Trường Đại học Khoa học Tự nhiên
Khoa Công nghệ thông tin
Bộ môn Khai thác dữ liệu và ứng dụng

BÁO CÁO

BÀI TẬP THỰC HÀNH 1:

TIỀN XỬ LÍ DỮ LIỆU VỚI WEKA

GVHD: Lê Ngọc Thành, Nguyễn Ngọc Thảo

TP. Hồ Chí Minh, ngày 04 tháng 04 năm 2019

MỤC LỤC

I. Thông tin thành viên:

1. Chu Phúc Nguyên 1512353
2. Võ Nhật Vinh 1612815

II. Báo cáo tổng quát:

Nhóm đã hoàn thành tất cả câu hỏi trong file bài tập giáo viên giao.

III. Báo cáo chi tiết:

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

a. Định nghĩa sự tích hợp dữ liệu.

Tích hợp dữ liệu là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu có sẵn cho quá trình khai phá 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?

- Xuất hiện vấn đề nhận diện thực thể trong 2 dataset. Với set heart-c.arff ở thuộc tính thứ 3 là cp trong khi ở set heart-h.arff là chest_pain
- Giải quyết: Đổi tên 2 thuộc tính này giống nhau. Nhóm đổi tên là chest_pain.

c. Có vấn đề dữ liệu dư thừa (redundancy) trong 2 dataset này hay không? Nếu có, giải quyết như thế nào?

Không có dư thừa dữ liệu trong 2 dataset.

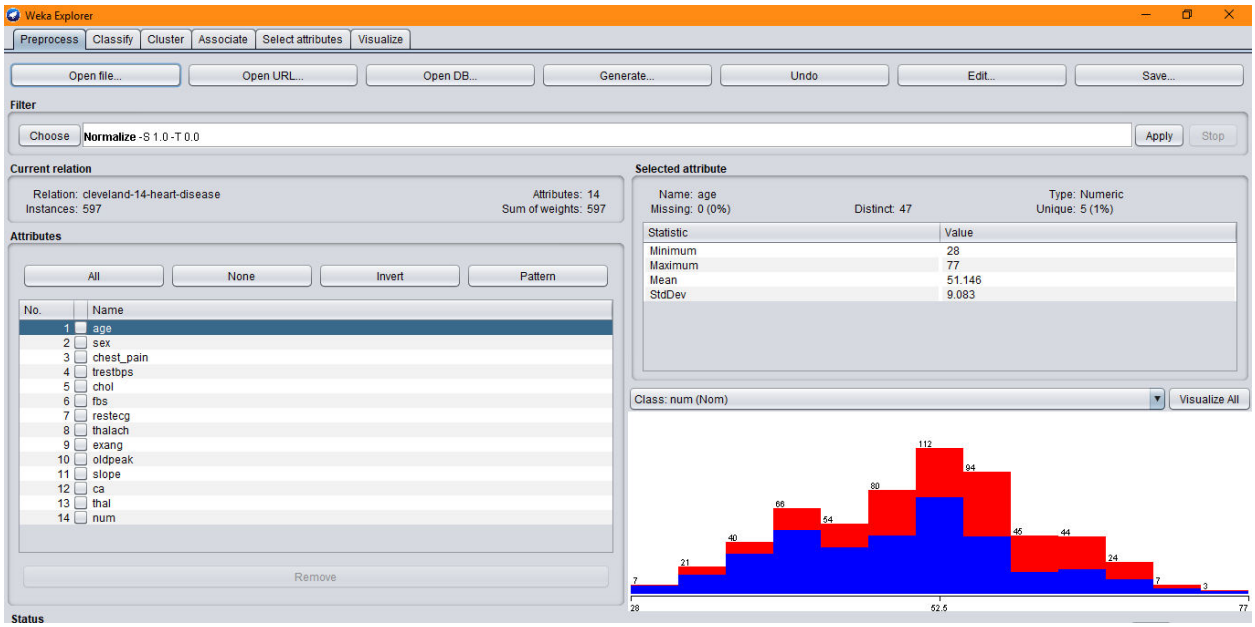
d. Có sự mâu thuẫn dữ liệu (data value conflicts) trong 2 dataset này hay không? Nếu có, giải quyết như thế nào?

- Có sự mâu thuẫn dữ liệu giữa 2 data cụ thể là ở thuộc tính thứ 11 'slope', ở heart-c.arff là {up, flat, down}, còn ở heart-h.arff là {down, flat, up}.
- Giải quyết: Ta thay đổi thứ tự các giá trị thuộc tính này đúng thứ tự với nhau.

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?

- 14 thuộc tính.
- 597 mẫu.

f. Chụp lại màn hình của cửa sổ Explorer của bạn.



Hình 1: Ảnh chụp màn hình cửa sổ Explorer sau khi nạp 2 dataset đã được tích hợp.

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

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?

- Trung bình: 51.146
- Độ lệch chuẩn: 9.083
- Giá trị nhỏ nhất: 28
- Giá trị lớn nhất: 77

b. Liệt kê five-number summary của thuộc tính này. Weka có cung cấp những con số này hay không?

- Five-number summary của thuộc tính age:
  - Giá trị nhỏ nhất: 28
  - Giá trị lớn nhất: 77
  - Trung vị: 52
  - First quartile: 44
  - Third quartile: 58
- Weka chỉ cung cấp giá trị lớn nhất và giá trị nhỏ nhất.

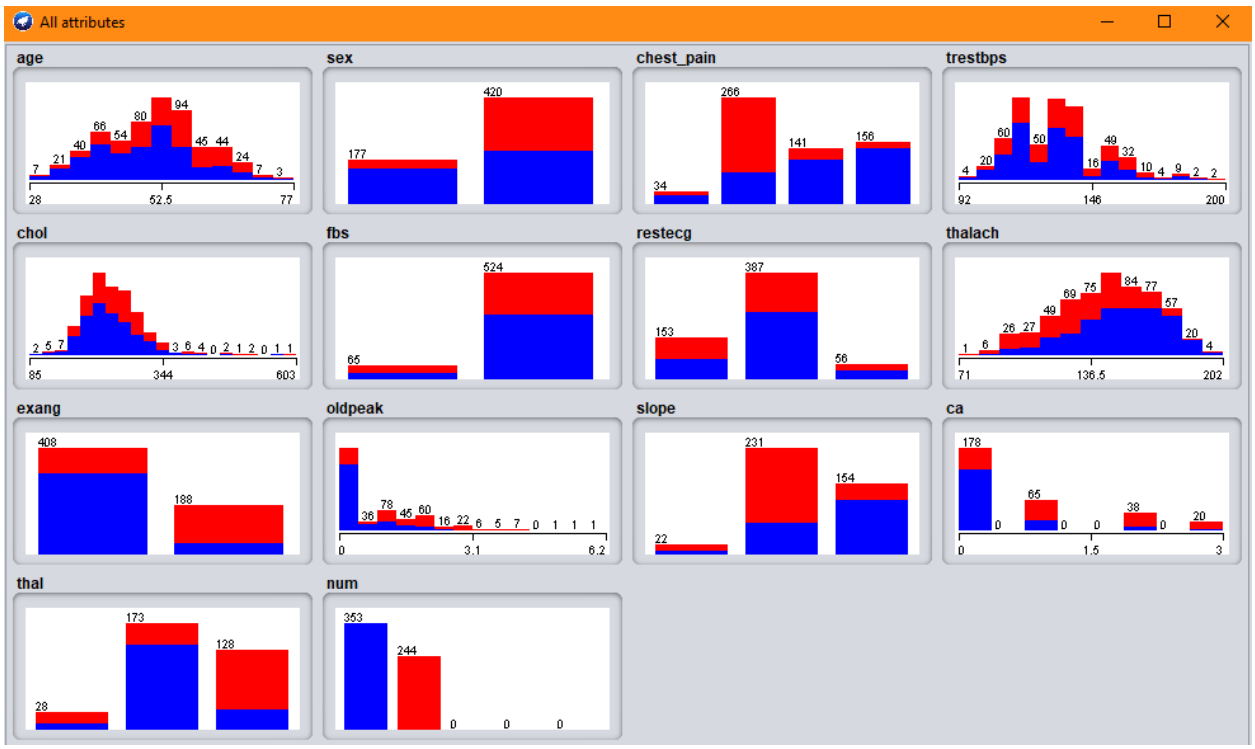
c. Cho biết thuộc tính nào là số (numeric), thuộc tính nào là có thứ tự (ordernal) và thuộc tính nào là rời rạc/danh sách (categorical/nomial).

- Các thuộc tính là số (numeric): age, trestbps, chol, thalach, oldpeak, ca.
- Các thuộc tính là có thứ tự (ordernal): chest_pain, restecg, slope, thal.
- Các thuộc tính rời rạc/danh sách (categorical/nomial): sex, fbs, exang, num.

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?

- Ý nghĩa của đồ thị: Biểu diễn các giá trị của thuộc tính, tần suất phân bố của dữ liệu.
- Tên đồ thị: Đồ thị phân phối dữ liệu (histogram)
- Ý nghĩa màu xanh: không bị bệnh (<50), màu đỏ: bị bệnh (>50_1)
- Đồ thị biểu diễn tần suất phân bố của dữ liệu đang xét và thông số thuộc tính.

e. Lần lượt xem xét các thuộc tính khác của dataset dưới dạng đồ thị. Dán các ảnh chụp màn hình vào bài làm.



Hình 2: Ảnh chụp màn hình các đồ thị của 14 thuộc tính.

f. Nhận xét của bạn từ những đồ thị đó?

- Các thuộc tính có kiểu dữ liệu là số (numeric) sẽ có dạng đồ thị các cột nối liền với nhau riêng thuộc tính **ca** có các missing value nên đồ thị có vẻ bị tách rời. Kiểu dữ liệu số không có label.
- Các kiểu dữ liệu rời rạc có thứ tự các cột rời nhau, và có label.

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).

- Thuật ngữ sử dụng trong textbook đặt tên cho đồ thị là: Scatter plot
- Các thuộc tính dẫn đến bệnh tim nhiều nhất: chest_pain, exang, slope, oldpeak.
- Đồ thị có khả năng dự đoán bệnh tim tốt nhất là exang.

h. Có những cặp thuộc tính khác nhau nào có vẻ như tương quan với nhau không?

Không phát hiện những cặp thuộc tính tương quan với nhau sau khi đã tính độ tương quan giữa các thuộc tính dựa theo công thức Pearson'sproductmoment coefficient. Các kết quả đều <0.5.

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

a. Bạn hãy cho biết có bao nhiêu thuộc tính trong những dataset trước khi xử lý?

Có 14 thuộc tính trong mỗi dataset trước khi xử lí dữ liệu.

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.

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,…
- Search Method. Để xác định một phương pháp (thứ tự) xét các thuộc tính.
  Vd: best-first, greedy step wise, ranking,…

Có 3 phương pháp tìm kiếm chọn lọc thuộc tính trong weka:

- BestFirst: Tìm kiếm không gian của các tập hợp thuộc tính bằng cách tăng tốc tham lam với một cơ sở quay lui. Đặt ra số lượng nút không cải thiện liên tiếp cho phép kiểm soát mức độ quay lui được thực hiện. BestFirst có thể bắt đầu với bộ thuộc tính trống và tìm kiếm chuyển tiếp hoặc bắt đầu với bộ thuộc tính đầy đủ và tìm kiếm ngược hoặc bắt đầu tại bất kỳ điểm nào và tìm kiếm theo cả hai hướng.

- GreedyStepWise: Thực hiện tìm kiếm tiến hoặc lùi tham lam thông qua không gian của các tập hợp thuộc tính. Có thể bắt đầu bằng không hoặc tất cả các thuộc tính hoặc từ một điểm tùy ý trong không gian. Dừng khi thêm, xóa bất kỳ thuộc tính còn lại nào dẫn đến giảm sự đánh giá. Cũng có thể tạo một danh sách các thuộc tính được xếp hạng bằng cách di chuyển

không gian giữa hai bên và ghi lại thứ tự các thuộc tính được chọn.

- Ranker: Xếp hạng các thuộc tính theo đánh giá cá nhân của nó. Sử dụng kết hợp với các đánh giá thuộc tính (SavingF, GainRatio, Entropy, v.v.).

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?

Trong textbook không có phương pháp ranker và best first.

4. Chuẩn bị dữ liệu – 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.

- Các phương pháp xử lý dữ liệu bị thiếu:
  - Ignore the tuple. (Bỏ qua các bộ các thuộc tính thiếu giá trị. Thường áp dụng trong các bài toán phân lớp. Hoặc khi tỷ lệ % các giá trị thiếu đối với các thuộc tính quá lớn.)
  - Fill the missing value manually. (Gán giá trị mặc định)
  - Use a global constant to fill the missing value. (Gán 1 giá trị chung cho thuộc tính)
  - Use the attribute mean to fill the missing value. (Gán giá trị trung bình của thuộc tính đó.)

- Use the attribute mean for all sample belonging to the same class as the given tuple. (Sử dụng trung bình thuộc tính cho tất cả các mẫu thuộc cùng một lớp với bộ dữ liệu đã cho.)
- Use the most probable value to fill the missing value(Gán giá trị có thể xảy ra nhất)

- Weka đã cài đặt những phương pháp: Add noise, Remove, Remove Percentage, Remove type, Remove Frequent values, Remove Misclassified, Remove Range, Remove Useless, Remove Missing Value, Subset by Expression, Interquartile Range, Remove with Values.
- Sử dụng phương pháp:ReplaceMissingValues trong weka (Thay thế giá trị thiếu dựa trên mean và mode)
- Giải thích: Ta thấy rằng trong tập dữ liệu có đến 50% là dữ liệu thiếu, nên không thể loại bỏ hết dữ liệu thiếu này vì sẽ làm mất tích khách quan của dữ liệu. Ta có thể dựa vào mean và mode của dữ liệu tập huấn luyện để suy ra được những dữ liệu còn thiếu, cách này theo đánh giá sẽ tạo tính khác quan hơn, nhất là khi dữ liệu bị thiếu nhiều.

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?

- Những phương pháp loại bỏ dữ liệu nhiễu:
  - Phân khoảng (binning): Sắp xếp dữ liệu và phân chia thành các khoảng (bins) có tần số xuất hiện giá trị như nhau. Sau đó, mỗi khoảng dữ liệu có thể được biểu diễn bằng trung bình, trung vị, hoặc các giới hạn … của các giá trị trong khoảng đó.

- Hồi quy (regression): Gắn dữ liệu với một hàm hồi quy.
- Phân cụm (clustering).

- Weka đã cài đặt những phương pháp:
  - Binning (Discretize filter trong weka).
  - Hồi quy (regression).
  - Phân cụm (clustering).

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.

- Phương pháp dò tìm dữ liệu tạp:
  - InterquartileRange.
  - Z-Score
- Dò tìm dữ liệu bằng weka:
  - Open dataset.
  - Trong mục Filter, **Choose -> filter -> unsupervised -> attributes -> ReplaceMissingValues -> Apply**.
  - Trong mục Filter, click **Choose**, click **Filter** button ở cuối cửa sổ.
  - Lúc này xuất hiện bảng Filter Capabilities. Check vào các ô **numeric atributes** và **numeric class**, Click **OK**.
  - Chọn **Choose -> filter -> unsupervised -> attributes -> InterQuartileRange**.
  - Click **Apply**.
  - Bây giờ ta có thể xem kết quả trong cửa sổ Viewer. Ở cột outlier, nếu chữ Yes thì dòng dữ liệu đó có dữ liệu tạp.

- Trong dataset có dữ liệu tạp:

- 62,female,asympt,160,164,f,left_vent_hyper,145,no,6.2,down,3,reversable_defect,>50_1,yes,yes
- 67,female,non_anginal,115,564,f,left_vent_hyper,160,no,1.6,flat,0,reversable_defect,<50,yes,no
- 48,female,atyp_angina,0,308,f,st_t_wave_abnormality,0,0,2,up,0,0,<50,yes,no

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.

Relation: cleveland-14-heart-disease-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-weka.filters.unsupervised.attribute.InterquartileRange-Rfirst-last-O3.0-

No.	1: age	2: sex	3: chest_pain	4: trestbps	5: chol	6: fbs	7: restecg	8: thalach	9: exang	10: oldpeak	11: slope	12: ca	13: thal	14: num	15: Outlier	16: Extreme'
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	'(52.5-57.4]'	female	asympt	'(189.2-inf)'	'(240.4-292.2]'	t	left_vent_hyper	'(123.4-136.5]'	yes	'(3.72-4.34]'	down	'(1.8-2.1]'	reversable_defect	(50	no	no
2	'(57.4-62.3]'	female	asympt	'(135.2-146]'	'(240.4-292.2]'	f	left_vent_hyper	'(149.6-162.7]'	no	'(3.1-3.72]'	down	'(1.8-2.1]'	normal	)50_1	no	no
3	'(62.3-67.2]'	female	typ_angina	'(146-156.8]'	'(188.6-240.4]'	f	normal	'(110.3-123.4]'	no	'(2.48-3.1]'	down	'(-inf-0.3]'	normal	)50_1	no	no
4	'(42.7-47.6]'	female	non_anginal	'(135.2-146]'	'(136.8-188.6]'	f	left_vent_hyper	'(149.6-162.7]'	yes	'(1.24-1.86]'	down	'(-inf-0.3]'	normal	(50	no	no
5	'(57.4-62.3]'	female	asympt	'(156.8-167.6]'	'(136.8-188.6]'	f	left_vent_hyper	'(136.5-149.6]'	no	'(5.58-inf)'	down	'(2.7-inf)'	reversable_defect	(50	no	no
6	'(42.7-47.6]'	female	asympt	'(124.4-135.2]'	'(292.2-344]'	t	left_vent_hyper	'(123.4-136.5]'	yes	'(2.48-3.1]'	flat	'(-inf-0.3]'	reversable_defect	(50	no	no
7	'(57.4-62.3]'	female	asympt	'(135.2-146]'	'(292.2-344]'	t	normal	'(97.2-110.3]'	no	'(1.86-2.48]'	flat	'(2.7-inf)'	normal	)50_1	no	no
8	'(62.3-67.2]'	female	asympt	'(167.6-178.4]'	'(188.6-240.4]'	t	normal	'(162.7-175.8]'	yes	'(0.62-1.24]'	flat	'(1.8-2.1]'	reversable_defect	(50	no	no
9	'(57.4-62.3]'	female	asympt	'(167.6-178.4]'	'(188.6-240.4]'	t	left_vent_hyper	'(136.5-149.6]'	yes	'(2.48-3.1]'	flat	'(1.8-2.1]'	fixed_defect	)50_1	no	no
10	'(47.6-52.5]'	female	atyp_angina	'(113.6-124.4]'	'(240.4-292.2]'	t	st_t_wave_abnormality	'(136.5-149.6]'	no	'(-inf-0.62]'	flat	'(0.6-0.9]'	normal	)50_1	no	no
11	'(52.5-57.4]'	female	atyp_angina	'(113.6-124.4]'	'(188.6-240.4]'	t	normal	'(136.5-149.6]'	no	'(-inf-0.62]'	flat	'(0.6-0.9]'	normal	(50	no	no
12	'(57.4-62.3]'	female	asympt	'(124.4-135.2]'	'(292.2-344]'	t	st_t_wave_abnormality	'(123.4-136.5]'	yes	'(1.24-1.86]'	flat	'(0.6-0.9]'	normal	(50	no	no
13	'(42.7-47.6]'	female	non_anginal	'(124.4-135.2]'	'(240.4-292.2]'	t	normal	'(162.7-175.8]'	no	'(-inf-0.62]'	flat	'(0.6-0.9]'	normal	)50_1	no	no
14	'(52.5-57.4]'	female	atyp_angina	'(135.2-146]'	'(292.2-344]'	f	left_vent_hyper	'(149.6-162.7]'	no	'(1.24-1.86]'	flat	'(-inf-0.3]'	normal	(50	no	no
15	'(47.6-52.5]'	female	non_anginal	'(113.6-124.4]'	'(188.6-240.4]'	f	normal	'(149.6-162.7]'	no	'(1.24-1.86]'	flat	'(-inf-0.3]'	normal	(50	no	no
16	'(62.3-67.2]'	female	asympt	'(146-156.8]'	'(188.6-240.4]'	f	left_vent_hyper	'(110.3-123.4]'	no	'(0.62-1.24]'	flat	'(2.7-inf)'	reversable_defect	(50	no	no
17	'(47.6-52.5]'	female	asympt	'(124.4-135.2]'	'(292.2-344]'	f	normal	'(136.5-149.6]'	yes	'(0.62-1.24]'	flat	'(-inf-0.3]'	reversable_defect	)50_1	no	no
18	'(52.5-57.4]'	female	asympt	'(124.4-135.2]'	'(240.4-292.2]'	f	left_vent_hyper	'(136.5-149.6]'	no	'(-inf-0.62]'	flat	'(-inf-0.3]'	normal	(50	no	no
19	'(42.7-47.6]'	female	non_anginal	'(102.8-113.6]'	'(136.8-188.6]'	f	normal	'(162.7-175.8]'	no	'(-inf-0.62]'	flat	'(-inf-0.3]'	normal	(50	no	no
20	'(57.4-62.3]'	female	asympt	'(146-156.8]'	'(240.4-292.2]'	f	left_vent_hyper	'(149.6-162.7]'	no	'(2.48-3.1]'	flat	'(1.8-2.1]'	reversable_defect	(50	no	no
21	'(57.4-62.3]'	female	asympt	'(135.2-146]'	'(292.2-344]'	f	left_vent_hyper	'(136.5-149.6]'	yes	'(0.62-1.24]'	flat	'(-inf-0.3]'	reversable_defect	(50	no	no

Hình 3: Ảnh chụp màn hình sau khi đã làm sạch dữ liệu.

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

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?

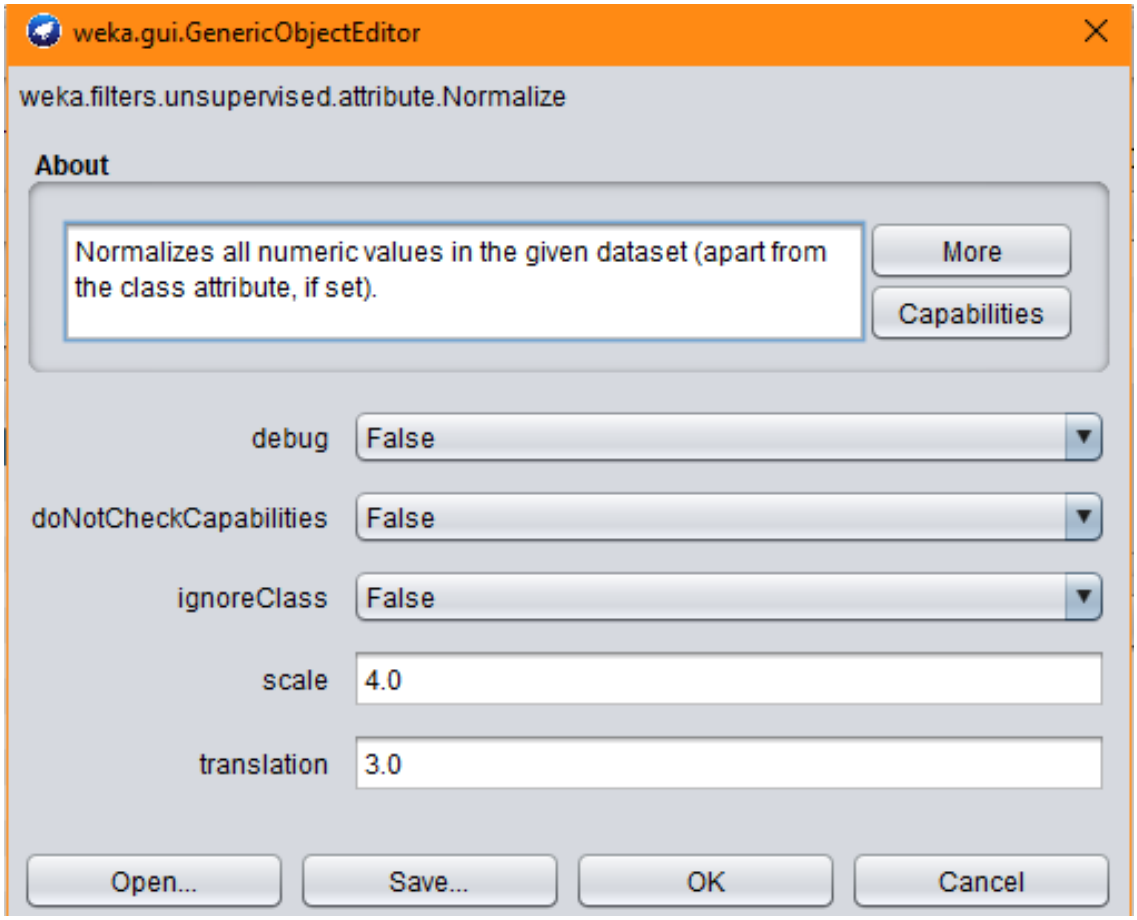
Bộ lọc của weka cho phép làm điều này là: Copy.
(weka.filters.unsupervised.attribute.Copy)

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.

- Bộ lọc cho phép chuẩn hóa thuộc tính trong weka: **Normalize, Standardize,**
- Các bộ lọc này có thể chuẩn hóa Min-max (Normalize), chuẩn hóa Z-score (Standardize), không chuẩn hóa thập phân.
- Cách thực hiện chuẩn hóa:
  - Chuẩn hóa Min-max:

+ Trong mục filter, chọn ***unsupervised->attribute->Normalize.***

+ Click chuột vào filter box, Cửa sổ sau hiển thị:



Hình 3: Cửa sổ Filter box của Normalize

+ Ở ô translation nhập giá trị min, scale nhập (max-min).

+ Click OK -> Apply

- Chuẩn hóa Z-Score: Trong mục filter, chọn **unsupervised->attribute->Standardize->Apply.**

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.

- Chọn phương pháp chuẩn hóa Min-max.
- Giải thích: Ta thấy các giá trị thuộc tính số thực ở đây đều dương, nên ta có thể chuẩn hóa Min-max về số dương trong đoạn[0,1].

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.

Viewer

Relation: cleveland-14-heart-disease-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0

No.	1: age	2: sex	3: chest_pain	4: trestbps	5: chol	6: fbs	7: restecg	8: thalach	9: exang	10: oldpeak	11: slope	12: ca	13: thal	14: num
	Numeric	Nominal	Nominal	Numeric	Numeric	Nominal	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Nominal
1	0.7142857...	male	typ_angina	0.4907...	0.28...	t	left_ve...	0.6030...	no	0.37096...	down	0.0	fixed_defect	(50
2	0.7959183...	male	asympt	0.6296...	0.38...	f	left_ve...	0.2824...	yes	0.24193...	flat	1.0	normal	)50_1
3	0.7959183...	male	asympt	0.2592...	0.27...	f	left_ve...	0.4427...	yes	0.41935...	flat	0.666666...	reversable_defect	)50_1
4	0.2653061...	fem...	atyp_angina	0.3518...	0.22...	f	left_ve...	0.7709...	no	0.22580...	up	0.0	normal	(50
5	0.6938775...	fem...	asympt	0.4444...	0.35...	f	left_ve...	0.6793...	no	0.58064...	down	0.666666...	normal	(50
6	0.7142857...	male	asympt	0.3518...	0.32...	f	left_ve...	0.5801...	no	0.22580...	flat	0.333333...	reversable_defect	(50
7	0.5102040...	male	asympt	0.4444...	0.22...	t	left_ve...	0.6412...	yes	0.5	down	0.0	reversable_defect	)50_1
8	0.5714285...	fem...	atyp_angina	0.4444...	0.40...	f	left_ve...	0.6259...	no	0.20967...	flat	0.0	normal	(50
9	0.5714285...	male	non_anginal	0.3518...	0.33...	t	left_ve...	0.5419...	yes	0.09677...	flat	0.333333...	fixed_defect	)50_1
10	0.7346938...	male	typ_angina	0.1666...	0.24...	f	left_ve...	0.5572...	yes	0.29032...	flat	0.0	normal	)50_1
11	0.6122448...	fem...	typ_angina	0.5370...	0.38...	t	left_ve...	0.6946...	no	0.16129...	up	0.0	normal	(50
12	0.6122448...	male	atyp_angina	0.2592...	0.38...	f	left_ve...	0.6793...	no	0.29032...	flat	0.0	normal	(50
13	0.6122448...	male	non_anginal	0.3703...	0.26...	f	left_ve...	0.7786...	no	0.51612...	up	0.666666...	reversable_defect	)50_1
14	0.6530612...	male	asympt	0.3518...	0.23...	f	left_ve...	0.4656...	yes	0.38709...	flat	0.666666...	reversable_defect	(50
15	0.2448979...	male	asympt	0.1666...	0.15...	f	left_ve...	0.3282...	yes	0.32258...	flat	0.0	reversable_defect	(50
16	0.3061224...	male	asympt	0.2592...	0.17...	f	left_ve...	0.3740...	yes	0.40322...	flat	0.0	reversable_defect	(50
17	0.5918367...	male	asympt	0.5370...	0.36...	f	left_ve...	0.3129...	yes	0.09677...	flat	0.333333...	fixed_defect	)50_1
18	0.7551020...	fem...	asympt	0.5370...	0.27...	f	left_ve...	0.3282...	no	0.16129...	flat	1.0	reversable_defect	(50
19	0.6734693...	fem...	asympt	0.3518...	0.47...	f	left_ve...	0.7480...	no	0.0	up	0.0	normal	(50
20	0.6122448...	male	non_anginal	0.1851...	0.27...	f	left_ve...	0.7175...	no	0.40322...	flat	0.333333...	reversable_defect	(50
21	0.4489795...	male	asympt	0.5370...	0.30...	f	left_ve...	0.4351...	no	0.41935...	flat	0.0	reversable_defect	(50
22	0.7551020...	fem...	non_anginal	0.4444...	0.64...	t	left_ve...	0.6564...	no	0.12903...	up	0.333333...	normal	(50
23	0.5102040...	male	non_anginal	0.3518...	0.21...	t	left_ve...	0.6183...	no	0.19354...	down	0.0	normal	)50_1
24	0.3265306...	male	asympt	0.1851...	0.39...	f	left_ve...	0.6259...	no	0.0	up	0.333333...	normal	)50_1
25	0.3265306...	male	atyp_angina	0.3518...	0.25...	f	left_ve...	0.8931...	no	0.0	up	0.0	normal	)50_1

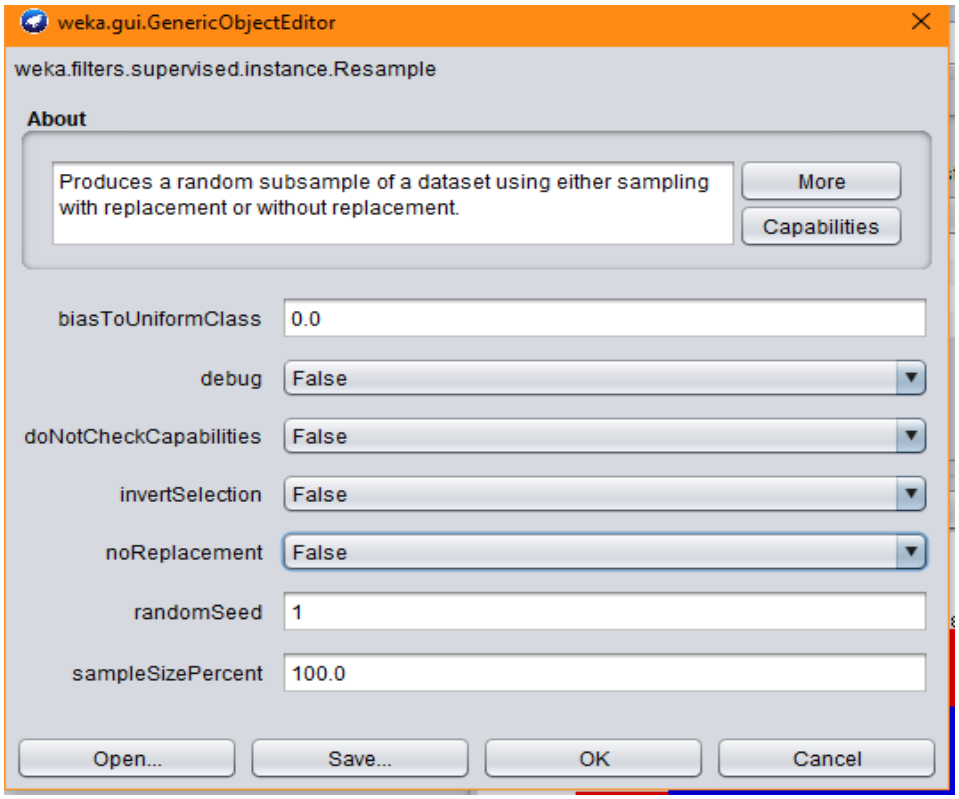
Hình 4: Hình ảnh của dữ liệu sau khi chuẩn hóa

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

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 ?

- Để lấy mẫu với các bộ lọc trong weka ta làm như sau:
  - Trong bộ lọc của weka, chọn **unsupervised->instance->Resample.**
  - Click chuột vào box filter, ở ô noReplacement: Chọn True nếu dùng phương pháp **Simple Random Sample Without Replacement,** chọn False nếu dùng phương pháp **Simple Random Sample With Replacement.**



Hình 5: Cửa sổ filter box của filter Resample.

- Weka có thể thực hiện được 2 phương pháp **Simple Random Sample Without Replacement** và **Simple Random Sample With Replacement.**

Tài liệu Tham khảo:

1. Paper - A study on missing handling values and noisy data using weka tool: https://issuu.com/ijsrd/docs/ijsrdv4i50673
2. Add new attribute calculated based on other attributes. Trích xuất: https://stackoverflow.com/questions/13288687/add-new-attribute-calculated-based-on-other-attributes
3. Creating samples without replacement from a dataset. Truy xuất: http://weka.8497.n7.nabble.com/Creating-Samples-without-replacement-from-a-dataset-td25398.html
4. Book - Data mining concepts and technique. 3rd Edition. Jiawei Han, Micheline Kamber, Jian Pei.