LAB02 – Association Rules

Mục tiêu của bài tập

- Sử dụng công cụ WEKA để thực hiện khai thác tập phổ biến và luật kết hợp bằng các giải thuật cơ bản như Apriori và FP-growth
- Rèn luyện kỹ năng lập trình bằng việc cài đặt giải thuật khai thác tập phổ biến và luật kết hợp.

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 ⟨**tên nhóm**⟩ (nếu tên nhóm có dấu và khoảng trắng thì bỏ dấu và viết dính liền), 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 25% điểm thực hành.

A – Nội dung thực hiện báo cáo viết (15 điểm)

Trong bài tập thực hành này, sinh viên làm việc trên Bảng Associate, một trong các chức năng chính của Giao diện Explorer, và giải thuật được đề cập là Apriori (weka.associations.Apriori).

Bảng Associate có giao diện tương tự với Bảng Classify, ngoại trừ việc Bảng Associate không có hộp Test options và trường chọn lựa thuộc tính lớp, vì cả hai tính năng này không liên quan đến luật kết hợp. Luật kết hợp không xem một thuộc tính nào đó có vai trò đặc biệt, điều này khác với cách bài toán phân lớp đối xử với thuộc tính lớp. Chức năng kiểm thử cũng không cần thiết trong tình huống này vì khai thác luật kết hợp được xem là tác vụ khảo sát dữ liệu hơn là dự đoán, và điều này cũng có nghĩa là ta không cần đánh giá độ chính xác ở đây.

Ta hãy khảo sát dữ liệu đầu ra của giải thuật khai thác luật kết hợp Apriori. Đọc một tập dữ liệu nào đó, ví dụ vote.arff (được mô tả bên dưới), vào Weka và chuyển sang Bảng Associate để chọn giải thuật Apriori. Sau khi nút Start được nhấn, Apriori bắt đầu xây dựng mô hình và ghi dữ liệu đầu ra vào trường Associator output. Phần đầu của kết quả (Run information) trình bày các thông số đã được chọn lựa tập dữ liệu đang sử dụng.

```
1=== Run information ===
Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -
c -1
Relation: vote
Instances: 435
Attributes: 17
handicapped-infants
water-project-cost-sharing
adoption-of-the-budget-resolution
physician-fee-freeze
el-salvador-aid
religious-groups-in-schools
anti-satellite-test-ban
aid-to-nicaraguan-contras
mx-missile
immigration
synfuels-corporation-cutback
education-spending
superfund-right-to-sue
crime
duty-free-exports
export-administration-act-south-africa
Class
```

Phần kế tiếp là thông tin về các thông số đạt được trong quá trình chạy giải thuật, các tập hạng mục phổ biến và luật kết hợp được phát sinh.

```
=== Associator model (full training set) ===
Apriori
=======
```

```
Minimum support: 0.45 (196 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 11
Generated sets of large itemsets:
Size of set of large itemsets L(1): 20
Size of set of large itemsets L(2): 17
Size of set of large itemsets L(3): 6
Size of set of large itemsets L(4): 1
Best rules found:
 1.  physician-fee-freeze=n  education-spending=n  202  ==>  Class=democrat  201
conf:(1)
 2. physician-fee-freeze=n 247 ==> Class=democrat 245    conf:(0.99)
```

Weka mặc định hiển thị 10 luật có độ tin cậy cao nhất, và có thể tùy chỉnh số luật được hiển thị thông qua danh sách tham số của giải thuật. Mỗi luật có dạng A ==> B, trong đó A và B chứa một vài giá trị thuộc tính. Lấy ví dụ luật 1, đây có nghĩa là nếu giá trị cho 'adoption-of-thebudget-resolution' is 'y' và giá trị cho 'physician-fee-freeze' là 'n' thì giá trị cho thuộc tính 'class' được dự đoán là 'democrat' (Lưu ý rằng không phải mọi luật đều có phần vế phải là thuộc tính lớp). Giá trị trước mũi tên là số mẫu áp dụng luật (khớp tiền đề), và giá trị sau mũi tên là số mẫu được luật dự đoán chính xác (hay nói cách khác là số mẫu thỏa mãn luật, là những mẫu khớp cả hai vế). Giá trị số nằm trong dấu ngoặc sau conf: là độ tin cậy của luật.

Dữ liệu thực nghiệm

Các thực nghiệm trong bài tập này sẽ làm việc trên ba tập dữ liệu như bên dưới

- **vote.arff:** Thông tin bỏ phiếu của các Hạ nghị sĩ Hoa kỳ về 16 vấn đề quan trọng. Mỗi mẫu trong tập dữ liệu biểu diễn thông tin bỏ phiếu của một Hạ nghị sĩ và đảng phái của người này. Thuộc tính Class chỉ áp dụng cho bài toán phân lớp.
- **weather.nominal.arff**: Dữ liệu về các quyết định đi chơi hay không đi chơi tùy thuộc vào tình hình thời tiết (quang cảnh, nhiệt độ, v.v.), có kích thước rất nhỏ. Tương tự như trên, ta không quan tâm thuộc tính lớp play trong ngữ cảnh này.
- **supermarket.arff:** Dữ liệu ghi nhận thói quen mua sắm của các khách hàng siêu thị. Hầu hết các thuộc tính biểu diễn cho từng nhóm mặt hàng cụ thể. Giá trị 't' tại một thuộc tính có nghĩa là khách hàng đã mua mặt hàng đó, và giá trị thiếu '?' cho tình huống ngược lại. Mỗi mẫu trong tập dữ liệu biểu diễn cho một khách hàng. Không có thuộc tính lớp trong tập dữ liệu vì điều này thường không cần thiết cho bài toán phân tích giỏ mua hàng.

Yêu cầu 1 – Khảo sát luật kết hợp trên dữ liệu vote.arff

1.1 Chạy giải thuật Apriori với tham số mặc định. 10 dòng cuối cùng của phần kết quả, bên dưới dòng "Best rules found", là 10 luật được chọn hiển thị trong số các luật đã phát sinh. Độ tin cậy (confidence) của luật 10 là 0.96. Giá trị này được tính như thế nào? Hãy trình bày công thức và thế giá trị cụ thể vào các phép tính.

1.2 Có bao nhiêu mẫu được xét khi cần tính độ hỗ trợ (support) của luật 8?

1.3 "Số lượng mẫu có thể áp dụng luật" nghĩa là gì? Giải thích bằng ví dụ là luật 7.

1.4 "Số lượng mẫu được dự đoán chính xác bởi luật" có nghĩa là gì? Giải thích thông qua ví dụ là luật 9.

1.5 Khảo sát phần mô tả tham số của Apriori bằng cách nhấn vào nút "More" trong cửa sổ tùy chọn tham số GenericObjectEditor. Hãy thử thay đổi số luật được hiển thị trong kết quả. Bạn có nghĩ rằng số luật được phát sinh có thể nhiều hơn 100 không? Giải thích. Chụp màn hình các tình huống số lượng luật hiển thị khác nhau làm minh chứng.

1.6 "Luật tốt nhất" nghĩa là gì? Tiêu chí nào được dùng để xác định các luật tốt nhất?

1.7 Luật nào nói về khả năng một Hạ nghị sĩ không bỏ phiếu cho 'el-salvador-aid' thì người này sẽ bỏ phiếu cho 'aid-to-nicaraguan-contras'? Tương tự, luật nào có kèm theo điều kiện về đảng phái của Hạ nghị sĩ này?

1.8 Một số luật trong 10 luật tốt nhất có vế phải là 'Class = democrat'. Vấn đề này nói lên điều gì về thói quen bỏ phiếu của các Hạ nghị sĩ đảng Dân chủ?

Yêu cầu 2 – Khảo sát luật kết hợp trên dữ liệu weather.nominal.arff

2.1 Xét luật **temperature=hot ==> humidity=normal**. Số lượng mẫu áp dụng luật là bao nhiêu? Số lượng mẫu thỏa mãn luật là bao nhiêu? Tính độ hỗ trợ và độ tin cậy của luật.

2.2 Xét luật **temperature=hot humidity=high ==> windy=TRUE**. Số lượng mẫu áp dụng luật là bao nhiêu? Số lượng mẫu thỏa mãn luật là bao nhiêu? Tính độ hỗ trợ và độ tin cậy của luật.

2.3 Điều chỉnh số lượng luật được hiển thị, giá trị minsup, và giá trị minconf, nếu cần thiết. Luật ở Câu 2.1 và Câu 2.2 nằm ở vị trí thứ mấy trong danh sách luật được tìm thấy? Chụp màn hình có hiển thị phần luật tương ứng làm minh chứng.

2.4 Một luật có thể có hai (hay nhiều) thuộc tính ở vế phải được không? Ví dụ,

outlook=sunny temperature=cool ==> humidity=normal play=yes

Hãy đưa ra 5 luật khác luật đã nêu ở trên để làm ví dụ.

Yêu cầu 3 – Khảo sát luật kết hợp trên dữ liệu supermarket.arff

Luật kết hợp được sử dụng chủ yếu cho việc phân tích dữ liệu để thăm dò. Hãy áp dụng Apriori để phát sinh luật kết hợp và dựa trên những luật này để có những nhận xét về thói quen mua sắm của khách hàng siêu thị. Cần phát sinh khoảng 30 luật.

Trong nhiều tình huống đa dạng, tác vụ phân tích sẽ quan tâm đến những luật chỉ có một thuộc tính đặc biệt ở vế phải, đó là thuộc tính lớp (hay tiêu chí phân loại nào đó). Các luật này được phát sinh bằng cách thiết lập tham số đầu tiên car (tức là class association rules) thành 'true' và tham số thứ hai thành giá trị chỉ mục (bắt đầu từ 0) của thuộc tính mà bạn cần nó xuất hiện ở vế phải.

3.1 Khảo sát một vài luật được phát sinh và mô tả một quan sát về thói quen mua sắm của khách hàng siêu thị mà bạn rút ra được từ việc khảo sát này. Dẫn chứng những luật kết hợp liên quan đến quan sát được đưa ra.

3.2 Tương tự câu hỏi trên, hãy mô tả quan sát thứ hai mà bạn có được, đồng thời dẫn chứng những luật kết hợp liên quan.

3.3 Các quan sát của bạn ở hai câu hỏi trên có gợi ý gì cho nhà quản lý siêu thị nhằm đưa ra hành động có ích cho việc kinh doanh? Nếu có, giải thích quan sát hữu ích như thế nào đến hoạt động của siêu thị. Nếu không, giải thích tại sao quan sát không có giá trị.

B – Nội dung thực hiện cài đặt (15 điểm)

Cài đặt chương trình đọc vào một tập dữ liệu giao dịch bất kỳ có định dạng *.csv, thực hiện khai thác tập phổ biến và luật kết hợp rồi xuất ra tập tin kết quả.

(2.0đ) Chương trình nhận dữ liệu đầu vào là **tập tin có định dạng *.csv** có cấu trúc như sau

- Giả sử tập dữ liệu có N hạng mục và M giao dịch chứa các hạng mục thuộc tập N hạng mục này. Dữ liệu sẽ được tổ chức thành bảng có M+1 dòng và N cột.
- Dòng đầu tiên chứa tên của N hạng mục, phân cách nhau bằng dấu phẩy (",").
- M dòng tiếp theo, mỗi dòng gồm N giá trị, phân cách nhau bằng dấu phẩy (","), nếu một hạng mục có trong giao dịch thì giá trị tương ứng là "y" (yes), ngược lại, giá trị là "n".

(6.0đ) Chương trình phát sinh dữ liệu đầu ra bao gồm hai tập tin, **FI.txt** và **AR.txt**, lần lượt chứa các tập phổ biến và luật kết hợp khai thác được từ dữ liệu đầu vào theo giá trị minsup và minconf đặc tả trong tham số dòng lệnh.

- Tập tin FI.txt chứa các tập phổ biến khai thác được với tham số minsup, có định dạng như sau
  - Dòng đầu tiên là một số nguyên không âm F_1 chỉ số lượng tập phổ biến 1-hạng mục.
  - F_1 dòng tiếp theo, mỗi dòng trình bày 1 tập phổ biến 1-hạng mục. Các hạng mục trong tập phổ biến cách nhau bởi khoảng trắng.
  - Độ hỗ trợ được ghi ở đầu dòng của tập phổ biến tương ứng, cách tập hạng mục bằng khoảng trắng. Giá trị độ hỗ trợ là số thực, làm tròn 2 chữ số sau dấu phẩy.
  - Thực hiện tương tự cho các tập phổ biến 2-hạng mục, 3-hạng mục,…

    Ví dụ:

    2 #có 2 tập phổ biến 1-hạng mục

    0.9 1 #tập phổ biến {1} có sup = 0.9

    0.83 2

    1 #có 1 tập phổ biến 2-hạng mục

    0.75 1 2 #tập phổ biến {1, 2} có sup = 0.75
- Tập tin AR.txt chứa các luật kết hợp được phát sinh từ các tập phổ biến trong FI.txt với tham số minconf, có định dạng tương tự như tập tin FI.txt, thay khái niệm tập k-hạng mục bằng luật kết hợp phát sinh từ tập k-hạng mục và thay giá trị độ hỗ trợ sup bằng giá trị độ tin cậy conf.

(2.0đ) Chương trình thực thi **giải thuật Apriori** với cú pháp tham số dòng lệnh như sau

<ID nhom> <input> <output FI> <output AR> <minsup> <minconf>

- <ID nhom>: tên của tập tin thực thi chương trình là ID của nhóm.
- <input>: tập tin dữ liệu giao dịch đầu vào
- <output FI>: tập tin đầu ra FI.txt chứa danh sách tập hạng mục phổ biến
- <output AR>: tập tin đầu ra AR.txt chứa danh sách luật kết hợp phát sinh được từ các tập phổ biến trong FI.txt
- <minsup>: giá trị độ phổ biến tối thiểu
- <minconf>: giá trị độ tin cậy tối thiểu

Chương trình **xử lý tuần tự các giao dịch theo thứ tự từ trên xuống**. Cần thể hiện thông báo ra màn hình console cho người dùng biết chương trình đang xử lý đến giai đoạn nào. Ví dụ: đang xây dựng cây FP-Tree, đang ghi tập tin FI.txt,…

(5.0đ) Chạy chương trình cài đặt với các tập dữ liệu đã cho ở Phần A. Thử nghiệm 5 kịch bản khác nhau (tức là thay đổi tập dữ liệu và/hoặc tham số của giải thuật). Đối chiếu kết quả phát sinh được với kết quả của WEKA-Apriori trên cùng bộ tham số.