



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP HCM

Khoa Công nghệ thông tin

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

BÀI TẬP 02

Khai thác mẫu phổ biến và luật kết hợp
Association Rules


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

A)	Nội dung thực hiện báo cáo viết (15 điểm)	2
1)	Yêu cầu 1 – Khảo sát luật kết hợp trên dữ liệu vote.arff.....	2
2)	Yêu cầu 2 – Khảo sát luật kết hợp trên dữ liệu weather.nominal.arff	5
3)	Yêu cầu 3 – Khảo sát luật kết hợp trên dữ liệu supermarket.arff	7
B)	Nội dung thực hiện cài đặt (15 điểm)	8
1)	Nhận dữ liệu đầu vào:	8
2)	Phát sinh dữ liệu đầu ra:.....	9
3)	Thực thi giải thuật Apriori.....	10
4)	Chạy chương trình – So sánh với phần A	11

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

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

- a) **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.**

Số mẫu được luật dự đoán chính xác chia cho số mẫu áp dụng luật(khớp tiền đề)

Ví dụ: trong luật 10:

`10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210`

Số mẫu được luật dự đoán chính xác: 210

Số mẫu áp dụng luật: 218

$210/218 = 0.96$

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

203

- c) **“Số lượng mẫu có thể áp dụng luật” nghĩa là gì? Giải thích bằng ví dụ là luật 7.**

“Số lượng mẫu có thể áp dụng luật” là số lượng mẫu thỏa vế trước của luật.

Xét luật 7:

`7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204`

Số mẫu thỏa el-salvador-aid=n là 208 mẫu

- d) **“Số lượng mẫu được dự đoán chính xác bởi luật” có nghĩa là gì? Giải thích thông qua ví dụ là luật 9.**

“Số lượng mẫu được dự đoán chính xác bởi luật” là số lượng mẫu thỏa vế trước và vế sau của luật.

Xét luật 9:

9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197

Có 204 mẫu thỏa el-salvador-aid=n và aid-to-nicaraguan-contras=y

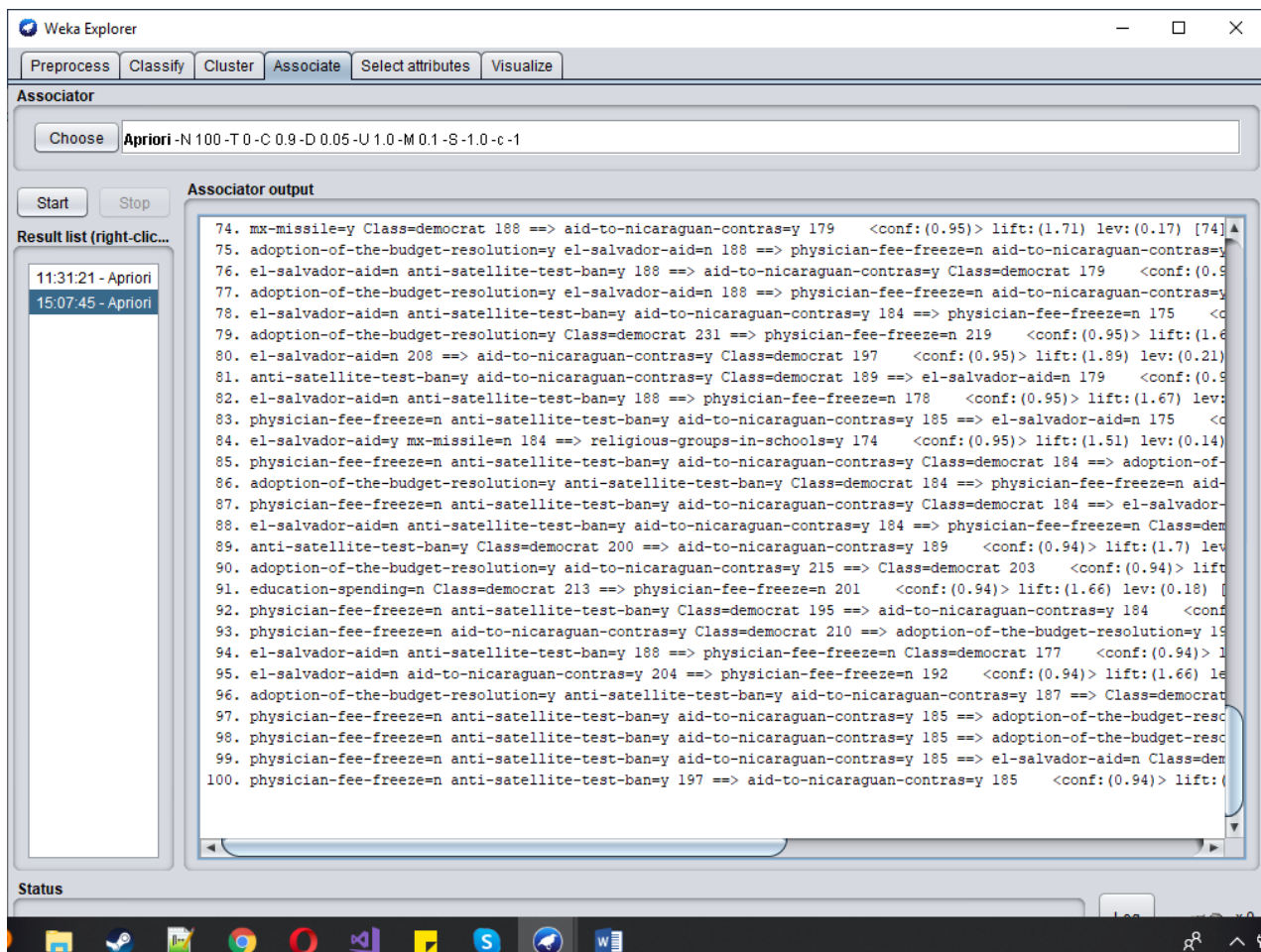
Trong 204 mẫu này có cả mẫu thuộc class democrat và class republican

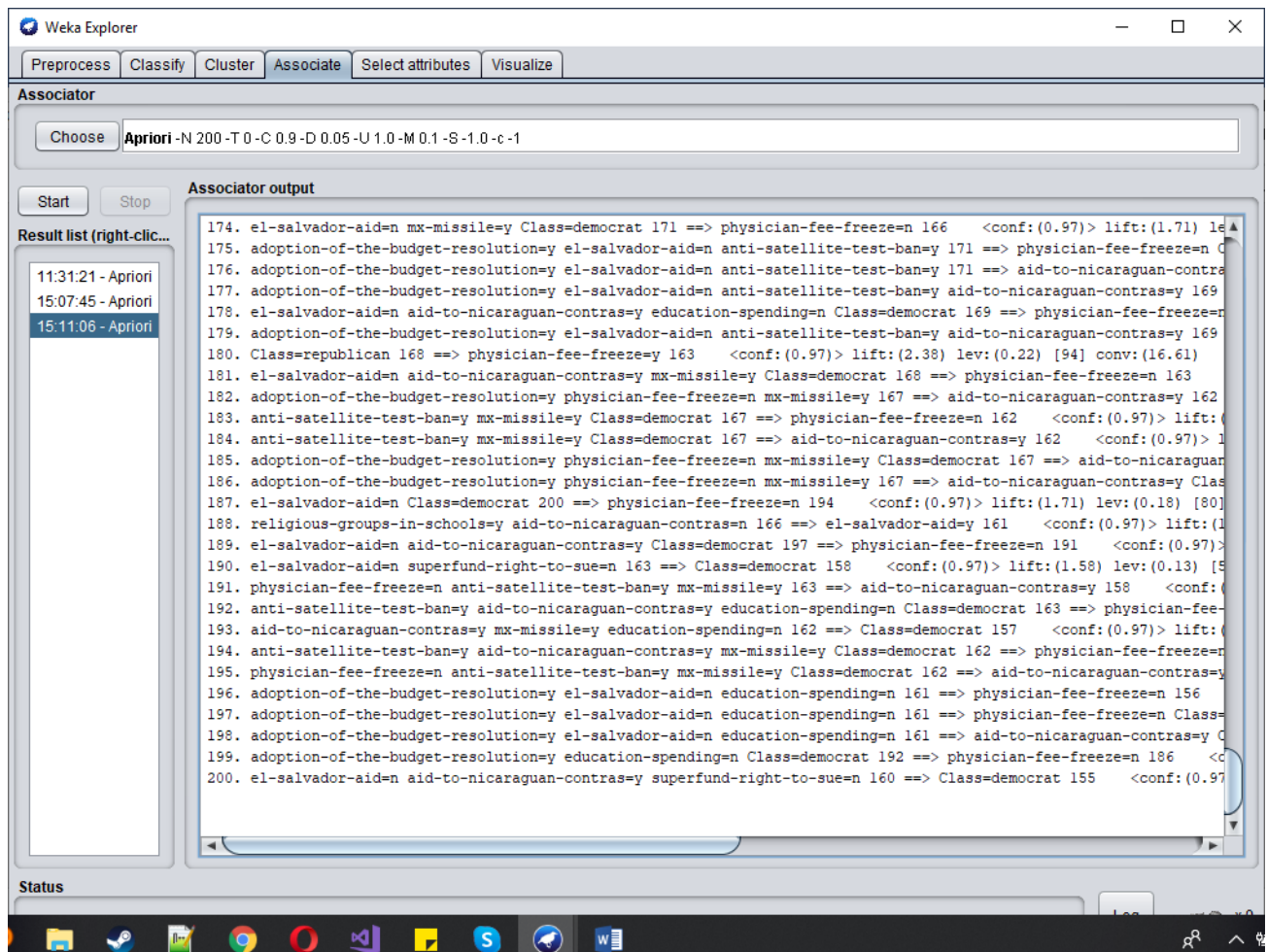
Trong đó có 197 mẫu thuộc class democrat

Tức là có 197 thỏa el-salvador-aid=n và aid-to-nicaraguan-contras=y và thỏa điều kiện class = democrat.

- e) 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.

Với Instances là 435 và 17 attributes, bên cạnh đó, việc ta có thể thay đổi minsupp và minconf thì số mẫu có thể nhiều hơn 100.





f) “Luật tốt nhất” nghĩa là gì? Tiêu chí nào được dùng để xác định các luật tốt nhất?

“Luật tốt nhất” là luật có độ tin cậy và độ phổ biến cao nhất. Dùng độ tin cậy và độ phổ biến để xác định các luật tốt nhất.

g) 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?

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’: luật 7

Luật nào có kèm theo điều kiện về đảng phái của Hạ nghị sĩ này: luật 6

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

Họ thường bỏ phiếu giống nhau – thói quen chia sẻ thông tin bỏ phiếu hoặc kêu gọi bỏ phiếu giống nhau.

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

- a) 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.

Số lượng mẫu áp dụng luật: 4

Số lượng mẫu thỏa mãn luật: 1

Độ hỗ trợ: 1/14

Độ tin cậy: 0.25

- b) 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.

Số lượng mẫu áp dụng luật: 3

Số lượng mẫu thỏa mãn luật: 1

Độ hỗ trợ: 1/14

Độ tin cậy: 0.33

- c) 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.

Luật của 2.1 `temperature=hot ==> humidity=normal` nằm ở vị trí 1338

Lab02 – Association Rules

Weka Explorer

Preprocess | Classify | Cluster | **Associate** | Select attributes | Visualize

Choose **Apriori** -N 5000 -T 0 -C 0.1 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c 1

Start Stop

Associator output

Result list (right-click...)

- 11:31:21 - Apriori
- 15:07:45 - Apriori
- 15:11:06 - Apriori
- 15:19:46 - Apriori
- 15:20:43 - Apriori
- 15:31:58 - Apriori
- 17:13:57 - Apriori
- 17:14:53 - Apriori
- 17:18:14 - Apriori

1322. windy=FALSE 8 ==> temperature=hot play=yes 2 <conf:(0.25)> lift:(1.75) lev:(0.06) [0] conv:(0.98)

1323. windy=FALSE 8 ==> temperature=mild humidity=high 2 <conf:(0.25)> lift:(0.88) lev:(-0.02) [0] conv:(0.82)

1324. windy=FALSE 8 ==> temperature=mild play=yes 2 <conf:(0.25)> lift:(0.88) lev:(-0.02) [0] conv:(0.82)

1325. windy=FALSE 8 ==> temperature=cool humidity=normal 2 <conf:(0.25)> lift:(0.88) lev:(-0.02) [0] conv:(0.82)

1326. windy=FALSE 8 ==> temperature=cool play=yes 2 <conf:(0.25)> lift:(1.17) lev:(0.02) [0] conv:(0.9)

1327. windy=FALSE 8 ==> humidity=high play=yes 2 <conf:(0.25)> lift:(1.17) lev:(0.02) [0] conv:(0.9)

1328. windy=FALSE 8 ==> humidity=high play=no 2 <conf:(0.25)> lift:(0.88) lev:(-0.02) [0] conv:(0.82)

1329. windy=FALSE 8 ==> outlook=sunny humidity=high play=no 2 <conf:(0.25)> lift:(1.17) lev:(0.02) [0] conv:(0.9)

1330. windy=FALSE 8 ==> outlook=overcast temperature=hot play=yes 2 <conf:(0.25)> lift:(1.75) lev:(0.06) [0] conv:(0.9)

1331. windy=FALSE 8 ==> outlook=rainy temperature=mild play=yes 2 <conf:(0.25)> lift:(1.75) lev:(0.06) [0] conv:(0.9)

1332. windy=FALSE 8 ==> outlook=rainy humidity=normal play=yes 2 <conf:(0.25)> lift:(1.75) lev:(0.06) [0] conv:(0.9)

1333. windy=FALSE 8 ==> temperature=cool humidity=normal play=yes 2 <conf:(0.25)> lift:(1.17) lev:(0.02) [0] conv:(0.9)

1334. temperature=cool 4 ==> outlook=sunny 1 <conf:(0.25)> lift:(0.7) lev:(-0.03) [0] conv:(0.64)

1335. outlook=overcast 4 ==> temperature=mild 1 <conf:(0.25)> lift:(0.58) lev:(-0.05) [0] conv:(0.57)

1336. temperature=cool 4 ==> outlook=overcast 1 <conf:(0.25)> lift:(0.88) lev:(-0.01) [0] conv:(0.71)

1337. outlook=overcast 4 ==> temperature=cool 1 <conf:(0.25)> lift:(0.88) lev:(-0.01) [0] conv:(0.71)

1338. temperature=hot 4 ==> humidity=normal 1 <conf:(0.25)> lift:(0.5) lev:(-0.07) [-1] conv:(0.5)

1339. temperature=hot 4 ==> windy=TRUE 1 <conf:(0.25)> lift:(0.58) lev:(-0.05) [0] conv:(0.57)

1340. temperature=cool 4 ==> play=no 1 <conf:(0.25)> lift:(0.7) lev:(-0.03) [0] conv:(0.64)

1341. temperature=hot 4 ==> outlook=sunny windy=TRUE 1 <conf:(0.25)> lift:(1.75) lev:(0.03) [0] conv:(0.86)

1342. temperature=hot 4 ==> outlook=sunny windy=FALSE 1 <conf:(0.25)> lift:(1.17) lev:(0.01) [0] conv:(0.79)

1343. temperature=mild humidity=high 4 ==> outlook=sunny 1 <conf:(0.25)> lift:(0.7) lev:(-0.03) [0] conv:(0.64)

1344. temperature=mild play=yes 4 ==> outlook=sunny 1 <conf:(0.25)> lift:(0.7) lev:(-0.03) [0] conv:(0.64)

1345. temperature=cool humidity=normal 4 ==> outlook=sunny 1 <conf:(0.25)> lift:(0.7) lev:(-0.03) [0] conv:(0.64)

1346. temperature=cool 4 ==> outlook=sunny humidity=normal 1 <conf:(0.25)> lift:(1.75) lev:(0.03) [0] conv:(0.86)

1347. temperature=cool 4 ==> outlook=sunny windy=FALSE 1 <conf:(0.25)> lift:(1.17) lev:(0.01) [0] conv:(0.79)

1348. temperature=cool 4 ==> outlook=sunny play=yes 1 <conf:(0.25)> lift:(1.75) lev:(0.03) [0] conv:(0.86)

1349. humidity=normal windy=FALSE 4 ==> outlook=sunny 1 <conf:(0.25)> lift:(0.7) lev:(-0.03) [0] conv:(0.64)

1350. temperature=hot 4 ==> outlook=overcast humidity=high 1 <conf:(0.25)> lift:(1.75) lev:(0.03) [0] conv:(0.86)

Status

Luật của 2.2 temperature=hot humidity=high ==> windy=TRUE nằm ở vị trí 1370

d) 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.

Có thể có hai (hay nhiều) thuộc tính ở vế phải.

Ví dụ:

temperature=hot play=yes 2 ==> outlook=overcast windy=FALSE 2
 outlook=overcast windy=FALSE 2 ==> temperature=hot play=yes 2
 outlook=overcast temperature=hot 2 ==> windy=FALSE play=yes 2
 temperature=cool windy=FALSE 2 ==> humidity=normal play=yes 2
 temperature=cool windy=FALSE 2 ==> humidity=normal play=yes 2

3) 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.

- a) **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.**

Phần lớn khách hàng hay mua “bread and cake” dù là mua cái gì trước đó.

```
biscuits=t 2605 ==> bread and cake=t 2083    <conf:(0
milk-cream=t 2939 ==> bread and cake=t 2337    <conf:
fruit=t 2962 ==> bread and cake=t 2325    <conf:(0.78
baking needs=t 2795 ==> bread and cake=t 2191    <con
frozen foods=t 2717 ==> bread and cake=t 2129    <con
vegetables=t 2961 ==> bread and cake=t 2298    <conf:
juice-sat-cord-ms=t 2463 ==> bread and cake=t 1869
```

- b) **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.**

Phần lớn khách hàng nếu mua rau củ thì cũng sẽ mua trái cây, và ngược lại.

```
vegetables=t 2961 ==> fruit=t 2207    <conf:(0.75)>
fruit=t 2962 ==> vegetables=t 2207    <conf:(0.75)>
```

- c) **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.**

Các quan sát ở trên có thể giúp nhà quản lý siêu thị đưa ra các dịch vụ có lợi cho việc kinh doanh.

Ở trường hợp 1: vì khách hàng hay mua “bread and cake” nên quản lý có thể đưa ra hình thức giảm giá khi mua “bread and cake” với bất kỳ các mặt hàng khác. Như vậy sẽ thúc đẩy các mặt hàng trước đây ko mua cùng với “bread and cake” lên.

Ở trường hợp 2: vì khách hàng mua trái cây thì sẽ mua rau củ và ngược lại, nên quản lý có thể đưa ra các gói hàng/combo cho rau củ và trái cây.

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.

1) Nhận dữ liệu đầu vào:

(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”.

Beef	Chicken	Milk		
Beef	Cheese			
Cheese	Boots			
Beef	Chicken	Cheese		
Beef	Chicken	Clothes	Cheese	Milk
Chicken	Clothes	Milk		
Chicken	Milk	Clothes		

Hình 1: Mẫu dữ liệu

Beef	Boots	Cheese	Chicken	Clothes	Milk
y	n	n	y	n	y
y	n	y	n	n	n
n	y	y	n	n	n
y	n	y	y	n	n
y	n	y	y	y	y
n	n	n	y	y	y
n	n	n	y	y	y

Hình 2: Chuyển đổi theo đề bài

File input.csv và input_raw.csv đính kèm

2) Phát sinh dữ liệu đầu ra:

(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 F1 chỉ số lượng tập phổ biến 1-hạng mục.
 - F1 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,...

```
0.57 Beef
0.57 Cheese
0.71 Chicken
0.43 Clothes
0.57 Milk
0.43 Beef, Cheese
0.43 Beef, Chicken
0.43 Chicken, Clothes
0.57 Chicken, Milk
0.43 Clothes, Milk
0.43 Chicken, Clothes, Milk
```

Hình 3: Tập tin FI.txt chứa tập phổ biến

- 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

bảng 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.

```
1.0 Clothes -> Chicken
0.8 Chicken -> Milk
1.0 Milk -> Chicken
1.0 Clothes -> Milk
1.0 Clothes -> Milk, Chicken
1.0 Clothes -> Chicken
1.0 Milk -> Chicken
0.8 Chicken -> Milk
1.0 Clothes -> Milk
1.0 Milk, Clothes -> Chicken
1.0 Chicken, Clothes -> Milk
```

Hình 4: Tập tin AR.txt chứa luật kết hợp (có các luật trùng nhau)

Chưa hoàn thiện

3) Thực thi giải thuật Apriori

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

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

- <ID nhóm>: 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,...

```
Microsoft Windows [Version 10.0.17763.678]
(c) 2018 Microsoft Corporation. All rights reserved.

D:\Source\CourseHCMUS\BigData\KTDLUD\Lab02\AssociationRules>python Team37_v2.py input.csv FI.txt AR.txt 0.3 0.8
----- TẬP PHỔ BIẾN -----
----- Tìm tập phổ biến 1 hạng mục -----
Tìm được: 5 tập phổ biến
0.57 Beef
0.57 Cheese
0.71 Chicken
0.43 Clothes
0.57 Milk
----- Tìm tập phổ biến 2 hạng mục -----
Tìm được: 5 tập phổ biến
0.43 Beef, Cheese
0.43 Beef, Chicken
0.43 Chicken, Clothes
0.57 Chicken, Milk
0.43 Clothes, Milk
----- Tìm tập phổ biến 3 hạng mục -----
Tìm được: 1 tập phổ biến
0.43 Chicken, Clothes, Milk
----- LUẬT KẾT HỢP -----
Tìm được: 5 luật kết hợp
1.0 Clothes -> Chicken
0.8 Chicken -> Milk
1.0 Milk -> Chicken
1.0 Clothes -> Milk
1.0 Clothes -> Chicken, Milk
Tìm được thêm: 6 luật kết hợp (có các luật trùng)!
1.0 Clothes -> Chicken
1.0 Milk -> Chicken
0.8 Chicken -> Milk
1.0 Clothes -> Milk
1.0 Milk, Clothes -> Chicken
1.0 Chicken, Clothes -> Milk
```

Hình 5: Thực thi giải thuật Apriori

File kết quả đính kèm

4) Chạy chương trình – So sánh với phần A

(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