BÀI TẬP THỰC HÀNH 1
TIỀN XỬ LÝ DỮ LIỆU VỚI WEKA

Mục tiêu

SV biết cách sử dụng công cụ khai thác dữ liệu Weka để tiến hành tiền xử lý dữ liệu.

Quy định

- Làm nhóm tối đa 2 người/nhóm.
- Thời hạn: **xem trên Moodle**
- Hình thức: thư mục bài làm có tên là **MSSV1_MSSV2**, bao gồm:
  - Document lưu ở dạng file *.doc(x) hoặc *.pdf: thông tin nhóm, những phần đã hoàn thành, những phần chưa hoàn thành, báo cáo trả lời các câu hỏi
  - Các file *.arff thu được sau các bước tiền xử lý dữ liệu

Giới thiệu

✓ Weka là một công cụ mã nguồn mở viết trên môi trường Java sử dụng trong khai thác dữ liệu, được phát triển bởi Trường Đại Học Waikato ở New Zealand và đã được sử dụng tại IAI Lab. Weka là một công cụ đắc lực cho việc học môn khai thác dữ liệu và ứng dụng bởi tính miễn phí, sinh viên có thể nghiên cứu sự khác biệt khi thực thi những mô hình khai thác dữ liệu khác nhau. Ngoài ra, các kết quả từ Weka có thể được công bố trên các tạp chí hay hội nghị uy tín nhất. Do vậy, Weka được xem là một môi trường phát triển thực tế được lựa chọn để nghiên cứu khai thác dữ liệu.

✓ Download Weka tại: http://www.cs.waikato.ac.nz/ml/weka/

✓ Cách sử dụng Weka: xem hướng dẫn chi tiết trong thư mục cài đặt

Cơ sở dữ liệu bệnh tim

Dataset về bệnh tim được lấy từ UCI repository (datasets-UCI.jar) gồm:

+ ***heart-h.arff***: dữ liệu của người Hung-ga-ri
+ ***heart-c.arff***: dữ liệu của vùng Cleveland

Các dataset này mô tả các thành phần của bệnh tim. Download dataset tại: http://prdownloads.sourceforge.net/weka/datasets-UCI.jar (1.1MB)

Mục tiêu của việc khai thác dữ liệu từ các dataset 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)

Nội Dung

Câu hỏi đặt ra là có thể dự đoán bệnh tim từ những dữ liệu đã biết khác của một bệnh nhân hay không. Tác vụ khai thác dữ liệu được chọn để trả lời câu hỏi này là ***phân lớp/dự đoán***, và một vài thuật toán khác nhau sẽ được sử dụng để tìm ra thuật toán cho kết quả dự đoán tốt nhất.

1. Chuẩn bị dữ liệu – Tích hợp dữ liệu (integration) [1]

Bước này hợp nhất 2 dataset lại thành 1. Bạn hãy cho biết:

a. Định nghĩa sự tích hợp dữ liệu.
b. Có vấn đề về nhận diện thực thể (*entity identification*) trong 2 dataset này hay không? Nếu có, giải quyết như thế nào?
c. Có vấn đề dữ liệu dư thừa (*redundancy*) trong 2 dataset này hay không? Nếu có, giải quyết như thế nào?
d. Có sự mâu thuẫn dữ liệu (*data value conflicts*) trong 2 dataset này hay không? Nếu có, giải quyết như thế nào?
e. Tích hợp 2 dataset này lại thành 1 dataset để chuẩn bị cho các câu hỏi tiếp theo. Nạp dataset sau khi tích hợp vào Explorer. Bạn có bao nhiêu mẫu? Bao nhiêu thuộc tính?
f. Chụp lại màn hình của cửa sổ Explorer của bạn.

2. Tóm tắt mô tả dữ liệu – Descriptive data summarization [2]

Trước khi tiền xử lý dữ liệu, một bước quan trọng là làm quen với dữ liệu

a. Trong **tab Preprocess**, xem xét thuộc tính age và trả lời câu hỏi: trung bình, độ lệch chuẩn, giá trị nhỏ nhất, lớn nhất của nó là gì?
b. Liệt kê ***five-number summary*** của thuộc tính này. Weka có cung cấp những con số này hay không?
c. Cho biết thuộc tính nào là số (*numeric*), thuộc tính nào là có thứ tự (*ordernal*) và thuộc tính nào là rời rạc/danh sách (*categorical*/*nomial*).
d. Giải thích ý nghĩa của đồ thị trong cửa sổ Explorer. Bạn đặt tên cho đồ thị này là gì? Màu xanh và màu đỏ có nghĩa gì (chú ý các pop-up hiện lên khi di chuyển chuột trên đồ thị). Đồ thị này biểu diễn cho cái gì?
e. Lần lượt xem xét các thuộc tính khác của dataset dưới dạng đồ thị. Dán các ảnh chụp màn hình vào bài làm.

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

f. Nhận xét của bạn từ những đồ thị đó?
g. Chuyển sang **tab Visualize**. Thuật ngữ sử dụng trong textbook để đặt tên cho các đồ thị là gì? Chọn ***jitter*** tối đa, chú ý cột ***num*** (cột cuối cùng), theo bạn các thuộc tính nào có vẻ như dẫn đến bệnh tim nhiều nhất? Dán vào bài làm hình ảnh đồ thị của thuộc tính mà bạn cho rằng có khả năng dự đoán bệnh tim tốt nhất (Y) như là một hàm của num(X).
h. Có những cặp thuộc tính khác nhau nào có vẻ như tương quan với nhau không?

3. Chuẩn bị dữ liệu – Chọn lọc dữ liệu (selection) [3]

Các dataset sử dụng trong bài tập đã được xử lý bằng các chọn ra tập các thuộc tính liên quan đến mục tiêu khai thác dữ liệu.

a. Bạn hãy cho biết có bao nhiêu thuộc tính trong những dataset trước khi xử lý?
b. Sử dụng **tab Select attributes**. Liệt kê những lựa chọn khác nhau của Weka để chọn lọc thuộc tính, giải thích ngắn gọn từng phương pháp.
c. So sánh với các phương pháp chọn lọc dữ liệu trong textbook, có phương pháp nào không có trong Weka hay phương pháp nào trong Weka không có trong textbook?

4. Chuẩn bị dữ liệu – Làm sạch dữ liệu (cleaning) [4]

Xử lý các dữ liệu thiếu, nhiễu, và mâu thuẫn. Sử dụng các bộ lọc trong Weka để làm sạch dữ liệu.

a. Các giá trị thiếu (*Missing values*): Liệt kê các phương pháp đã học để xử lý dữ liệu thiếu. Weka đã cài đặt những phương pháp nào? Bạn hãy chọn 1 phương pháp để xử lý giá trị thiếu trong dataset, giải thích tại sao bạn chọn phương pháp đó. Cài đặt 1 phương pháp khác mà bạn thích nếu nó không có trong Weka
b. Dữ liệu nhiễu (*Noisy data*): Liệt kê các phương pháp đã học để loại bỏ các dữ liệu nhiễu, Weka đã cài đặt những phương pháp nào?
c. Dò tìm dữ liệu tạp (*Outlier detection*): Liệt kê các phương pháp đã học để dò tìm dữ liệu tạp. Bạn dò tìm dữ liệu tạp bằng Weka như thế nào? Có dữ liệu tạp trong dataset đã cho hay không? Nếu có, liệt kê một số dữ liệu tạp.
d. Lưu dataset đã làm sạch vào file ***heart-cleaned.arff*** và dán vào bài làm 1 ảnh chụp cho thấy ít nhất 10 dòng của dữ liệu với tất cả các cột.

5. Chuẩn bị dữ liệu – Chuyển đổi dữ liệu (Transformation) [5]

[3] Xem [3], phần 7
[4] Xem [4], phần 2.3 – Data Cleaning

Trong số các kỹ thuật chuyển đổi dữ liệu, sử dụng các bộ lọc của Weka để tìm hiểu các kỹ thuật sau:

a. Xây dựng thuộc tính – *Attribute construction*: ví dụ, thêm một thuộc tính là tổng của 2 thuộc tính khác. Bộ lọc nào của Weka cho phép làm điều này?
b. Chuẩn hóa – *Normalize* một thuộc tính. Bộ lọc nào của Weka cho phép làm điều này? Bộ lọc đó có thể chuẩn hóa Min-max không, chuẩn hóa Z-score hay chuẩn hóa thập phân hay không? Cho biết cụ thể cách thức thực hiện những chuẩn hóa này trong Weka.
c. Chọn 1 phương pháp và tiến hành chuẩn hóa tất cả các thuộc tính là số thực, giải thích sự lựa chọn của bạn.
d. Lưu dataset đã chuẩn hóa vào file ***heart-normal.arff*** và chụp ảnh màn hình cho thấy ít nhất 10 dòng dữ liệu với tất cả các cột.

6. Chuẩn bị dữ liệu – Rút gọn dữ liệu (Reduction) [6]

Các cơ sở dữ liệu thường rất lớn, không thể thao tác trực tiếp được. Các kỹ thuật rút gọn dữ liệu được áp dụng để tiền xử lý dữ liệu. Trong **tab Preprocess**, bên cạnh việc chọn lọc thuộc tính, một phương pháp để rút gọn dữ liệu là chọc lọc các dòng trong một dataset, hay còn gọi là lấy mẫu (*sampling*). Làm cách nào để lấy mẫu với các bộ lọc của Weka? Nó có thể thực hiện 2 phương pháp chính là: **Simple Random Sample Without Replacement,** và **Simple Random Sample With Replacement** hay không?

Tài liệu tham khảo

[1] Slide lý thuyết

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

[3] Hướng dẫn sử dụng Explorer trong Weka

[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

[5] Xem [4], phần 2.4.2 – Data Transformation
[6] Xem [4], phần 2.5 – Data Reduction