LAB04 – k-means Clustering

Mục tiêu của bài tập

- Trải nghiệm tác vụ gom cụm dữ liệu bằng cách áp dụng giải thuật k-means được hỗ trợ bởi công cụ WEKA và tự cài đặt giải thuật k-means.
- Rèn luyện kỹ năng phân tích dữ liệu thông qua việc tiến hành thực nghiệm và nhận xét trên kết quả thu được.

Quy định

- Thời gian thực hiện: **2 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**. Các ngôn ngữ khác 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 30 rồi quy đổi về tỉ lệ 20% điểm thực hành.

LƯU Ý: BÀI TẬP NÀY SẼ THỰC HIỆN TRÊN WEKA 3.8

A – 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

No.	1: age	2: gender	3: chest-pain-type	4: blood-pressure	5: cholesterol	6: Fasting-blood-sugar120	7: resting-ecg	8: maximum-heart-rate	9: angina	10: peak	11: slope	12: #colored-vessels	13: thal	14: **class**
	Numeric	Nominal	Nominal	Numeric	Numeric	Nominal	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Nominal
1		Male	Asymptomatic	130.0	206.0	FALSE	Hyp	132.0	TRUE	2.4	Flat	2.0	Rev	Sick
2		Male	AbnormalAngina	130.0	266.0	FALSE	Normal	171.0	FALSE	0.6	Up	0.0	Nor...	Healthy
3	64.0	Male	Angina	110.0	211.0	FALSE	Hyp	144.0	TRUE	1.8	Flat	0.0	Nor...	Healthy
4	63.0	Male	Asymptomatic	130.0	254.0	FALSE	Hyp	147.0	FALSE	1.4	Flat	1.0	Rev	Sick
5	53.0	Male	Asymptomatic	140.0	203.0	TRUE	Hyp	155.0	TRUE	3.1	Down	0.0	Rev	Sick
6	58.0	Female	Angina	150.0	283.0	TRUE	Hyp	162.0	FALSE	1.0	Up	0.0	Nor...	Healthy
7	58.0	Male	AbnormalAngina	120.0	284.0	FALSE	Hyp	160.0	FALSE	1.8	Flat	0.0	Nor...	Sick
8	58.0	Male	NoTang	132.0	224.0	FALSE	Hyp	173.0	FALSE	3.2		2.0	Rev	Sick
9	63.0	Male	Angina	145.0	233.0	TRUE	Hyp	150.0	FALSE	2.3	Down	0.0	Fix	Healthy
10		Male	Asymptomatic	160.0	286.0	FALSE	Hyp	108.0	TRUE	1.5	Flat	3.0	Nor...	Sick

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 k-mean của WEKA (weka.clusterers.SimpleKMeans) để kiểm chứng phát biểu này.

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

1.1 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.

1.2 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).

1.3 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?

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).

Dưới đây là kết quả chạy Apriori.

```
=== Run information ===

Scheme:       weka.associations.Apriori -I -N 10 -T 0 -C 0.7 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     cardiology-weka.filters.unsupervised.attribute.RenameAttribute-find([\s\S]+)-replacegender-R2-weka.fil
Instances:    303
Attributes:   2
              gender
              heart-condition
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (30 instances)
Minimum metric <confidence>: 0.7
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 4

Large Itemsets L(1):
gender=Male 207
gender=Female 96
heart-condition=Sick 138
heart-condition=Healthy 165

Size of set of large itemsets L(2): 3

Large Itemsets L(2):
gender=Male heart-condition=Sick 114
gender=Male heart-condition=Healthy 93
gender=Female heart-condition=Healthy 72

Best rules found:

 1. heart-condition=Sick 138 ==> gender=Male 114    <conf:(0.83)> lift:(1.21) lev:(0.07) [19] conv:(1.75)
 2. gender=Female 96 ==> heart-condition=Healthy 72    <conf:(0.75)> lift:(1.38) lev:(0.07) [19] conv:(1.75)
```

2.1 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

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

a. Nam giới có nguy cơ mắc bệnh về tim cao hơn.

b. Nữ giới có nguy cơ mắc bệnh về tim thấp hơn.

2.2 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.

2.3 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).

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

3.1 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).

3.2 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).

3.3 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%)

3.4 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

B – 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.

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

(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)

(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.

(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).