TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

BÁO CÁO BÀI THỰC HÀNH 2: KHAI THÁC LUẬT KẾT HỢP

GVHD: Lê Ngọc Thành, Nguyễn Ngọc Thảo

TP. Hồ Chí Minh, ngày 20 tháng 04 năm 2019

Mục lục

I. Thông tin thành viên: 3
II. Báo cáo tổng quát: 3
III. Báo cáo chi tiết: 3
A. Lý thuyết 3
1. Phương pháp cải tiến quá trình tìm luật kết hợp từ tập phổ biến 3
2. Thuật toán Apriori và FP-Growth 4
B. Thực hành 9
1. Tạo tập tin plants.csv 9
2. Tìm hiểu tập dữ liệu plants 9
3. Tạo tập tin plants.arff 13
4. Khai thác tập phổ biến bằng thuật toán Apriori 13
5. Khai thác luật kết hợp bằng thuật toán FP-Growth 14
THAM KHẢO: 15

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:

A. Lý thuyết

1. Hãy tìm hiểu trong tài liệu tham khảo và trình bày chi tiết một phương pháp cải tiến quá trình tìm luật kết hợp từ tập phổ biến. Giải thích vì sao nó hiệu quả hơn.

- Trong tài liệu tham khảo ta chọn thuật toán khai thác luật thu gọn.
- **Điểm khác biệt:**

  + Ở thuật toán khai thác luật truyền thống là luật sinh ra giữa các X, Y thuộc FI, với X là con Y.

  + Ở thuật toán khai thác luật thu gọn luật kết hợp là luật X suy từ tập phổ biến sang tập phổ biến cha Y của nó trên tập FI, với |Y|=|X|+1.

- **Thuật toán:**

```
SHORTEN_AR()
   SORT (FI)//SX tập FI tăng theo k-itemset
   AR = ∅
   for each Y ∈ FI  with |Y| > 1 do
      for each X ∈ FI with |X| = |Y|-1 do
         if  X ⊂ Y  then
        conf = Sup(Y)/ Sup(X)
        if  conf  ≥ minConf  then
           AR = AR ∪ {X →Y \X (Sup(Y), conf)}
   return AR
```

- **Giải thích thuật toán:**
  + Đầu tiên ta sắp xếp tập FI tăng dần.
  + Bắt đầu duyệt tập phổ biến Y thuộc FI.
  + Duyệt tập con X thuộc FI với điều kiện |X|=|Y|-1.
  + Nếu X là tập con của Y thỏa mãn Conf(X->Y\X) >= minConf thì thêm vào tập thu gọn.

- **Ưu diểm (Tính hiệu quả):**
  + Tốc độ nhanh hơn.
  + Số lượng luật dư thừa ít hơn (không chứa luật bắt cầu).

2. Cho CSDL sau và minsupp=50%, minconf=100%

TID	Items_bought

100	I, B, F, D, E, C, H, J
200	F, C, F, G, A, D, C
300	B, J, D, A, H
400	E, A, B, E, G

a) Sử dụng thuật toán Apriori để tìm tất cả các tập phổ biến. Sử dụng thuật toán FpGrowth để tìm tất cả các tập phổ biến. So sánh kết quả. Liệt kê tập phổ biến tối đại, tập phổ biến đóng.

minsupp=50% → sup.count >= 2

Thuật toán Apriori:

- Tập C1: {{A}, {B}, {C}, {D}, {E}, {F}, {G}, {H}, {I}, {J}}

Tập Item L1	Số lần xuất hiện
A	3
B	3
C	2
D	3
E	2
F	2
G	2
H	2
J	2

- Tập C2: {{A, B}, {A, C}, {A, D}, {A, E}, {A, F}, {A, G}, {A, H}, {A, J}, {B, C}, {B, D}, {B, E}, {B, F}, {B, G}, {B, H}, {B, J}, {C, D}, {C, E}, {C, F}, {C, G}, {C, H}, {C, J}, {D, E}, {D, F}, {D, G}, {D, H},

{D, J}, {E, F}, {E, G}, {E, H}, {E, J}, {F, G}, {F, H}, {F, J}, {G, H}, {G, J}, {H, J}}

Tập Item L2	Số lần xuất hiện
A, B	2
A, D	2
A, G	2
B, D	2
B, E	2
B, H	2
B, J	2
C, D	2
C, F	2
D, F	2
D, H	2
D, J	2
H, J	2

- Tập C3: {{A, B, D}, {A, B, G}, {A, D, G}, {B, D, E}, {B, D, H}, {B, D, J}, {C, D, F}, {D, F, H}, {D, F, J}, {D, H, J}, {B, E, H}, {B, E, J}, {B, H , J}}

Tập Item L3	Số lần xuất hiện
B, D, H	2
B, D, J	2
C, D, F	2
D, H, J	2
B, H, J	2

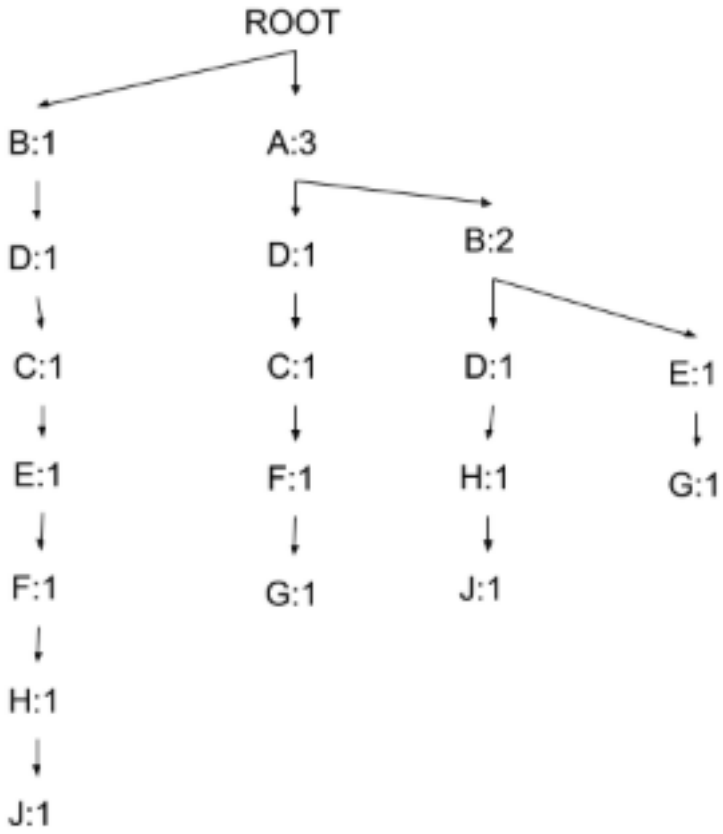
- Tập C4: {B, D, H, J}

Tập Item L4	Số lần xuất hiện
B, D, H, J	2

- Tập C5 rỗng.

Thuật toán FP-Growth:

- Danh sách các tập phổ biến L là: <(A:3), (B:3), (D:3), (C:2), (E:2), (F:2), (G:2), (H:2), (J:2)>
- Cây FP:



- Bảng kết quả của tất cả các item:

Item	Cơ sở điều kiện	Cây điều kiện FP
J	{(B:1, D:1, C:1, E:1, F:1, H:1), (A:1, B:1, D:1, H:1)}	{(B:2, D:2, H:2)}|J
H	{(B:1, D:1, C:1, E:1, F:1), (A:1, B:1, D:1)}	{(B:2, D:2)}|H
G	{(A:1, D:1, C:1, F:1), (A:1, B:1, E:1)}	{(A:2)}|G
F	{(B:1, D:1, C:1, E:1), (A:1, D:1, C:1)}	{(D:2, C:2)}|F
E	{(B:1, D:1, C:1), (A:1, B:1)}	{(B:2)}|E
C	{(B:1, D:1), (A:1, D:1)}	{(D:2)}|C
D	{(B:1), (A:1), (A:1, B:1)}	{(A:2),(B:2)}|D
B	{(A:2)}	{(A:2)}|B
A	{}	{}

- Ta thu được các tập phổ biến: A, B, C, D, E, F, G, H, J, AB, AD, BD, CD, BE, DF, CF, AG, BH, DH, JB, JD, JH, DCF, BDH, BDJ, BHJ, DHJ, BDHJ.
- So sánh kết quả: 2 thuật toán cho kết quả giống nhau.
- Các tập phổ biến tối đại: AB, AD, AG, BE, CDF, BDHJ.
- Các tập phổ biến đóng: A, B, D, AB, AD, AG, BE, CDF, BDHJ.

b) Tìm tất cả LKH có dạng (item 1^ item 2 -> item 3) thỏa mãn ngưỡng minsupp và minconf đã cho.

Các luật kết hợp thỏa minsupp và minconf là: {(B,D->H); (B,H->D), (D,H->B), (B,D->J), (B,J->D), (D,J->B), (C,D->F), (C,F->D), (D,F->C), (D,H->J), (D,J->H), (H,J->D), (B,H->J), (B,J->H), (J,H->B)}

c) Ứng dụng cải tiến của câu 1 vào việc tìm các luật kết hợp ở câu b thỏa mãn ngưỡng minconf. So sánh hiệu quả về thời gian thực hiện với kết quả ở câu b).

Ở câu 2b, các luật sinh ra có dạng item1^item2 -> item3. Ta xét có luật có 3 item, ở thuật toán rút gọn không sinh ra các luật bắc cầu, các luật có 1 item. Ví dụ: B->D,H. Vì vậy sẽ không tốn thêm chi phí xét conf cho các luật này nên thời gian chạy nhanh hơn.

B. Thực hành

1. Hãy chuyển dữ liệu trong tập tin plants.data từ dạng giao dịch sang dạng nhị phân:

- Mỗi dòng là một loài cây.

- Cột đầu tiên là tên loài cây, các cột tiếp theo là các vùng phân bố.

- Giá trị nhị phân gồm y và n. y đại diện cho sự xuất hiện của cây trong vùng phân bố và n là không xuất hiện.

- File theo yêu cầu được lưu với tên plants.csv.

2. Trả lời các câu hỏi sau:

a. Có tất cả bao nhiêu loài cây.

Có 34781 loài cây

b. Có tất cả bao nhiêu vùng phân bố.

Ở tập tin stateabbr.txt có 69 vùng phân bố. Ở tập tin plants.data lại có 70 vùng phân bố (có thêm thành phố gl). Do dữ liệu không thống nhất và thành phố gl chỉ là tên

viết tắt nên nhóm không lấy thành phố này và thống nhất có 69 vùng phân bố.

c. Số loài cây trên mỗi vùng phân bố

Vùng	Số loài cây
ab Alabama	3408
ak Alaska	2969
ar Arkansas	4610
az Arizona	6778
ca California	11676
co Colorado	5465
ct Connecticut	4391
de Delaware	3630
dc District of Columbia	3080
fl Florida	6621
ga Georgia	5942
hi Hawaii	3804
id Idaho	5129
il Illinois	5167
in Indiana	4440
ia Iowa	3652
ks Kansas	3869
ky Kentucky	4555
la Louisiana	5154
me Maine	3969
md Maryland	5108
ma Massachusetts	4963
mi Michigan	4734

mn Minnesota	4929
ms Mississippi	4815
mo Missouri	4638
mt Montana	4800
ne Nebraska	3281
nv Nevada	5670
nh New Hampshire	3635
nj New Jersey	4822
nm New Mexico	6403
ny New York	5773
nc North Carolina	5926
nd North Dakota	2682
oh Ohio	4772
ok Oklahoma	4651
or Oregon	7028
pa Pennsylvania	181
pr Puerto Rico	4781
ri Rhode Island	3295
sc South Carolina	5432
sd South Dakota	3185
tn Tennessee	4900
tx Texas	8483
ut Utah	6041
vt Vermont	3713
va Virginia	5638
vi Virgin Islands	2185
wa Washington	5654
wv West Virginia	4062

wi Wisconsin	4321
wy Wyoming	471
al Alberta	5702
bc British Columbia	4875
mb Manitoba	3023
nb New Brunswick	2856
lb Labrador	1433
nf Newfoundland	2188
nt Northwest Territories	2024
ns Nova Scotia	2844
nu Nunavut	979
on Ontario	5068
Prince Edward Island	5515
qc Québec	4272
sk Saskatchewan	2846
yt Yukon	2100
dengl Greenland (Denmark)	479
fraspm St. Pierre and Miquelon (France)	1210

d. Vùng phân bố có ít loài cây nhất, cho biết số lượng và tỉ lệ %.

- Vùng phân bố ít loài cây nhất: pa Pennsylvania (181 loài)
- Tỉ lệ %: 0.52%

e. Vùng phân bố có nhiều loài cây nhất, cho biết số lượng và tỉ lệ %.

- **V**ùng phân bố nhiều loài cây nhất: ca California (11676 loài)
- Tỉ lệ %: 33.57%

f. Trung bình một vùng phân bố có bao nhiêu loài cây.

Trung bình một vùng phân bố có 4346 loài cây.

3. Chúng ta chuẩn bị áp dụng giải thuật Apriori trên dữ liệu này, giả sử khi khai thác tập phổ biến và luật kết hợp ta chỉ quan tâm đến các vùng mà một loài cây có xuất hiện ở đó (các giá trị 'y') → dữ liệu được xem như dữ liệu giao dịch. Giải thuật Apriori trong Weka khi thực hiện sẽ bỏ qua các giá trị thiếu, chỉ cần loại các giá trị 'n' ra khỏi dữ liệu.

- Hãy thay thế toàn bộ giá trị 'n' thành '?'.

- Thuộc tính đầu tiên (tên loài cây) không cần thiết trong bài toán khai thác tập phổ biến, hãy xóa nó đi.

- Dữ liệu thỏa các yêu cầu trên được lưu với tên với tên plants.arff

4. Khai thác tập phổ biến: Sử dụng thuật toán Apriori trong Weka để khai thác tất cả tập hạng mục có độ phổ biến từ 0.1 trở lên.

- Bảng kết quả định lượng:

Kích thước	Số lượng
1 hạng mục	49
2 hạng mục	167
3 hạng mục	116

4 hạng mục	25
5 hạng mục	2

- Danh sách tất cả các tập phổ biến thỏa yêu cầu được lưu trong tập tin FI.doc

5. Khai thác luật kết hợp:

- Với mỗi tập phổ biến có kích thước lớn nhất theo kết quả của câu 4:

o Sử dụng thuật toán FP-Growth trong Weka để khai thác tất cả các luật kết hợp.

o Có độ tin cậy (Confidence) từ 0.95 trở lên.

o Chứa tất cả các hạng mục thuộc tập phổ biến đang xét.

- Bảng kết quả định lượng:

Tập hạng mục phổ biến	Số lượng luật
ga=y, al=y, va=y, sc=y	3579
nc=y, al=y, sc=y, ms=y	3572
nc=y, al=y, va=y, sc=y	3608
ga=y, nc=y, sc=y, ms=y	3612
ga=y, nc=y, al=y, ms=y	3635
al=y, va=y, sc=y	3682
ga=y, nc=y, al=y, va=y	3694
nc=y, sc=y, ms=y	3698

- Danh sách tất cả các luật kết hợp thỏa yêu cầu được lưu trong tập tin AR.docx

THAM KHẢO:

[1] Slide lý thuyết bài 3

[2] Textbook: J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition - Chapter 5: Mining Frequent Patterns, Associations and Correlations (5.2.1 đến 5.2.4)

[3] J. Han, H. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. In: Proc. Conf. on the Management of Data (SIGMOD'00, Dallas, TX). ACM Press, New York, NY, USA 2000

[4] Ong Xuan Hong, Apriori và Fp-Growth với tập dữ liệu plants, web: https://ongxuanhong.wordpress.com/2015/08/24/apriori-va-fp-growth-voi-tap-du-lieu-plants/?fbclid=IwAR3c3GGYRy837Tdtn86NksSAI5LgDK2Ob4Qr2K1H2DxzBeAU2FpnnZwmFBY