TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TPHCM

Khoa Công nghệ thông tin

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

BÀI TẬP 4

Gom cụm

Clustering

Giảng viên hướng dẫn: Nguyễn Ngọc Thảo

Người thực hiện: Nhóm 37

Hà Tiến Đạt – 18424023

Vũ Mạnh Hùng -18424029

TP.HCM – 09/2019

Nội dung

I) Nội dung thực hiện báo cáo viết (15 điểm) 2
1) Yêu cầu 1 – Tiền xử lý dữ liệu (5.0 điểm) 2
2) Yêu cầu 2 – Gom cụm có xét thông tin phân lớp (5.0 điểm) 4
3) Yêu cầu 3 – Đánh giá kết quả gom cụm (5.0 điểm) 6
II) Nội dung thực hiện cài đặt (15 điểm) 7
1) (2.0đ) Chương trình nhận dữ liệu đầu vào là tập tin cardiology-cleaned.arff 7
2) (2.0đ) Chương trình phát sinh dữ liệu đầu ra là tập tin <k>-clusters.txt, với k là số cụm cần gom. Tập tin có nội dung tương tự như những gì hiển thị trong Clusterer Output của 7
3) (1.0đ) Chương trình thực thi giải thuật k-means với cú pháp tham số dòng lệnh là 8
4) (10.0đ) Đối chiếu kết quả phát sinh được với kết quả của WEKA (đã thực hiện ở phần Nội dung thực hiện báo cáo viết) trên cùng giá trị k (từ 2 đến 5). 8

I) Nội dung thực hiện báo cáo viết (15 điểm)

Dữ liệu thực nghiệm

Tập dữ liệu bệnh học tim cardiology (tập tin cardiology.arff được gửi kèm với đề bài), bao gồm các mẫu chứa thông số bệnh học của bệnh nhân có vấn đề về tim (Sick) và không có vấn đề về tim (Healthy).

Thông tin cơ bản về tập dữ liệu bao gồm

- Số mẫu: 303
- Số lượng thuộc tính; 14, cả dạng rời rạc và dạng số

Các bác sĩ khoa tim cho rằng "nam giới có nguy cơ mắc bệnh về tim cao, trong khi nữ giới có nguy cơ mắc bệnh về tim thấp". Nhiệm vụ của sinh viên là áp dụng giải thuật kmean của WEKA (weka.clusterers.SimpleKMeans) để kiểm chứng phát biểu này.

1) Yêu cầu 1 – Tiền xử lý dữ liệu (5.0 điểm)

a) Liệt kê các thuộc tính có xảy ra trường hợp dữ liệu thiếu. Cho biết số giá trị thiếu ở mỗi thuộc tính đã nhận diện.

Các thuộc tính xảy ra trường hợp thiếu dữ liệu: age, stope
Trong đó: age có 3 giá bị thiếu, stope có 5 giá trị bị thiếu

b) Để tránh việc thiếu dữ liệu ảnh hưởng đến chất lượng phân tích, hãy áp dụng bộ lọc thích hợp trong nhóm Unsupervised/Attribute để thay thế mọi giá trị thiếu trong tập dữ liệu gốc. Đồng thời, đổi tên thuộc tính class thành heart-condition. Lưu dữ liệu sau tiền xử lý vào tập tin cardiology-cleaned.arff và nộp lại tập tin.Tập tin sai định dạng hoặc thiếu dòng dữ liệu (-1đ). Không nộp tập tin (-2đ).

Đã hoàn thành.

c) Các giá trị thiếu đã lần lượt được thay thế bằng những giá trị gì? Giải thích lý do vì sao chúng được thay thế như vậy?

No.	1: age Numeric
1	
2	
3	64.0
4	63.0
5	53.0
6	58.0
7	58.0
8	58.0
9	63.0
10	
11	67.0
12	37.0

No.	1: age Numeric
1	54.3...
2	54.3...
3	64.0
4	63.0
5	53.0
6	58.0
7	58.0
8	58.0
9	63.0
10	54.3...

Selected attribute

Name: age Type: Nume
Missing: 3 (1%) Distinct: 41 Unique: 4 (1%)

Statistic	Value
Minimum	29
Maximum	77
Mean	54.323
StdDev	9.087

Các giá trị bị thiếu trong attribute age được thay thế thành: 54.3. Đây là giá trị trung bình của thuộc tính age trước khi sử dụng filter.

7	58.0	1.8	Flat
8	58.0	3.2	
9	63.0	2.3	Down
10		1.5	Flat
11	67.0	2.6	Flat
12	37.0	3.5	Down
13	41.0	1.4	Up
14	56.0	0.8	Up
15	62.0	3.6	Down
16	57.0	0.6	
17	57.0	0.4	Flat
18	56.0	1.3	Flat
19	56.0	0.6	Flat
20	44.0	0.0	Up
21	50.0	1.6	Flat
22	58.0	0.0	
23	66.0	2.6	Down
24	43.0	1.5	
25	40.0	2.0	Flat
26	69.0	1.8	
27	60.0	1.4	Up

7	58.0	1.8	Flat
8	58.0	3.2	Flat
9	63.0	2.3	Down
10	54.3...	1.5	Flat
11	67.0	2.6	Flat
12	37.0	3.5	Down
13	41.0	1.4	Up
14	56.0	0.8	Up
15	62.0	3.6	Down
16	57.0	0.6	Flat
17	57.0	0.4	Flat
18	56.0	1.3	Flat
19	56.0	0.6	Flat
20	44.0	0.0	Up
21	50.0	1.6	Flat
22	58.0	0.0	Flat
23	66.0	2.6	Down
24	43.0	1.5	Flat
25	40.0	2.0	Flat
26	69.0	1.8	Flat
27	60.0	1.4	Up

Các giá trị bị thiếu của attribute stope được thay thế thành flat. Đây là giá trị có số lượng nhiều nhất trong thuộc tính

Selected attribute

Name: slope Type: Nominal
Missing: 5 (2%) Distinct: 3 Unique: 0 (0%)

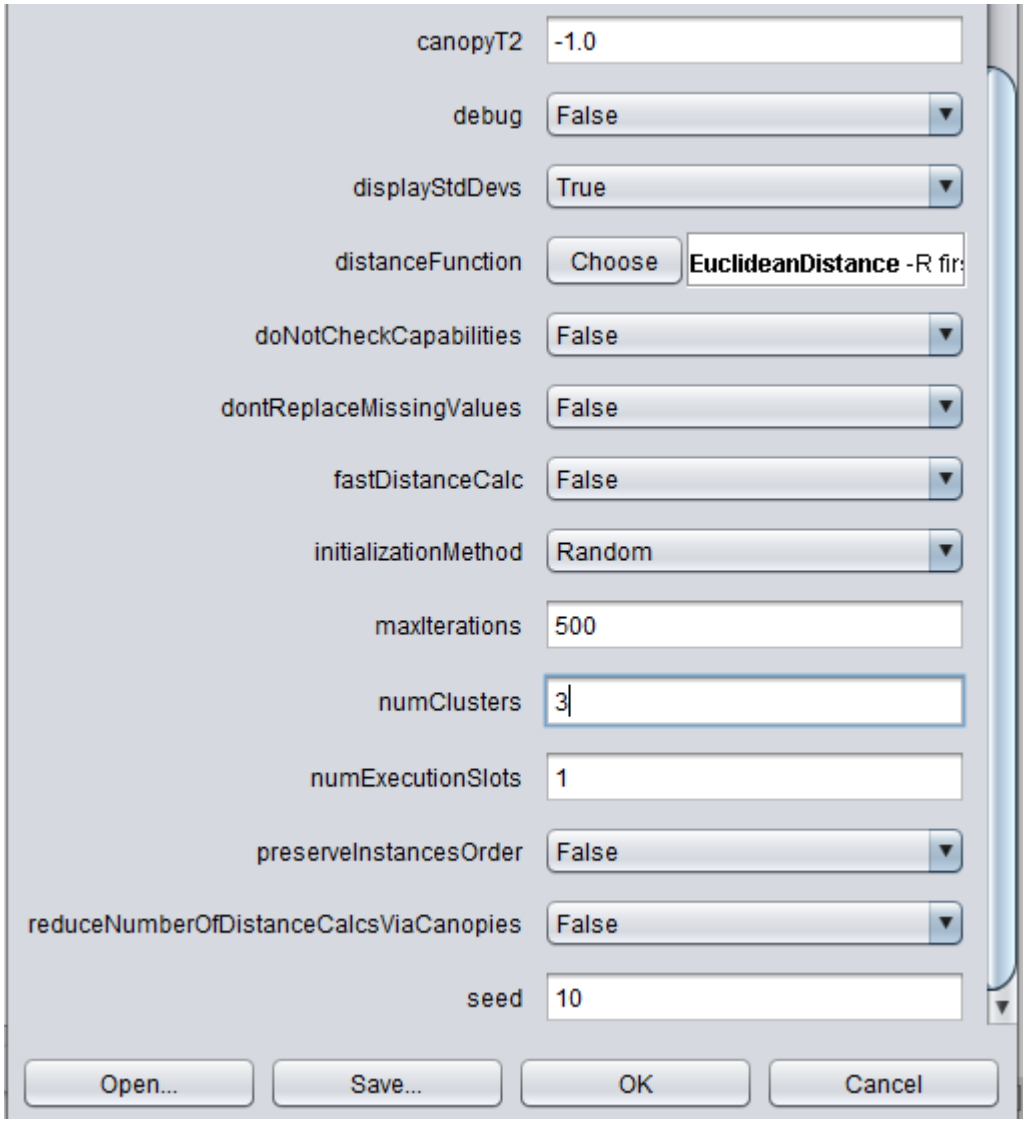
No.	Label	Count	Weight
1	Flat	140	140.0
2	Up	137	137.0
3	Down	21	21.0

2) Yêu cầu 2 – Gom cụm có xét thông tin phân lớp (5.0 điểm)

Như đã đề cập, nhiệm vụ của sinh viên trong bài tập này là áp dụng giải thuật SimpleKMeans của WEKA để kiểm chứng phát biểu của các bác sĩ khoa tim.
Một thực nghiệm độc lập với giải thuật Apriori đã được thực hiện trên thuộc tính gender và heart-condition (chính là thuộc tính class đã đổi tên) để tìm luật kết hợp thể hiện mối liên hệ giữa giới tính (Male/Female) và bệnh tim (Healhy/Sick).

a) Áp dụng giải thuật SimpleKMeans lên dữ liệu tiền xử lý (cardiology-cleaned) với thông số chỉ định như bên dưới

- Cluster mode: Use training set
- Xét cả thuộc tính lớp heart-condition
- Số lượng cụm: 3
- Display the standard deviation
- Distance function: Euclidean Distance



```
Final cluster centroids:
                                          Cluster#
Attribute                    Full Data           0           1           2
                               (303.0)     (110.0)      (76.0)     (117.0)
=============================================================================
age                           54.3233     51.1848     55.8816     56.2619
                             +/-9.0418   +/-8.7777   +/-9.5624   +/-8.1609

gender                           Male        Male      Female        Male
  Male                  207.0 ( 68%)  86.0 ( 78%)  18.0 ( 23%) 103.0 ( 88%)
  Female                 96.0 ( 31%)  24.0 ( 21%)  58.0 ( 76%)  14.0 ( 11%)

heart-condition               Healthy     Healthy     Healthy        Sick
  Sick                  138.0 ( 45%)  13.0 ( 11%)  13.0 ( 17%) 112.0 ( 95%)
  Healthy               165.0 ( 54%)  97.0 ( 88%)  63.0 ( 82%)   5.0 (  4%)
```

```
Time taken to build model (full training data) : 0.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      110 ( 36%)
1       76 ( 25%)
2      117 ( 39%)
```

Tổng cộng 303 thực thể, mang giới tính Nam(68%) và Healthy(54%)
Kết quả gom thành 3 nhóm:
Nhóm 0: 101 thực thể, có giới tính Nam(78%) và Healthy(89%)
Nhóm 1: 76 thực thể, có giới tính Nữ (76%) và Healthy(82%)
Nhóm 2: 117 thực thể, có giới tính Nam(88%) và Sick(95%)

- Dựa vào kết quả thực thi, hãy thông dịch kết quả tâm cụm tìm được để rút ra nhận định về các vấn đề sau
- Nam giới có nguy cơ mắc bệnh về tim cao hơn.

  Xét tổng thể: Tỉ lệ Nam 68% vượt trội so với Nữ, tuy nhiên tỉ lệ giữa Sick và Healthy không vượt trội quá nhiều. Bên cạnh đó, sau khi gom nhóm, có thể thấy rõ từ nhóm 2 tỉ lệ nam và nguy cơ mắc bệnh tim là hoàn toàn vượt trội

  ⇨ Đồng tình với nhận định "Nam giới có nguy cơ mắc bệnh về tim cao hơn"
- Nữ giới có nguy cơ mắc bệnh về tim thấp hơn.

  Tương tự như trên

  ⇨ Đồng tình với nhận định "Nữ giới có nguy cơ mắc bệnh về tim thấp hơn"

b) Nhận định từ kết quả gom cụm có nhất quán với mối liên hệ giữa giới tính và bệnh tim được Apriori cung cấp như trên? Giải thích lý do.

Apriori cho ra kết quả:
nếu là sick thì sẽ là male, độ tin cậy 0.83
nếu là female thì sẽ là healthy, độ tin cậy 0.75

⇨ Kết quả gom cụm nhất quán với kết quả ở apriori về mối liên hệ giữa giới tính và bệnh tim

c) Khảo sát với giá trị k thay đổi từ 2 đến 5. Lưu lại thông tin hiển thị ở Clusterer Output vào các tập tin <k>-clusters.txt, và nộp lại các tập tin. Tập tin sai định dạng hoặc thiếu dòng dữ liệu (-1đ). Không nộp tập tin (-2đ).

Đã hoàn thành

3) Yêu cầu 3 – Đánh giá kết quả gom cụm (5.0 điểm)

a) Áp dụng giải thuật k-means lên dữ liệu tiền xử lý (cardiology-cleaned) với thông số chỉ định như bên dưới để thu được 3 cụm dữ liệu

- Cluster mode: Use training set
- Bỏ qua thuộc tính lớp heart-condition
- Số lượng cụm: 3
- Display the standard deviation
- Distance function: Euclidean Distance

Lưu lại thông tin Clusterer Output vào tập tin 3-ts-classignored.txt và nộp lại tập tin. Tập tin sai định dạng hoặc thiếu dòng dữ liệu (-1đ). Không nộp tập tin (-2đ).
Đã hoàn thành.

b) Áp dụng giải thuật k-means lên dữ liệu tiền xử lý (cardiology-cleaned) với thông số chỉ định như bên dưới để thu được 3 cụm dữ liệu

- Cluster mode: Classes to cluster evaluation
- Bỏ qua thuộc tính lớp heart-condition
- Số lượng cụm: 3
- Display the standard deviation
- Distance function: Euclidean Distance

Lưu lại thông tin Clusterer Output vào tập tin 3-ce-classignored.txt và nộp lại tập tin. Tập tin sai định dạng hoặc thiếu dòng dữ liệu (-1đ). Không nộp tập tin (-2đ).
Đã hoàn thành.

c) Giả sử bạn nhận được kết quả gom cụm như bên dưới. Hãy diễn giải thông tin này.

Clustered Instances
0 113 (37%)
1 81 (27%)

2 109 (36%)
Số lượng và tỉ lệ thực thể của các nhóm:
Nhóm 0: 113 thực thể - 37%
Nhóm 1: 81 thực thể - 27%
Nhóm 2: 109 thực thể - 36%

d) Giả sử bạn nhận được kết quả gom cụm như bên dưới. Hãy diễn giải thông tin này.

Class attribute: heart-condition
Classes to Clusters:
0 1 2 <-- assigned to cluster
22 20 96 | Sick
91 61 13 | Healthy
Cluster 0 <-- Healthy
Cluster 1 <-- No class
Cluster 2 <-- Sick

Thuộc tính lớp là heart-condition
Chi tiết cho từng nhóm
Nhóm 0: 22 Sick, 91 Healthy
Nhóm 1: 20 Sick, 61 Healthy
Nhóm 2: 96 Sick, 13 Healthy
Xác định lớp cho từng nhóm:
Nhóm 0: lớp Healthy
Nhóm 1: không có
Nhóm 2: lớp Sick

II) Nội dung thực hiện cài đặt (15 điểm)

Cài đặt chương trình đọc vào tập dữ liệu cardiology-cleaned.arff, thực hiện gom cụm bằng giải thuật k-means rồi xuất ra tập tin kết quả.

1) (2.0đ) Chương trình nhận dữ liệu đầu vào là tập tin cardiology-cleaned.arff

```
D:\Source\CourseHCMUS\BigData\KTDLUD\Lab04\Source>python 37_Lab04_Clustering.py cardiology-cleaned.arff 2-clusters.txt 2
Đọc dữ liệu từ file arff

Lấy tên thuộc tính:
age, gender, chest-pain-type, blood-pressure, cholesterol, Fasting-blood-sugar<120, resting-ecg, maximum-heart-rate, angina, peak, slope, #colored-vessels, thal, heart-condition

Lấy dữ liệu:
(54.323333, b'Male', b'Asymptomatic', 130., 206., b'FALSE', b'Hyp', 132., b'TRUE', 2.4, b'Flat', 2., b'Rev', b'Sick')
(54.323333, b'Male', b'AbnormalAngina', 130., 266., b'FALSE', b'Normal', 171., b'FALSE', 0.6, b'Up', 0., b'Normal', b'Healthy')
(64., b'Male', b'Angina', 110., 211., b'FALSE', b'Hyp', 144., b'TRUE', 1.8, b'Flat', 0., b'Normal', b'Healthy')
(63., b'Male', b'Asymptomatic', 130., 254., b'FALSE', b'Hyp', 147., b'FALSE', 1.4, b'Flat', 1., b'Rev', b'Sick')
(53., b'Male', b'Asymptomatic', 140., 203., b'TRUE', b'Hyp', 155., b'TRUE', 3.1, b'Down', 0., b'Rev', b'Sick')
(58., b'Female', b'Angina', 150., 283., b'TRUE', b'Hyp', 162., b'FALSE', 1., b'Up', 0., b'Normal', b'Healthy')
(58., b'Male', b'AbnormalAngina', 120., 284., b'FALSE', b'Hyp', 160., b'FALSE', 1.8, b'Flat', 0., b'Normal', b'Sick')
(58., b'Male', b'NoTang', 132., 224., b'FALSE', b'Hyp', 173., b'FALSE', 3.2, b'Flat', 2., b'Rev', b'Sick')
(63., b'Male', b'Angina', 145., 233., b'TRUE', b'Hyp', 150., b'FALSE', 2.3, b'Down', 0., b'Fix', b'Healthy')
(54.323333, b'Male', b'Asymptomatic', 160., 286., b'FALSE', b'Hyp', 108., b'TRUE', 1.5, b'Flat', 3., b'Normal', b'Sick')
(67., b'Male', b'Asymptomatic', 120., 229., b'FALSE', b'Hyp', 129., b'TRUE', 2.6, b'Flat', 2., b'Rev', b'Sick')
(37., b'Male', b'NoTang', 130., 250., b'FALSE', b'Normal', 187., b'FALSE', 3.5, b'Down', 0., b'Normal', b'Healthy')
(41., b'Female', b'AbnormalAngina', 130., 204., b'FALSE', b'Hyp', 172., b'FALSE', 1.4, b'Up', 0., b'Normal', b'Healthy')
(56., b'Male', b'AbnormalAngina', 120., 236., b'FALSE', b'Normal', 178., b'FALSE', 0.8, b'Up', 0., b'Normal', b'Healthy')
(62., b'Female', b'Asymptomatic', 140., 268., b'FALSE', b'Hyp', 160., b'FALSE', 3.6, b'Down', 2., b'Normal', b'Sick')
```

2) (2.0đ) Chương trình phát sinh dữ liệu đầu ra là tập tin <k>-clusters.txt, với k là số cụm cần gom. Tập tin có nội dung tương tự như những gì hiển thị trong Clusterer Output của

WEKA 3.8 khi chạy chế độ Use training set, bao gồm

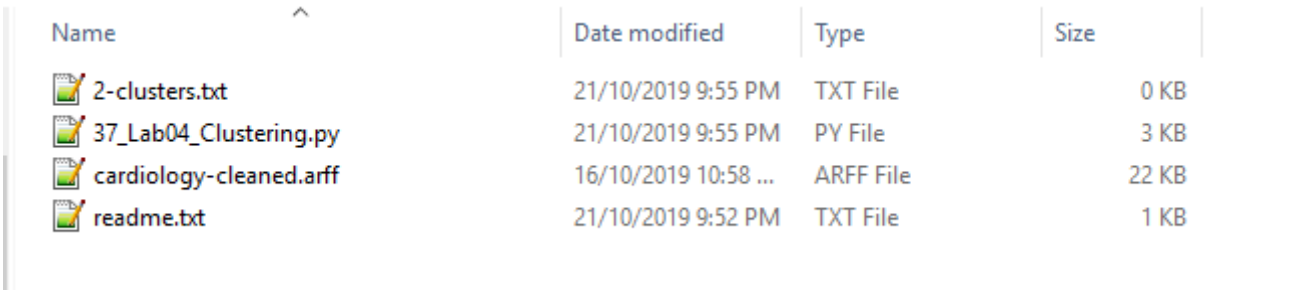
- Giá trị Sum of squared errors
- Thông tin gom cụm (Final cluster centroids)

3) (1.0đ) Chương trình thực thi giải thuật k-means với cú pháp tham số dòng lệnh là

<ID nhom> <input> <output_model> <k>

- <ID nhom>: tên của tập tin thực thi chương trình là ID của nhóm.
- <input>: tập tin dữ liệu đầu cardiology-cleaned.arff
- <output_model>: tập tin đầu ra k-clusters.txt
- <k>: số lượng cụm cần gom

Chương trình xử lý tuần tự các mẫu theo thứ tự từ trên xuống. Cần thể hiện ra màn hình console cho người dùng biết chương trình đang xử lý đến giai đoạn nào. Ví dụ: đang tính vòng lặp 1, đang tính vòng lặp 2, đang tính độ lỗi SSE, v.v.
Chương trình xuất ra giá trị độ đo đánh giá thuộc tính theo chiến lược đã chọn ra màn hình console trong quá trình tính toán.



Name	Date modified	Type	Size
2-clusters.txt	21/10/2019 9:55 PM	TXT File	0 KB
37_Lab04_Clustering.py	21/10/2019 9:55 PM	PY File	3 KB
cardiology-cleaned.arff	16/10/2019 10:58 ...	ARFF File	22 KB
readme.txt	21/10/2019 9:52 PM	TXT File	1 KB

4) (10.0đ) Đối chiếu kết quả phát sinh được với kết quả của WEKA (đã thực hiện ở phần Nội dung thực hiện báo cáo viết) trên cùng giá trị k (từ 2 đến 5).