Lab 01 - Preprocessing Data

Mining - Term I/2021-2022

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

- Tối đa 2 thành viên/nhóm.
- Thời gian: 3 tuần (xem chi tiết trên Moodle).
- Thư mục bài làm: nếu nhóm có 1 sinh viên thì đặt tên là **<MSSV>**, nếu 2 sinh viên thì đặt tên là **<MSSV1>_<MSSV2>**, bao gồm các nội dung sau:
  - Báo cáo trả lời các câu hỏi tự luận, định dạng **pdf**. Trang đầu tiên ghi thông tin nhóm, tỉ lệ thực hiện của mỗi thành viên và các câu hỏi chưa làm được.
  - Mã nguồn chương trình cài đặt: đặt trong thư mục **Source**, bao gồm các file mã nguồn liên quan trong bài tập lập trình. Ngôn ngữ: **Python 3**. Các ngôn ngữ khác tối đa 80% số điểm.
- Nén thư mục bài làm thành định dạng **zip** rồi nộp trên Moodle.
- Tổng điểm tối đa là 12/10 (2 điểm bonus). Nhóm nào quá 10 điểm sẽ được tính là 10 điểm.

1 Yêu cầu 1: Cài đặt Weka (1 điểm)

Đầu tiên các bạn truy cập trang chủ của Weka để tải và cài đặt ứng dụng tương thích với hệ điều hành của mình: https://www.cs.waikato.ac.nz/ml/weka/downloading.html.
Nếu trong quá trình thực nghiệm với Weka mà gặp lỗi out-of-memory, các bạn tham khảo link sau để giải quyết: https://waikato.github.io/weka-wiki/faqs/OutOfMemoryException/.
Weka nhìn chung dễ sử dụng và có tài liệu hướng dẫn đầy đủ. Giao diện Weka thân thiện và trực quan. Nhà phát triển thông qua các tooltip để trang bị hướng dẫn ngữ cảnh cho những chức năng trong Weka, điều này tiện lợi khi người dùng cần tìm hiểu thêm về các tham số.

Weka có 3 giao diện đồ họa (GUI): Explorer, Experimenter và Knowledge Flow Interface. Bên cạnh đó còn có Command Line Interface (CLI) với Java API (nếu bạn quan tâm đến API, hãy xem tài liệu hướng dẫn tại link https://www.youtube.com/watch?v=q3Gf6kqaJWA).
Trong bài tập thực hành này, sinh viên sẽ sử dụng chức năng Explorer để truy cập vào các phương pháp tiền xử lý dữ liệu được Weka hỗ trợ với tên gọi bộ lọc (filters).
Câu hỏi báo cáo:

- Sau khi cài đặt xong. Sinh viên chụp hình giao diện chức năng Explorer **cùng màn hình desktop** và báo cáo lại ảnh chụp.
- Sinh viên tìm thư mục **data** trong thư mục cài đặt của Weka và mở một tập dữ liệu bất kì (có phần mở rộng là **arff**). Giải thích ý nghĩa các nhóm điều khiển **Current relation**, **Attributes** và **Selected attribute** trong tab **Preprocess**. Giải thích ngắn gọn ý nghĩa 5 tab trong giao diện Explorer của Weka.

2 Yêu cầu 2: Làm quen với Weka (6 điểm)

Trong phần này các bạn sẽ tìm hiểu khái quát cơ sở dữ liệu Ung thư vú (**breast_cancer.arff**). Mục tiêu của việc khai thác dữ liệu từ dataset này là để hiểu rõ hơn các nhân tố nguy hiểm đối với bệnh Ung thư vú.

2.1 Đọc dữ liệu vào Weka (2 điểm)

Khởi động chức năng Weka Explorer và đọc tập dữ liệu **breast_cancer.arff** (có thể tìm thấy trong thư mục **data** của thư mục cài đặt Weka).
Sau khi đọc dữ liệu thành công, quan sát cửa sổ Explorer và trả lời các câu hỏi sau:

1. Tập dữ liệu có bao nhiêu mẫu (instances)?
2. Tập dữ liệu có bao nhiêu thuộc tính (attributes)?
3. Thuộc tính nào được dùng làm lớp (class)? Có thể thay đổi thuộc tính dùng làm lớp hay không? Nếu có thì bằng cách nào?
4. Tìm hiểu chi tiết từng thuộc tính trong khung Attributes và cho biết: có bao nhiêu thuộc tính bị thiếu dữ liệu (missing values)? Thuộc tính nào thiếu dữ liệu ít nhất/nhiều nhất? Trình bày tổng quát các cách để giải quyết vấn đề missing values.

5. 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? Đồ thị này biểu diễn cho cái gì?

Chụp màn hình và đánh dấu những nội dung tương ứng để làm minh chứng.

2.2 Khám phá tập dữ liệu Weather (2 điểm)

Tìm tập dữ liệu **weather.numeric.arff** trong thư mục cài đặt của Weka, sau đó mở bằng giao diện Explorer của Weka.

1. Tập dữ liệu có bao nhiêu thuộc tính? Bao nhiêu mẫu? Phân loại các thuộc tính theo kiểu dữ liệu (categorical/numeric). Thuộc tính nào là lớp?
2. Liệt kê ***five-number summary*** của thuộc tính **temperature** và **humidity**. Weka có cung cấp những giá trị này không?
3. Lần lượt xem xét các thuộc tính khác của dataset dưới dạng đồ thị. Dán các ảnh chụp màn hình vào bài làm.
4. Chuyển sang tab **Visualize**. Thuật ngữ sử dụng trong textbook để đặt tên cho các đồ thị ở đây là gì? Chọn jitter tối đa để thấy tổng quan hơn về phân bố dữ liệu. Theo bạn có những cặp thuộc tính khác nhau nào có vẻ như tương quan với nhau không?

2.3 Khám phá tập dữ liệu Tín dụng Đức (2 điểm)

Tìm tập dữ liệu **credit-g.arff** trong thư mục cài đặt của Weka, sau đó mở bằng giao diện Explorer của Weka.

1. Nội dung của phần ghi chú (comment) trong **credit-g.arff** (khi mở bằng 1 text editor bất kì) nói về điều gì? Tập dữ liệu có bao nhiêu mẫu? Bao nhiêu thuộc tính? Mô tả 5 thuộc tính bất kì (phải vừa có cả thuộc tính rời rạc và thuộc tính liên tục).
2. Tên của thuộc tính lớp là gì? Đánh giá phân bố của các lớp, tức là cân bằng hay lệch về một lớp?
3. 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.
4. Cần sử dụng bộ lọc nào để chọn ra 5 thuộc tính có tương quan cao nhất với thuộc tính lớp? Mô tả các bước làm, kèm theo hình chụp từng bước và kết quả cuối cùng.

3 Cài đặt tiền xử lý dữ liệu (5 điểm)

- Bạn hãy cài đặt một chương trình tiền xử lý dữ liệu đơn giản trên giao diện console. 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 này nhận đầu vào là một tệp CSV với dòng đầu là tên các cột của dataset (ngăn cách nhau bởi dấu phẩy), các dòng sau là các dòng dữ liệu của dataset (trên mỗi dòng các trường cũng được ngăn cách nhau bằng dấu phẩy). Để đơn giản, bạn không cần phải kiểm tra tính hợp lệ của dữ liệu. Tuy nhiên dataset có thể chứa mọi loại dữ liệu và cả giá trị bị thiếu (được quy định bằng chuỗi rỗng). Tùy thuộc vào yêu cầu chức năng mà chương trình có thể xuất kết quả ra màn hình hoặc lưu vào file.
- Chương trình phải hỗ trợ các chức năng sau (mỗi chức năng 0.5 điểm):
  1. Liệt kê các cột bị thiếu dữ liệu.
  2. Đếm số dòng bị thiếu dữ liệu.
  3. Điền giá trị bị thiếu bằng phương pháp mean, median (cho thuộc tính numeric) và mode (cho thuộc tính categorical). Lưu ý: khi tính mean, median hay mode các bạn bỏ qua giá trị bị thiếu.
  4. Xóa các dòng bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước (Ví dụ: xóa các dòng bị thiếu hơn 50% giá trị các thuộc tính).
  5. Xóa các cột bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước (Ví dụ: xóa các cột bị thiếu giá trị thuộc tính ở hơn 50% số mẫu).
  6. Xóa các mẫu bị trùng lặp.
  7. Chuẩn hóa một thuộc tính numeric bằng phương pháp min-max và Z-score.
  8. Tính giá trị biểu thức thuộc tính: ví dụ đối với một tập dữ liệu có chứa 2 thuộc tính $width$ và $height$ thì biểu thức $width * height$ sẽ trả về tập dữ liệu cũ với một thuộc tính mới có giá trị ở mỗi mẫu là tích của thuộc tính *width* và *height* trong mẫu tương ứng, với điều kiện cả 2 giá trị $width$ và $height$ đều không bị thiếu, trong trường hợp bị thiếu thì giá trị biểu thức coi như bị thiếu. Lưu ý: biểu thức có thể có nhiều thuộc tính và nhiều phép toán bao gồm cộng, trừ, nhân, chia.
- Cú pháp tham số dòng lệnh do sinh viên tự quy định. Ví dụ gợi ý:
  - Chức năng 1:

```
python3 list-missing.py data.csv
```

  - Chức năng 3:

```
python3 impute.py data.csv --method=mean --columns length price --out=result.csv
```

- Sau khi cài đặt xong, sinh viên chạy các chức năng của chương trình với bộ dữ liệu **house-prices.csv** được đính kèm rồi báo cáo lại kết quả từng chức năng. Với mỗi chức năng nên đưa ra nhiều trường hợp (ví dụ thử chuẩn hóa nhiều cột khác nhau).
- 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**.
- Nhóm nào trình bày code rõ ràng, dễ hiểu và có chú thích đầy đủ tất cả các chức năng sẽ được **cộng thêm 1 điểm**.

4 Tài liệu tham khảo cho sinh viên

1. Slide lý thuyết.
2. Trang chủ của Weka: http://www.cs.waikato.ac.nz/ml/weka/
3. Textbook: J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition - Chapter 2: Data Preprocessing.
4. Textbook: I. H. Witten and E. Frank: Data mining, Practical Machine Learning Tools and Techniques.

Mọi thắc mắc các bạn gửi mail về trợ giảng **Lê Minh Nhật** (minhnhatvt2@gmail.com) hoặc **Nguyễn Khánh Toàn** (ktoan271199@gmail.com).