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: **3 tuần** (xem ngày cụ thể trên Moodle)
- Tổ chức thư mục bài làm: thư mục có tên là **⟨ID nhóm⟩**, 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 3**. Ngôn ngữ C/C++ 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.

Yêu cầu tiên quyết: Cài đặt Weka

Điều cần làm trước tiên là truy cập trang chủ của Weka để tải về và cài đặt ứng dụng

https://www.cs.waikato.ac.nz/ml/weka/downloading.html

Hãy **chọn phiên bản Weka 3.6** vì nó đã được kiểm thử tính ổn định và hỗ trợ các giải thuật khai thác dữ liệu phổ biến được đề cập trong hầu hết các tài liệu khai thác dữ liệu.

Đôi khi bạn sẽ gặp phải tình trạng Weka ngưng hoạt động do lỗi out-of-memory, đặc biệt là khi xử lý tập dữ liệu lớn hoặc chạy tác vụ tiền xử lý dữ liệu có quá nhiều tính toán. Điều này liên quan đến kích thước heap mặc định được cấp cho Java Virtual Machine (JVM). Hãy tham khảo trong link bên dưới và chọn giải pháp phù hợp với hệ điều hành hiện tại.

https://weka.wikispaces.com/OutOfMemoryException

Giao diện Weka: Chức năng Explorer

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 ba 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 bên dưới).

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).

Sinh viên lưu ý các dòng *in nghiêng màu đỏ* vì đó là những yêu cầu cần phải trình bày vào báo cáo hoặc dữ liệu cần phải nộp để hoàn thành bài tập thực hành.

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

Yêu cầu 1 – Tạo tập tin arff từ tập dữ liệu Course Ratings

Hình 1. trình bày tập dữ liệu Course Ratings (nguồn: Daumé, 2015:11). Tập dữ liệu gồm 20 mẫu phỏng vấn sinh viên, trong mỗi mẫu, sinh viên trả lời một số câu hỏi (tương ứng với các thuộc tính Easy?, AI?, Sys?, Thy?, và Morning?) và đánh giá mức độ yêu thích môn học của mình (Rating, các giá trị {0, +1 ,+2} là thích – "liked" và các giá trị {-2, -1} là ghét – "hated").

Rating	Easy?	AI?	Sys?	Thy?	Morning?
+2	y	y	n	y	n
+2	y	y	n	y	n
+2	n	y	n	n	n
+2	n	n	n	y	n
+2	n	y	y	n	y
+1	y	y	n	n	n
+1	y	y	n	y	n
+1	n	y	n	y	n
0	n	n	n	n	y
0	y	n	n	y	y
0	n	y	n	y	n
0	y	y	y	y	y
-1	y	y	y	n	y
-1	n	n	y	y	n
-1	n	n	y	n	y
-1	y	n	y	n	y
-2	n	n	y	y	n
-2	n	y	y	n	y
-2	y	n	y	n	n
-2	y	n	y	n	y

Hình 1. Tập dữ liệu Course Ratings (nguồn: Daumé, 2015:11).

Chuẩn bị dữ liệu

Sử dụng trình soạn thảo văn bản ASCII bất kỳ để tạo tập tin arff từ tập dữ liệu ở Hình 1.

Đặt tên tập tin là "course_rating.arff" và nộp lại cho giáo viên.

!!!Chú ý: Weka mặc định thuộc tính cuối cùng là thuộc tính lớp, do đó thứ tự khai báo các thuộc tính trong tập tin arff sẽ hơi khác với nội dung thể hiện trong Hình 1.

Đọc dữ liệu vào Explorer

Khởi động chức năng Weka Explorer và đọc tập dữ liệu (tab Preprocess → Open file)

Câu hỏi khởi động

1. *Tập dữ liệu được đọc vào Weka thành công hay không? Nếu có, trả lời câu hỏi tiếp theo. Nếu không, cho biết lỗi gặp phải và bạn đã sửa lỗi đó như thế nào?*
2. *Sau khi đọc dữ liệu thành công, quan sát thông tin thể hiện trên giao diện Explorer và trả lời những câu hỏi sau đây*

- *Tên của mối quan hệ (relation) trong dữ liệu là gì?*
- *Tập dữ liệu có bao nhiêu mẫu (instances)?*
- *Tập dữ liệu có bao nhiêu thuộc tính (attributes)?*
- *Thuộc tính nào trong tập dữ liệu là thuộc tính lớp (class)?*

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

3. *Bạn có nhận thấy điều gì đáng chú ý khi quan sát các thông tin thống kê và đồ thị trình diễn?*
4. *Điều gì xảy ra nếu thuộc tính lớp là thuộc tính đầu tiên bên trái? So sánh nội dung các đồ thị trình diễn do chức năng Visualize All (nằm góc dưới bên phải trong tab Preprocess) cung cấp, trong trường hợp thuộc tính lớp là cột đầu tiên và trong trường hợp thuộc tính lớp là cột cuối cùng. Lưu ý rằng tại bước này ta chỉ quan sát thông tin trình diễn dữ liệu chứ không làm bất kỳ thao tác gì ảnh hưởng đến nội dung dữ liệu.*

Sau khi làm quen với chức năng trình diễn dữ liệu, hãy thử các thao tác sau: xóa (deletion), đảo ngược (inversion), hủy thao tác để trở về trạng thái trước đó (undoing), và chỉnh sửa (editing). Bạn có thể xóa một thuộc tính bằng cách nhấn chọn checkbox và nhấn nút **Remove**. **All** giúp chọn mọi thuộc tính, **None** để không chọn thuộc tính nào, và **Invert** đảo ngược lựa chọn hiện hành. Bạn có thể hủy thao tác đã làm bằng cách nhấn nút **Undo**. Nút **Edit** mở trình soạn thảo cho phép bạn khảo sát dữ liệu, tìm giá trị và chỉnh sửa chúng, và xóa mẫu hay thuộc tính. Các hướng dẫn ngữ cảnh tương ứng sẽ hiện ra khi bạn nhấp chuột phải vào giá trị hoặc tiêu đề cột.

5. *Với mỗi thao tác nêu trên, tùy ý thực hiện minh họa và chụp màn hình làm minh chứng.*

Yêu cầu 2 – Khảo sát tập dữ liệu Weather (thuộc tính rời rạc)

Tải tập dữ liệu Weather tại link bên dưới hoặc tìm trong thư mục cài đặt Weka

http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/weather.nominal.arff

Đọc dữ liệu vào Explorer. Nhấn nút Edit từ dãy nút trên đầu tab Preprocess → Cửa số Viewer được mở và nội dung trong cửa sổ hiển thị mọi mẫu trong tập dữ liệu.

6. *Tập dữ liệu có bao nhiêu mẫu? Bao nhiêu thuộc tính? Tên của các thuộc tính này là gì? Các thuộc tính này có loại gì? Thuộc tính nào là lớp?*

   Nhấn vào tên thuộc tính ở bảng **Attributes** bên trái và xem thông tin về **Selected attribute** ở bảng bên phải, chẳng hạn như các giá trị thuộc tính và có bao nhiêu mẫu trong tập dữ liệu mang giá trị thuộc tính cụ thể. Thông tin này cũng được trình diễn bằng biểu đồ histogram ở bên dưới bảng **Selected attribute**.

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

7. *Chức năng của cột đầu tiên ở cửa số Viewer là gì? Lớp của mẫu thứ 8 trong tập dữ liệu là gì?*

Các bộ lọc của Weka

Bây giờ bạn hãy xóa một thuộc tính ra khỏi tập dữ liệu bằng bộ lọc Remove, hay tên đầy đủ là weka.filters.unsupervised.attribute.Remove. Các bộ lọc Weka được tổ chức theo hệ thống phân cấp có gốc là weka. Các phương pháp unsupervised không cần người dùng thiết lập thuộc tính lớp, trong khi phương pháp supervised cần có điều đó. Các bộ lọc được chia nhỏ hơn thành nhóm phương pháp hoạt động chủ yếu trên thuộc tính (attribute) và nhóm dựa trên mẫu (instance). Nhấn nút **Choose** ở tab Preprocess để mở menu chọn bộ lọc và lần theo cấu trúc phân cấp thể hiện trên menu. Khi chọn thành công, dòng văn bản "Remove" sẽ xuất hiện ở trường nằm kế nút Choose. Nhấp chuột vào vùng văn bản này để mở cửa sổ Generic Object Editor , đây là cửa sổ giúp chỉ định tham số cho mọi công cụ trong Weka. Trong trường hợp này, cửa số chứa lời mô tả ngắn về bộ lọc Remove, nhấn nút **More** để đọc mô tả đầy đủ hơn. Nhập giá trị 3 vào trường attributeIndices và nhấn nút OK → cửa sổ tùy chọn đóng. Bây giờ nhấn nút **Apply** ở ngay bên phải để áp dụng bộ lọc lên dữ liệu. Bộ lọc loại bỏ thuộc tính có chỉ mục 3 ra khỏi tập dữ liệu. Lưu ý rằng hành động này chỉ ảnh hưởng đến dữ liệu trong bộ nhớ chứ không tác động đến tập dữ liệu thực sự. Dữ liệu vừa thay đổi có thể được lưu thành tập tin arff bằng cách nhấn nút **Save** và nhập vào tên tập tin. Bạn có thể hủy hành động hiệu chỉnh dữ liệu bằng cách nhấn nút Undo, một lần nữa lưu ý rằng điều này chỉ ảnh hưởng đến dữ liệu trong bộ nhớ.

Đặt tên tập tin là "course_rating_remove3.arff" và nộp lại cho giáo viên.

Những gì mô tả trên đây minh họa cách áp dụng bộ lọc vào dữ liệu. Trong trường hợp cụ thể của Remove, ta còn cách khác đơn giản hơn, đó là thay vì kích hoạt bộ lọc, ta chọn thuộc tính thông qua check box ở bảng Attribute và nhấn nút Remove ở bên dưới danh sách thuộc tính.

8. *Sử dụng bộ lọc weka.unsupervised.instance.RemoveWithValues để loại bỏ mọi mẫu có giá trị thuộc tính humidity là high.*

   Để thực hiện, đầu tiên chọn bộ lọc có tên đã chỉ định, sau đó nhấn chuột vào vùng văn bản để mở cửa sổ Generic Object Editor và tìm cách thiết lập các thông số một cách phù hợp.

   Chụp màn hình cửa sổ Viewer để thể hiện dữ liệu trước và sau khi áp dụng bộ lọc.

   Chụp màn hình cửa sổ Generic Object Editor để thể hiện tham số đã thiết lập.

Lưu ý rằng, nếu muốn sử dụng tiếp dữ liệu, hãy hủy hành động thay đổi trên dữ liệu vừa làm và kiểm tra rằng dữ liệu đã trở về trạng thái nguyên thủy, điều này tránh mọi sai sót về sau.

Yêu cầu 3 – Khảo sát tập dữ liệu Iris (thuộc tính số liên tục)

Tải tập dữ liệu Iris tại link bên dưới hoặc tìm trong thư mục cài đặt Weka

http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/iris.arff

Đọc dữ liệu vào Explorer. Nhấn nút Edit từ dãy nút trên đầu tab Preprocess → Cửa số Viewer được mở và nội dung trong cửa sổ hiển thị mọi mẫu trong tập dữ liệu.

9. *Mô tả tập dữ liệu. Nội dung của phần ghi chú (comment) nói về điều gì? Tập dữ liệu có bao nhiêu mẫu? Bao nhiêu thuộc tính? Miền giá trị của thuộc tính petallength là gì? Tập dữ liệu có bao nhiêu thuộc tính số và bao nhiêu thuộc tính rời rạc? Tên của thuộc tính lớp là gì? Đánh giá phân bố của các lớp, tức là cân bằng hay lệch về một lớp?*

Tab Visualize

Bây giờ hãy quan sát chức năng trình diễn dữ liệu của Weka ở tab Visualize. Chức năng này làm việc tốt với dữ liệu số. Nhấp chuột vào biểu đồ điểm đầu tiên trên dòng thứ hai của ma trận biểu đồ để mở cửa sổ Visualizing iris hiển thị biểu đồ phóng to, với các trục như đã chọn. Các mẫu được thể hiện bằng dấu x nhỏ, có màu sắc tùy thuộc vào lớp của mẫu. Trục x biểu diễn thuộc tính **sepallength** và trục y biểu diễn **petalwidth**. Nhấp chuột đôi vào một trong các dấu x sẽ kích hoạt cửa sổ **Instance Info**, trong đó liệt kê giá trị của mọi thuộc tính cho mẫu đã chọn. Đóng cửa sổ Instance Info để trở về cửa sổ Visualizing iris. Các trường tùy chọn ở phần trên của cửa sổ cho biết thuộc tính nào được chọn cho trục x và trục y. Đổi trục x thành petalwidth và trục y thành petallengh. Trường thể hiện Color: class (**Num)** dùng để đổi ký hiệu màu.

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

Mỗi thanh đủ màu ở bên phải scatter plot biểu diễn một thuộc tính, các mẫu được đặt ở vị trí thích hợp theo trục x và ngẫu nhiên theo trục y. Nhấp chuột vào thanh sẽ chọn sử dụng thuộc tính tương ứng cho trục x của scatter plot, và nhấp chuột phải sẽ làm điều tương tự cho trục y. Sử dụng các thanh này để thay đổi thuộc tính ở trục x và trục y thành sepallength và petalwidth. **Jitter slider** dịch chuyển dấu x cho mỗi mẫu một cách ngẫu nhiên tính từ vị trí thật của nó và nhờ đó giúp nhìn rõ trong tình huống nhiều mẫu nằm trùng lên nhau.

Hãy trải nghiệm một chút bằng cách di chuyển Jitter slider. Nút **Select Instance** và các nút **Reset, Clear** và **Save** cho phép hiệu chỉnh dữ liệu. Các mẫu tương ứng sẽ được chọn và các mẫu khác sẽ bị loại bỏ. Hãy thử lựa chọn Rectangle: Chọn một vùng bằng cách nhấp chuột trái và rê chuột. Nút Reset sẽ chuyển thành nút Submit. Nhấp vào đó khiến cho mọi mẫu nằm ngoài hình chữ nhật sẽ bị loại. Bạn có thể dùng nút **Save** để lưu dữ liệu đã hiệu chỉnh vào một tập tin, trong khi Reset sẽ phục hồi tập dữ liệu về tình trạng ban đầu.

Tùy ý minh họa và chụp màn hình và đánh dấu những nội dung tương ứng để làm minh chứng.

Yêu cầu 4 – Khảo sát tập dữ liệu Weather (thuộc tính số liên tục)

Tải tập dữ liệu Weather (numeric) tại link bên dưới hoặc tìm trong thư mục cài đặt Weka

http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/weather.arff

10. *Thực hiện lại những câu hỏi trong Yêu cầu 3. Mô tả lại những quan sát mà bạn có được trong quá trình khảo sát dữ liệu.*

B – Nội dung thực hiện cài đặt (25 điểm)

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.py

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.py