ) Liệt kê các phương pháp đã học để dò tìm dữ liệu tạp (*outlier detection*). WEKA hỗ trợ những phương pháp nào cho vấn đề này?
d. (1đ) Tập dữ liệu có gặp phải các vấn đề nêu trên hay không? Nếu có, liệt kê một số giá trị đại diện cho từng trường hợp và mô tả lựa chọn của bạn để giải quyết vấn đề (bạn có thể chọn bộ lọc của WEKA hoặc tự đề xuất phương pháp riêng).
e. (1đ) Lưu dữ liệu đã làm sạch vào tập tin **heart-cleaned.arff**. Chụp hình các phần dữ liệu có sự thay đổi trước và sau khi làm sạch.

Cần nộp lại tập tin **heart-cleaned.arff**. Tập tin sai định dạng hoặc thiếu dữ liệu (-1đ). Không nộp tập tin (-2đ).

5. Chuyển đổi dữ liệu (Transformation) [5] (3.0 điểm)

Tìm hiểu các bộ lọc của WEKA hỗ trợ cho vấn đề chuyển đổi dữ liệu. Làm việc trên tập tin **heart-cleaned.arff**

a. (1đ) Bộ lọc nào của WEKA cho phép xây dựng thuộc tính (*attribute construction*), ví dụ, thêm một thuộc tính là tổng của 2 thuộc tính khác?
b. (1đ) Bộ lọc nào của WEKA cho phép chuẩn hóa thuộc tính (*normalization*)? Bộ lọc này có thể chuẩn hóa Min-max, chuẩn hóa Z-score hay chuẩn hóa thập phân không? Nếu có, cho biết cụ thể cách thực hiện những chuẩn hóa này trong WEKA. Nếu không, mô tả giải pháp chuẩn hóa mà WEKA hỗ trợ.
c. (1đ) Chọn một bộ lọc chuẩn hóa trong WEKA và tiến hành chuẩn hóa tất cả các thuộc tính là số thực. Lưu dữ liệu đã chuẩn hóa vào tập tin **heart-normal.arff**. Chụp hình ít nhất 10 dòng dữ liệu với tất cả thuộc tính số thực để thể hiện rõ sự thay đổi sau chuẩn hóa.

Cần nộp lại tập tin **heart-normal.arff**. Tập tin sai định dạng hoặc thiếu dữ liệu (-1đ). Không nộp tập tin (-2đ).

[4] Xem [4], phần 2.3 – Data Cleaning
[5] Xem [4], phần 2.4.2 – Data Transformation

6. Rút gọn dữ liệu (Reduction) [6] (1.0 điểm)

Các tập dữ liệu thường rất lớn, không thể thao tác trực tiếp được, và do đó cần áp dụng các kỹ thuật rút gọn dữ liệu khi tiền xử lý dữ liệu. Một chiến lược khác để rút gọn dữ liệu là chọc lọc các dòng dữ liệu, hay còn gọi là lấy mẫu (*sampling*).

a. (1.0đ) Bộ lọc nào của WEKA cho phép lấy mẫu? Nó có thể thực hiện **Simple Random Sample Without Replacement,** và **Simple Random Sample With Replacement** hay không? Nếu có, cho biết cụ thể cách thực hiện những kỹ thuật này trong WEKA. Nếu không, mô tả giải pháp lấy mẫu mà WEKA hỗ trợ.

Tài liệu tham khảo

[1] Slide bài giảng lý thuyết lý thuyết

[2] Trang chủ của WEKA: http://www.cs.waikato.ac.nz/ml/weka/

[3] Hướng dẫn sử dụng Explorer trong Weka (Moodle)

[4] Textbook: J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition - Chapter 2: Data Preprocessing

[5] I. H. Witten and E. Frank: Data mining, Practical Machine Learning Tools and Techniques

[6] Xem [4], phần 2.5 – Data Reduction

B – Nội dung thực hiện cài đặt

1. Tiền xử lý dữ liệu trên tập dữ liệu tổng quát với một số chức năng đơn giản (15 điểm)

Cài đặt chương trình đọc vào một tập dữ liệu bất kỳ, thực hiện một tác vụ tiền xử lý dữ liệu và xuất ra tập tin kết quả. Chương trình hoạt động theo cơ chế console và các yêu cầu người dùng được đặc tả thông qua tham số dòng lệnh.

- Chương trình nhận đầu vào là một **tập tin CSV (.csv)** và tạo đầu ra cũng là một tập tin CSV. Định dạng tập tin này có thể mở được bằng Microsoft Excel hoặc các text editor thông dụng. Nội dung tập tin có dòng đầu tiên là tên các thuộc tính và các dòng tiếp theo là các mẫu dữ liệu.
- Chương trình hỗ trợ các chức năng
    a) Chuẩn hóa min-max trên danh sách thuộc tính chỉ định.
    b) Chuẩn hóa Z-scores trên danh sách thuộc tính chỉ định.
    c) Rời rạc hóa dữ liệu bằng phương pháp chia giỏ theo độ rộng trên danh sách thuộc tính chỉ định.
    d) Rời rạc hóa dữ liệu bằng phương pháp chia giỏ theo độ sâu trên danh sách thuộc tính chỉ định.
    e) Xóa các mẫu dữ liệu thiếu giá trị trên danh sách thuộc tính chỉ định.
    f) Điền giá trị bị thiếu trên danh sách thuộc tính chỉ định, giá trị được điền là giá trị trung bình (mean) của thuộc tính nếu đó là thuộc tính số hoặc điền giá trị có tần số xuất hiện cao nhất (mode) nếu là thuộc tính rời rạc.
- Cú pháp tham số dòng lệnh do sinh viên tự quy định. Ví dụ gợi ý:
  *chức năng a):
  preprocess --original.csv --output processed.csv --task remove --propList {id, name}
  *chức năng d):
  preprocess --original.csv --output processed.csv --task equalSizeDiscretize --bin 5 --propList {age, salary}
  *chức năng f):
  preprocess -- original.csv --output processed.csv --task removeMissingInstance --propList {age}
- Sinh viên được sử dụng thư viện để đọc/ghi tập tin CSV và xử lý tham số dòng lệnh. **Tất cả các phần còn lại đều phải tự cài đặt.**
- Đặt tên chương trình: **<ID nhóm>_B1.<phần mở rộng>**. Ví dụ: 1_B1.exe

2. Tiền xử lý dữ liệu trên tập dữ liệu cụ thể cho trước (10 điểm)

Bài tập này làm việc trên tập dữ liệu bệnh tim **countries.txt** (đính kèm đề bài). Nội dung tập tin chứa mô tả của các quốc gia, trong đó mỗi quốc gia được mô tả bởi một số thông tin như mã (country), tên (name), tên đầy đủ (longName), ngày thành lập (foundingDate), thủ đô (capital), thành phố lớn nhất (largestCity), dân số (population), diện tích (area).

Một số vấn đề cần lưu ý:

- Dữ liệu rỗng (chỉ có mã (country)).
- Trùng lắp thông tin (hai quốc gia có mã khác nhau nhưng các thông tin khác giống nhau).
- Một số mẫu thiếu dữ liệu trên một vài thuộc tính.
- Diện tích (area) có đơn vị không thống nhất (km: km^2 và mi: $mile^2$)

Cài đặt chương trình chuyển tập tin trên thành tập tin CSV (.csv), trong đó:

1. Xóa các mẫu rỗng.
2. Xóa các mẫu bị trùng lắp
3. Chuyển diện tích về km^2
4. Sử dụng chương trình đã cài đặt ở phần B-1. để xóa các mẫu bị thiếu diện tích.

- Chương trình xuất ra kết quả sau khi thực hiện các bước trên và lưu ở tập tin **<ID nhóm>**_B2.csv. Ví dụ **1**_B2.csv. Cần nộp lại tập tin.
- Cú pháp tham số dòng lệnh do sinh viên tự quy định. Ví dụ gợi ý:

  preprocess --input C:/data/countries.txt --output D:/output/data.csv
- Sinh viên được sử dụng thư viện để đọc/ghi tập tin CSV và xử lý tham số dòng lệnh. **Tất cả các phần còn lại đều phải tự cài đặt.**
- Đặt tên chương trình: **<ID nhóm>_B2.<phần mở rộng>**. Ví dụ: 1_B2.exe