Data Mining Nguyen Manh Tuan

khai thác dữ liệu và ứng dụng (Trường Đại học Công nghiệp Thành phố Hồ Chí Minh)

Người gởi : Nguyễn Mạnh Tuấn
MSSV: CH0601093
Ngày gởi : 30/06/2007

DATA MINING

Bài tập 1:

1. Cho tập các hoá đơn O={o1, o2, o3, o4, o5}, mỗi hóa đơn chứa các mặt hàng như sau:
   o1={i1,i3,i4} ; o2={i1,i3,i4} ; o3={i3,i5} ; o4={i4 , i5} ; o5={i2,i3,i5}
   Cho ngưỡng phổ biến tối thiểu minsupp=0,4 hãy:
   a. Tìm các tập phổ biến tối đại theo ngưỡng minsupp=0,4
   b. Tìm tất cả các luật kết hợp có độ phổ biến tối thiểu là 0,4 và độ tin cậy tối thiểu là 0,8
2. Sử dụng cây định danh để tìm các luật phân lớp từ bảng quyết định sau đây:

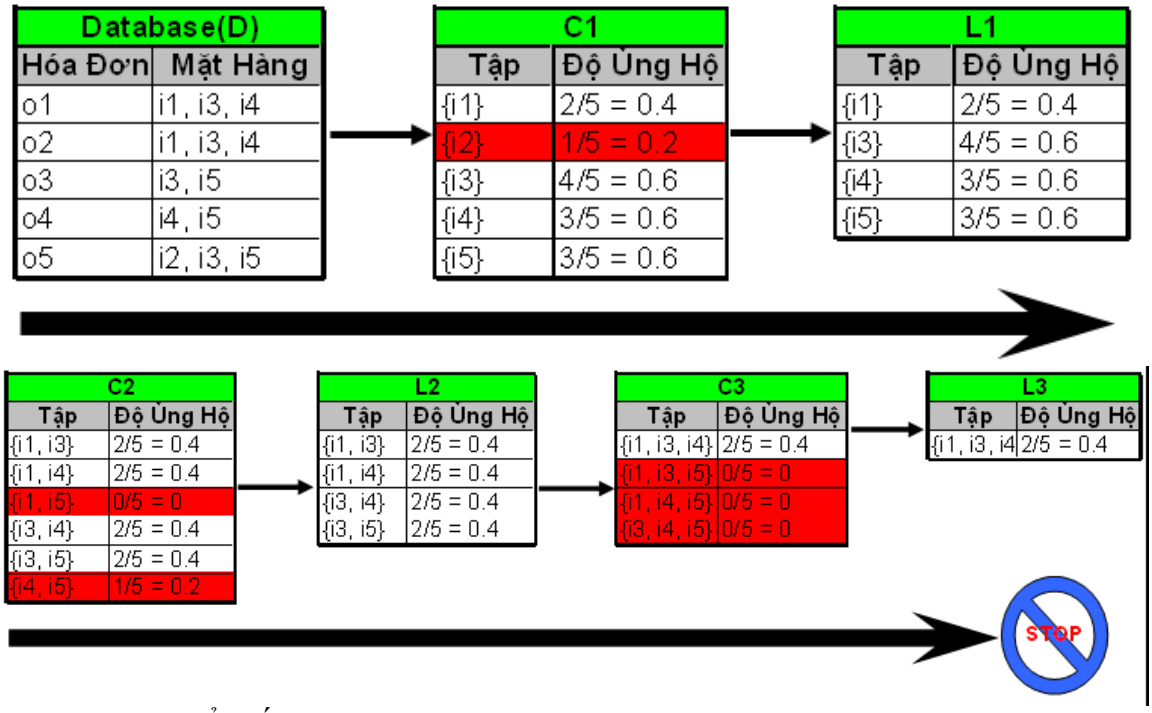
#	**Trời**	**Áp Suất**	**Gió**	**Kết quả**
1	Trong	Cao	Bắc	Không mưa
2	Mây	Cao	Nam	Mưa
3	Mây	Trung bình	Bắc	Mưa
4	Trong	Thấp	Bắc	Không mưa
5	Mây	Thấp	Bắc	Mưa
6	Mây	Cao	Bắc	Mưa
7	Mây	Thấp	Nam	Không mưa
8	Trong	Cao	Nam	Không mưa

Bạn có suy nghĩ gì về việc dùng luật kết hợp để làm luật phân lớp.
Bảng dữ liệu lúc đó sẽ có các cột <Trời, Trong>, <Trời, mây>, < Ápsuất, Cao> < Ápsuất,trungbình>, <Ápsuất, Thấp>

Giải bài tập 1:

1. Luật kết hợp

a.Tìm các tập phổ biến tối đại theo ngưỡng minsupp=0.4



Database(D)	
Hóa Đơn	Mặt Hàng
o1	i1, i3, i4
o2	i1, i3, i4
o3	i3, i5
o4	i4, i5
o5	i2, i3, i5

C1	
Tập	Độ Ủng Hộ
{i1}	2/5 = 0.4
{i2}	1/5 = 0.2
{i3}	4/5 = 0.6
{i4}	3/5 = 0.6
{i5}	3/5 = 0.6

L1	
Tập	Độ Ủng Hộ
{i1}	2/5 = 0.4
{i3}	4/5 = 0.6
{i4}	3/5 = 0.6
{i5}	3/5 = 0.6

C2	
Tập	Độ Ủng Hộ
{i1, i3}	2/5 = 0.4
{i1, i4}	2/5 = 0.4
{i1, i5}	0/5 = 0
{i3, i4}	2/5 = 0.4
{i3, i5}	2/5 = 0.4
{i4, i5}	1/5 = 0.2

L2	
Tập	Độ Ủng Hộ
{i1, i3}	2/5 = 0.4
{i1, i4}	2/5 = 0.4
{i3, i4}	2/5 = 0.4
{i3, i5}	2/5 = 0.4

C3	
Tập	Độ Ủng Hộ
{i1, i3, i4}	2/5 = 0.4
{i1, i3, i5}	0/5 = 0
{i1, i4, i5}	0/5 = 0
{i3, i4, i5}	0/5 = 0

L3	
Tập	Độ Ủng Hộ
{i1, i3, i4	2/5 = 0.4

Vậy các tập phổ biến thu được là :

L1 = **{i1}, {i3}, {i4}, {i5}}**

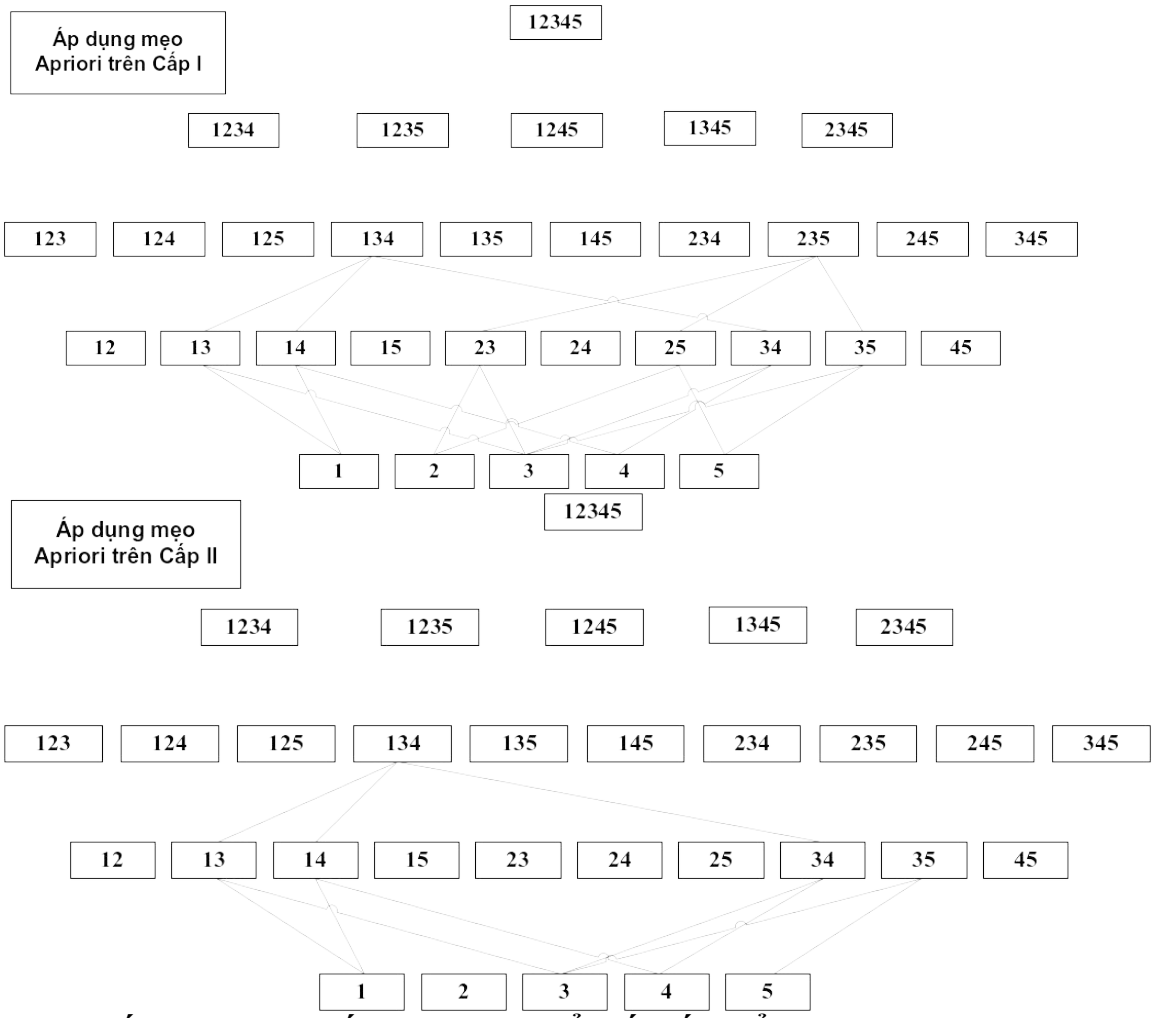
L2 = {{i1, i3}, {i1, i4}, {i3, i4}, {i3, i5}}

L3 = {{i1, i3, i4}}

Vậy tập phổ biến **tối đại** là: **{{i3, i5}, {i1, i3, i4}}**

Chú ý: chúng ta có thể áp dụng heuristic tại bước L2 -> C3, vì tạo ra C3 phải có 3 phần tử nên ta chỉ cần quan tâm đến 2 tập 3 phần tử xuất tập D: {i1, i3, i4} và {i2, i3, i5}, cho nên tại bước này ta chỉ cần chọn tập {i1, i3, i4} để tìm tập phổ biến.

Hay chúng ta dùng mẹo Apriori:



b. Tìm tất cả các luật kết hợp có độ phổ biến tối thiểu là 0,4 và độ tin cậy tối thiểu là 0,8

Tập Phổ Biến (S)	Luật	Độ Tin Cậy
{i1, i3}	I1 => i3	Supp(S)/Supp(i1) = (2/5)/(2/5) = 1
	I3 => i1	Supp(S)/Supp(i3) = (2/5)/(4/5) = 0.5
{i1, i4}	I1 => i4	Supp(S)/Supp(i1) = (2/5)/(2/5) = 1
	I4 => i1	Supp(S)/Supp(i4) = (2/5)/(3/5) = 0.67
{i3, i4}	I3 => i4	Supp(S)/Supp(i3) = (2/5)/(4/5) = 0.5
	I4 => i3	Supp(S)/Supp(i4) = (2/5)/(3/5) = 0.67
{i3, i5}	I3 => i5	Supp(S)/Supp(i3) = (2/5)/(4/5) = 0.5
	I5 => i3	Supp(S)/Supp(i5) = (2/5)/(3/5) = 0.67
{i1, i3, i4}	I1, i3 => i4	Supp(S)/Supp(i1, i3) = (2/5)/(2/5) = 1
	I1, i4 => i3	Supp(S)/Supp(i1, i4) = (2/5)/(2/5) = 1
	I3, i4 => i1	Supp(S)/Supp(i3, i4) = (2/5)/(2/5) = 1
	I1 => i3, i4	Supp(S)/Supp(i1) = (2/5)/(2/5) = 1

	I3 => i1, i4	Supp(S)/Supp(i3) = (2/5)/(4/5) = 0.5
	I4 => i1, i3	Supp(S)/Supp(i4) = (2/5)/(3/5) = 0.67

Tất cả các luật kết hợp có độ phổ biến tối thiểu là 0,4 và độ tin cậy tối thiểu là 0,8:

- Luật 1 : i1 => i3
- Luật 2 : i1 => i4
- Luật 3: i1,i3 => i4
- Luật 4: i1,i4 => i3
- Luật 5: i3,i4 => i1
- Luật 6: i1 => i3, i4

2. Luật Phân Lớp

#	Trời	Áp Suất	Gió	Kết quả
1	Trong	Cao	Bắc	Không mưa
2	Mây	Cao	Nam	Mưa
3	Mây	Trung bình	Bắc	Mưa
4	Trong	Thấp	Bắc	Không mưa
5	Mây	Thấp	Bắc	Mưa
6	Mây	Cao	Bắc	Mưa
7	Mây	Thấp	Nam	Không mưa
8	Trong	Cao	Nam	Không mưa

Dùng thuật tóan ID3 để phân hoạch:

Ký hiệu: P: Mưa; N: Không mưa

$I(P, N) = -P/(P+N)*LOG_2(P/(P+N)) - N/(P+N)*LOG_2(/(P+N))$

Ta có: P = 4 và N = 4

=> I(4, 4) = 1

- Tính entropy cho thuộc tính [**Trời**]:

Trời	pi	ni	I(pi, ni)
Trong	0	3	0
Mây	4	1	0.72
E (Trời) = 3/8*0 + 5/8*0.72 = 0.45			
Gain(Trời) = I(4,4) - E (Trời) = 0.55			

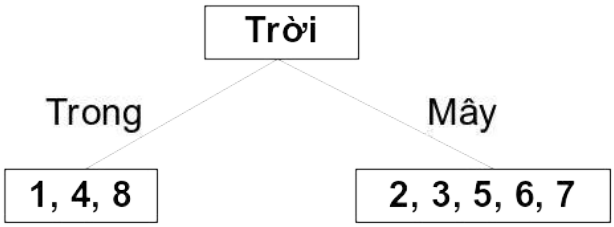
- Tính entropy cho thuộc tính [**Áp Suất**]:

Áp Suất	pi	ni	I(pi, ni)
Cao	2	2	1.00
Trung Bình	1	0	0.00
Thấp	1	2	0.92
E (Áp Suất) = 4/8*1 + 1/8*0 + 3/8*0.92 = 0.84			
Gain(Áp Suất) = I(4,4) - E (Áp Suất) = 0.155			

- Tính entropy cho thuộc tính [**Gió**]:

Gió	pi	ni	I(pi, ni)
Bắc	3	2	0.97
Nam	1	2	0.92
E (Gió) = 4/8*0.97 + 3/8*0.92 = 0.95			
Gain(Trời) = I(4,4) - E (Gió) = 0.05			

Ta nhận thấy Gain của thuộc tính [Trời] là lớn nhất, nên ta dùng thuộc tính [Trời] để phân lớp:



- Phân lớp nhánh [Trời - Trong]:
  - Bảng dữ liệu của nhánh:

#	Áp Suất	Gió	Kết quả
1	Cao	Bắc	Không mưa
4	Thấp	Bắc	Không mưa
8	Cao	Nam	Không mưa

  Với kết quả này ta không cần phân lớp nữa cho lớp [Trời - Trong].
- Phân lớp nhánh [Trời - Mây]:
  - Bảng dữ liệu của nhánh:

#	Áp Suất	Gió	Kết quả
2	Cao	Nam	Mưa
3	Trung bình	Bắc	Mưa
5	Thấp	Bắc	Mưa
6	Cao	Bắc	Mưa
7	Thấp	Nam	Không mưa

  I(4,1) = 0.72
  - Tính entropy cho thuộc tính [Áp Suất]:

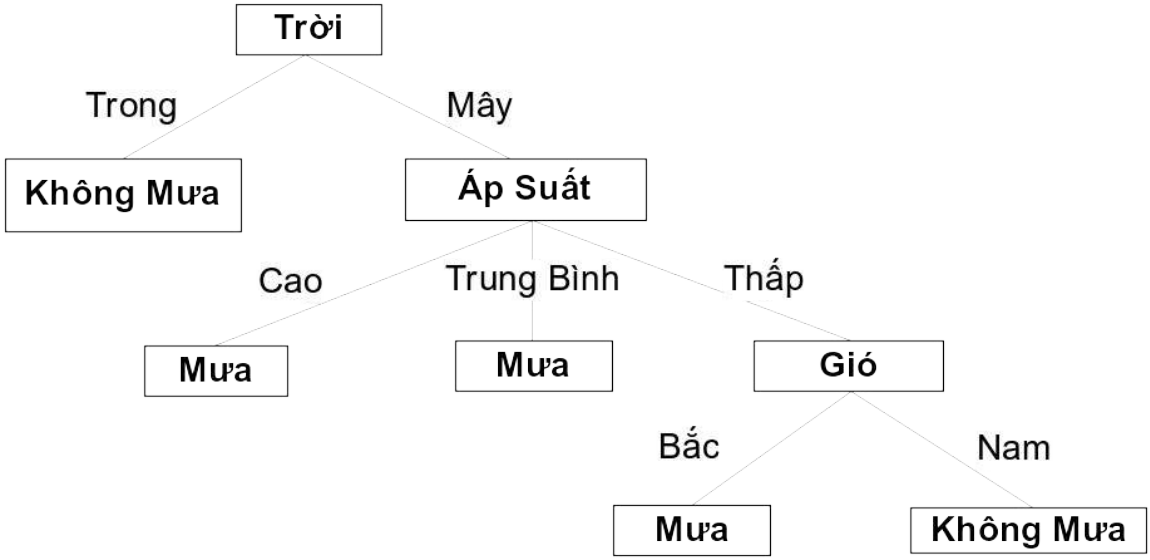
Áp Suất	Pi	ni	I(pi, ni)
Cao	2	0	0.00
Trung Bình	1	0	0.00
Thấp	1	1	1.0
E (Áp Suất) = 2/5*0 + 1/5*0 + 2/5*1 = 0.4			
Gain(Áp Suất) = I(4,1) - E (Áp Suất) = 0.32			

  - Tính entropy cho thuộc tính [Gió]:

Gió	pi	Ni	I(pi, ni)
Bắc	3	0	0.00
Nam	1	1	1.0

E (Gió) = 3/5*0 + 2/5*1 = 0.4
Gain(Trời) = I(4,1) – E (Gió) = 0.32

Thuộc tính [Gió] và [Áp Suất] có Gain bằng nhau, vì [Áp Suất] có nhiều thuộc tính hơn nên ta chọn [Áp Suất] để phân tích:



Vậy, ta có các luật sau:

- Trời trong => Không mưa
- Trời mây, Áp Suất Cao => Mưa
- Trời mây, Áp Suất Trung Bình => Mưa
- Trời mây, Áp Suất Thấp, Gió Bắc => Mưa
- Trời mây, Áp Suất Thấp, Gió Nam => Không mưa

Bài Tập 2: Luật Phân Lớp

Dùng thuật tóan ID3 và Naïve Bayes để tìm luật phân lớp trong bảng sau đây.

TT	Màu Tóc	Chiều Cao	Cân Nặng	Dùng Thuốc?	Kết Quả
1	Đen	Tầm thước	Nhẹ	Không	Bị rám
2	Đen	Cao	Vừa phải	Có	Không
3	Râm	Thấp	Vừa phải	Có	Không
4	Đen	Thấp	Vừa phải	Không	Bị rám
5	Bạc	Tầm thước	Nặng	Không	Bị rám
6	Râm	Cao	Nặng	Không	Không
7	Râm	Tầm thước	Nặng	Không	Không
8	Đen	Thấp	Nhẹ	Có	Không

So sánh kết qủa.

Bài giải:

- Sử dụng thuật toán ID3:

Ký hiệu: P: Bị Rám; N: Không

$I(P, N) = -P/(P+N)*LOG_2(P/(P+N)) - N/(P+N)*LOG_2(/(P+N))$

Ta có: P = 3 và N = 5

=> I(3, 5) = 0.95

- Tính entropy cho thuộc tính [**Màu Tóc**]:

Màu Tóc	pi	ni	I(pi, ni)
Đen	2	2	1.00
Râm	0	3	0.00
Bạc	1	0	0.00
E (Màu Tóc) = 4/8*1 + 3/8*0 + 1/8*0 = 0.5			
Gain(Màu Tóc) = I(3,5) - E (Màu Tóc) = 0.45			

- Tính entropy cho thuộc tính [**Chiều Cao**]:

Chiều Cao	Pi	ni	I(pi, ni)
Cao	0	2	0.00
Tầm thước	2	1	0.92
Thấp	1	2	0.92
E (Chiều Cao) = 2/8*0 + 3/8*0.92 + 3/8*0.92 = 0.69			
Gain(Chiều Cao) = I(3,5) - E (Chiều Cao) = 0.26			

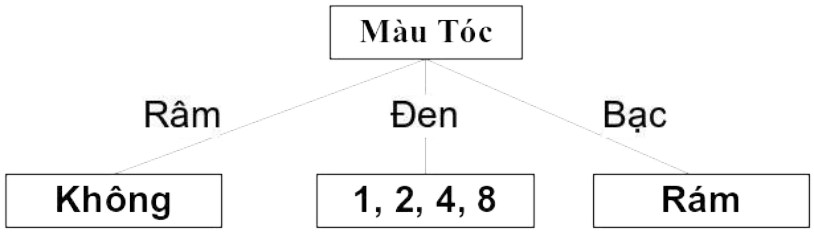
- Tính entropy cho thuộc tính [**Cân Nặng**]:

Cân Nặng	pi	ni	I(pi, ni)
Nặng	1	2	0.92
Vừa phải	1	2	0.92
Nhẹ	1	1	1.00
E (Cân Nặng) = 3/8*0.92 + 3/8*0.92 + 2/8*1 = 0.94			
Gain(Cân Nặng) = I(3,5) - E (Cân Nặng) = 0.01			

- Tính entropy cho thuộc tính [**Dùng Thuốc**]:

Dùng Thuốc	Pi	ni	I(pi, ni)
Không	3	2	0.97
Có	0	3	0.00
E (Dùng Thuốc) = 5/8*0.97 + 3/8*0 = 0.6			
Gain(Dùng Thuốc) = I(3,5) - E (Dùng Thuốc) = 0.35			

Ta nhận thấy Gain của thuộc tính [**Màu Tóc**] là lớn nhất, nên ta dùng thuộc tính [**Màu Tóc**] để phân lớp:



- Phân lớp nhánh [Màu Tóc - Đen]:
  - Bảng dữ liệu của nhánh:

TT	Chiều Cao	Cân Nặng	Dùng Thuốc?	Kết Quả
1	Tầm thước	Nhẹ	Không	Bị rám
2	Cao	Vừa phải	Có	Không
4	Thấp	Vừa phải	Không	Bị rám
8	Thấp	Nhẹ	Có	Không

I(2,2) = 1

  - Tính entropy cho thuộc tính [**Chiều Cao**]:

Chiều Cao	**Pi**	**Ni**	**I(pi, ni)**
Cao	0	1	0.00
Tầm thước	1	0	0.00
Thấp	1	1	1.00
E (Chiều Cao) = ¼*0 + 1/4*0 + 2/4*1.00 = 0.5			
Gain(Chiều Cao) = I(2,2) - E (Chiều Cao) = 0.5			

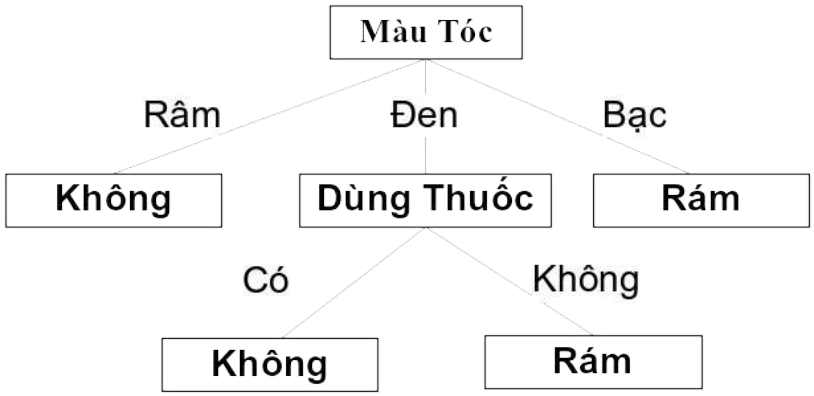
  - Tính entropy cho thuộc tính [**Cân Nặng**]:

Cân Nặng	**pi**	**Ni**	**I(pi, ni)**
Nặng	0	0	0.00
Vừa phải	1	1	1.00
Nhẹ	1	1	1.00
E (Cân Nặng) = 0/4*0 + 2/4*1 + 2/4*1 = 1			
Gain(Cân Nặng) = I(2,2) - E (Cân Nặng) = 0			

  - Tính entropy cho thuộc tính [**Dùng Thuốc**]:

Dùng Thuốc	**Pi**	**Ni**	**I(pi, ni)**
Không	2	0	0.00
Có	0	2	0.00
E (Dùng Thuốc) = 2/4*0 + 2/4*0 = 0			
Gain(Dùng Thuốc) = I(2,2) - E (Dùng Thuốc) = 1			

Ta nhận thấy Gain của thuộc tính [**Dùng Thuốc**] là lớn nhất, nên ta dùng thuộc tính [**Dùng Thuốc**] để phân lớp:



Vậy, ta có các luật sau:

- Màu Tóc râm => Không Rám
- Màu Tóc bạc => Rám
- Màu Tóc đen, dùng thuốc => Không Rám
- Màu Tóc đen, không dùng thuốc => Rám

- **Sử dụng thuật tóan Naive Bayes**
  - P(P) = 3/8 = 0.375
  - P(N) = 5/8 = 0.625

Màu Tóc	
P(Đen/P) = 2/3	P(Đen/N) = 2/5
P(Râm/P) = 0/3	P(Râm/N) = 3/5
P(Bạc/P) = 1/3	P(Bạc/N) = 0/5

Chiều Cao	
P(Cao/P) = 0/3	P(Cao/N) = 2/5
P(Tầm Thước/P) = 2/3	P(Tầm Thước/N) = 1/5
P(Thấp/P) = 1/3	P(Thấp/N) = 2/5

Cân Nặng	
P(Nặng/P) = 1/3	P(Nặng/N) = 2/5
P(Vừa Phải/P) = 1/3	P(Vừa Phải/N) = 2/5
P(Nhẹ/P) = 1/3	P(Nhẹ/N) = 1/5

Dùng Thuốc	
P(Không/P) = 3/3	P(Không/N) = 2/5
P(Có/P) = 0/3	P(Có/N) = 3/5

  - Phân lớp X: Xét các mẫu chưa tìm thấy

- X1 = {Đen, Tầm thước, Nhẹ, Có}
- P(X1,P).P(P) = P(Đen,P)*P(Tầm Thước,P)*P(Nhẹ, P)*P(Có,P)*P(P) = 2/3*2/3*1/3*0/3*0.375 = 0
- P(X1,N).P(N) = P(Đen,N)*P(Tầm Thước,N)*P(Nhẹ, N)*P(Có,N)*P(N) = 2/5*1/5*1/5*3/5*0.625 = 0.006
  Mẫu X2 được phân vào lớp N

- X2 = {Đen, Tầm thước, Vừa Phải, Không}
- P(X2,P).P(P) = P(Đen,P)*P(Tầm Thước,P)*P(Vừa Phải, P)*P(Không,P)*P(P) = 2/3*2/3*1/3*3/3*0.375 = 0.05555
- P(X2,N).P(N) = P(Đen,N)*P(Tầm Thước,N)*P(Vừa Phải, N)*P(Không,N)*P(N) = 2/5*1/5*2/5*2/5*0.625 = 0.008
  Mẫu X2 được phân vào lớp P

- X3 = {Đen, Tầm thước, Vừa Phải, Có}
- P(X3,P).P(P) = P(Đen,P)*P(Tầm Thước,P)*P(Vừa Phải, P)*P(Có,P)*P(P) = 2/3*2/3*1/3*0/3*0.375 = 0
- P(X3,N).P(N) = P(Đen,N)*P(Tầm Thước,N)*P(Vừa Phải, N)*P(Có,N)*P(N) = 2/5*1/5*2/5*3/5*0.625 = 0.012
  Mẫu X3 được phân vào lớp N

- X4 = {Đen, Tầm thước, Nặng, Không}
- P(X4,P).P(P) = P(Đen,P)*P(Tầm Thước,P)*P(Nặng, P)*P(Không,P)*P(P) = 2/3*2/3*1/3*3/3*0.375 = 0.055
- P(X4,N).P(N) = P(Đen,N)*P(Tầm Thước,N)*P(Nặng, N)*P(Không,N)*P(N) = 2/5*1/5*2/5*2/5*0.625 = 0.008
  Mẫu X4 được phân vào lớp P

- X5 = {Đen, Tầm thước, Nặng, Có} phân vào lớp N vì P(Có,P) = 0
- X6 = {Đen, Cao,...,...} phân vào lớp N vì P(Cao/P) = 0
- X7 = {Đen, Thấp, Nặng, Không}
- P(X7,P).P(P) = P(Đen,P)*P(Thấp,P)*P(Nặng, P)*P(Không,P)*P(P) = 2/3*1/3*1/3*3/3*0.375 = 0.028
- P(X7,N).P(N) = P(Đen,N)*P(Thấp,N)*P(Nặng, N)*P(Không,N)*P(N) = 2/5*2/5*2/5*2/5*0.625 = 0.016
  Mẫu X7 được phân vào lớp P
- X8 = {Đen, Thấp, Nặng, Có} phân vào lớp N vì P(Có,P) = 0
- X9 = {Đen, Thấp, Nhẹ, Không}

- P(X10,P).P(P) = P(Đen,P)*P(Thấp,P)*P(Nhẹ, P)*P(Không,P)*P(P) = 2/3*1/3*1/3*3/3*0.375 = 0.028
- P(X10,N).P(N) = P(Đen,N)*P(Thấp,N)*P(Nhẹ, N)*P(Không,N)*P(N) = 2/5*2/5*1/5*2/5*0.625 = 0.008 Mẫu X10 được phân vào lớp P
- X11 = {Đen, Thấp, Nhẹ, Có} phân vào lớp N vì P(Có,P) = 0
- X12 = {Râm,...,...,...} phân vào lớp N vì P(Râm,P) = 0
- X12 = {Bạc,...,...,...} phân vào lớp P vì P(Bạc,N) = 0

- Rút ra các phân lớp:
  - Màu tóc Râm => Không Rám
  - Màu tóc Bạc => Rám
  - Chiều cao Cao => Không Rám
  - Có dùng Thuốc => Không Rám
  - Màu tóc Đen, Tầm Thước, Vừa Phải, Không => Rám
  - Màu tóc Đen, Tầm Thước, Nhẹ, Không => Rám
  - Màu tóc Đen, Tầm Thước, Nặng, Không => Rám
  - Màu tóc Đen, Thấp, Vừa Phải, Không => Rám
  - Màu tóc Đen, Thấp, Nhẹ, Không => Rám
  - Màu tóc Đen, Thấp, Nặng, Không => Rám
- Rút gọn các phân lớp:
  - Màu tóc Râm => Không Rám
  - Màu tóc Bạc => Rám
  - Chiều cao Cao => Không Rám
  - Có dùng Thuốc => Không Rám
  - Màu tóc Đen, Tầm Thước, Không => Rám
  - Màu tóc Đen, Thấp, Không => Rám

Kết Luận: dùng phương pháp Naive Bayes phức tạp hơn phương pháp ID3, tuy nhiên có khai phá ra được nhiều luật mới hơn ID3.

Bài Tập 3: Tập Thô
Hãy rút gọn bảng quyết định sau đây:

Bảng : Số liệu quan sát về hiện tượng rám nắng.

TT	Tên người	Màu tóc	Chiều cao	Cân nặng	Dùng thuốc?	Kết quả
		M	C	N	T	R
1	Hoa	Đen	Tầm thước	Nhẹ	Không	Bị rám
2	Lan	Đen	Cao	Vừa phải	Có	Không
3	Xuân	Râm	Thấp	Vừa phải	Có	Không
4	Hạ	Đen	Thấp	Vừa phải	Không	Bị rám
5	Thu	Bạc	Tầm thước	Nặng	Không	Bị rám

6	Đông	Râm	Cao	Nặng	Không	Không
7	Mơ	Râm	Tầm thước	Nặng	Không	Không
8	Đào	Đen	Thấp	Nhẹ	Có	Không

Khi dùng cây quyết định, ta có các luật:

- IF Tóc đen
  Người đó dùng thuốc
  THEN Không sao cả
- IF Người tóc đen
  Không dùng thuốc
  THEN Họ bị rám nắng
- IF Người tóc bạc
  THEN Bị rám nắng
- IF Người tóc râm
  THEN Không sao cả

Sau khi rút gọn (để có các reducts) , bạn có luật gì ????
So sánh với kết quả tạo luật từ cây quyết định

Giải:

Ta có ma trận phân biệt:

	Hoa	**Lan**	**Xuân**	**Hạ**	**Thu**	**Đông**	**Mơ**
Lan	C, N, T						
Xuân	M,C,N,T						
Hạ	λ	C, T	M,T				
Thu	λ	M,C,N,T	M,C,N,T				
Đông	M, C, N	λ	λ	M,C,N	M,C		
Mơ	M,N	λ	λ	M,C,N	M	λ	
Đào	C,T	λ	λ	N,T	M,C,N,T	λ	λ

=> F(M,C,N,T) = (C$^{\vee}$N$^{\vee}$T) $\wedge$ (M$^{\vee}$C$^{\vee}$N$^{\vee}$T) $\wedge$ (M$^{\vee}$N$^{\vee}$C) $\wedge$ (M$^{\vee}$N) $\wedge$ (C$^{\vee}$T) $\wedge$ (M$^{\vee}$T) $\wedge$ (N$^{\vee}$T) $\wedge$ (M$\wedge$C) $\wedge$ (M) = M $\wedge$ (C$^{\vee}$N$^{\vee}$T) $\wedge$ (C$^{\vee}$T) $\wedge$ (N$^{\vee}$T) = (M$\wedge$T) $^{\vee}$ (M$\wedge$C$\wedge$N)

- **Ta có rút gọn sau:**
  Reduct1 = {M, T}
  Reduct2 = {M, C, N}
  Core = {Kết quả} = {M, C, N} ∩ {M, T} = {M}
- Tìm các luật
  - Với C = {Kết quả} ta có các lớp tương đương sau:

- X1 = {1,4,5} Bị rám
- X2 = {2,3,6,7,8} Không bị rám

- Xét Reduct1 = {M, T} với {M, T} => K, ta có các lớp tương đương:

  Z1 = {1, 4} Đen, Không dùng thuốc
  Z2 = {2, 8} Đen, Có dùng thuốc
  Z3 = {3} Râm, Có dùng thuốc
  Z4 = {5} Bạc, Không dùng thuốc
  Z5 = {6, 7} Râm, Không dùng thuốc

  - Xét phân lớp X1 {1,4,5} Bị Rám:
    - X1 ∩ Z1 = {1, 4} # ⊘ và Z1 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Đen, Không dùng thuốc thì Bị rám.
    - X1 ∩ Z2 = ⊘ => không có luật phân lớp.
    - X1 ∩ Z3 = ⊘ => không có luật phân lớp.
    - X1 ∩ Z4 = {5} # ⊘ và Z4 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Bạc, Không dùng thuốc thì Bị rám.
    - X1 ∩ Z5 = ⊘ => không có luật phân lớp.

  Vậy ta có các luật sau :

  L1: Nếu tóc Đen, Không dùng thuốc thì Bị rám.
  L5: Nếu tóc Bạc, Không dùng thuốc thì Bị rám.

  - Xét phân lớp X2 {2,3,6,7,8} Không Bị Rám:
    - X1 ∩ Z1 = ⊘ => không có luật phân lớp.
    - X1 ∩ Z2 = {2, 8} # ⊘ và Z2 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Đen, Có dùng thuốc thì Không Bị rám.
    - X1 ∩ Z3 = {3} # ⊘ và Z3 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Râm, Có dùng thuốc thì Không Bị rám.
    - X1 ∩ Z4 = ⊘ => không có luật phân lớp.
    - X1 ∩ Z5 = {6,7} # ⊘ và Z5 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Râm, Không dùng thuốc thì Không Bị rám.
    - X1 ∩ Z5 = ⊘ => không có luật phân lớp.

  Vậy ta có các luật sau :

  L2: Nếu tóc Đen, Có dùng thuốc thì Không Bị rám.
  L3: Nếu tóc Râm, Có dùng thuốc thì Không Bị rám.
  L6: Nếu tóc Râm, Không dùng thuốc thì Không Bị rám.

  Vậy đối với trường hợp **{M,T} -> {K} .** Ta có các luật sau :

  L1: Nếu tóc Đen, Không dùng thuốc thì Bị rám.
  L5: Nếu tóc Bạc, Không dùng thuốc thì Bị rám.

L2: Nếu tóc Đen, Có dùng thuốc thì Không bị rám.
L3: Nếu tóc Râm, Có dùng thuốc thì Không bị rám.
L6: Nếu tóc Râm, Không dùng thuốc thì Không bị rám.

Rút gọn luật L3 và L6. Ta có các luật còn lại:
L1: Nếu tóc Đen, Không dùng thuốc thì Bị rám.
L5: Nếu tóc Bạc, Không dùng thuốc thì Bị rám.
L2: Nếu tóc Đen, Có dùng thuốc thì Không bị rám.
L3: Nếu tóc Râm thì Không bị rám.

- Xét Reduct2 = {M, C, N} với {M, C, N} => K, ta có các lớp tương đương:
  Z1 = {1} Đen, Tầm thước, Nhẹ
  Z2 = {2} Đen, Cao, Vừa phải
  Z3 = {3} Râm, Thấp, Vừa phải
  Z4 = {4} Đen, Thấp, Vừa phải
  Z5 = {5} Bạc, Tầm thước, Nặng
  Z6 = {6} Râm, Cao, Nặng
  Z7 = {7} Râm, Tầm thước, Nặng
  Z8 = {8} Đen, Thấp, Nhẹ
  - Xét phân lớp X1 {1,4,5} Bị Rám:
    - X1 ∩ Z1 = {1} # ⊘ và Z1 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Đen, Tầm thước, Nhẹ thì Bị rám.
    - X1 ∩ Z2 = ⊘ => không có luật phân lớp.
    - X1 ∩ Z3 = ⊘ => không có luật phân lớp.
    - X1 ∩ Z4 = {4} # ⊘ và Z4 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Đen, Thấp, Vừa phải thì Bị rám.
    - X1 ∩ Z5 = {5} # ⊘ và Z5 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Bạc, Tầm thước, Nặng thì Bị rám.
    - X1 ∩ Z6 = ⊘ => không có luật phân lớp.
    - X1 ∩ Z7 = ⊘ => không có luật phân lớp.
    - X1 ∩ Z8 = ⊘ => không có luật phân lớp.

  Vậy ta có các luật sau :
  L'1: Nếu tóc Đen, Tầm thước, Nhẹ thì Bị rám.
  L'4: Nếu tóc Đen, Thấp, Vừa phải thì Bị rám
  L'5: Nếu tóc Bạc, Tầm thước, Nặng thì Bị rám.
  - Xét phân lớp X2 {2,3,6,7,8} Không Bị Rám:
    - X1 ∩ Z1 = ⊘ => không có luật phân lớp.
    - X1 ∩ Z2 = {2} # ⊘ và Z2 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Đen, Cao, Vừa phải thì Không Bị rám.

- X1 ∩ Z3 = {3} # ⊘ và Z3 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Râm, Thấp, Vừa phải thì Không Bị rám.
- X1 ∩ Z4 = ⊘ => không có luật phân lớp.
- X1 ∩ Z5 = ⊘ => không có luật phân lớp.
- X1 ∩ Z6 = {6} # ⊘ và Z6 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Râm, Cao, Nặng thì Không Bị rám.
- X1 ∩ Z7 = {7} # ⊘ và Z7 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Râm, Tầm thước, Nặng thì Không Bị rám.
- X1 ∩ Z8 = {8} # ⊘ và Z8 ⊆ X1 nên ta có luật phân lớp đúng chính xác 100%. Vậy ta có luật : Nếu tóc Đen, Thấp, Nhẹ thì Không Bị rám.

Vậy ta có các luật sau :

L'2: Nếu tóc Đen, Cao, Vừa phải thì Không bị rám.
L'3: Nếu tóc Râm, Thấp, Vừa phải thì Không bị rám.
L'6: Nếu tóc Râm, Cao, Nặng thì Không bị rám.
L'7: Nếu tóc Râm, Tầm thước, Nặng thì Không bị rám.
L'8: Nếu tóc Đen, Thấp, Nhẹ thì Không bị rám.

Vậy đối với trường hợp **{M,T} -> {K} .** Ta có các luật sau :

L'1: Nếu tóc Đen, Tầm thước, Nhẹ thì Bị rám.
L'4: Nếu tóc Đen, Thấp, Vừa phải thì Bị rám
L'5: Nếu tóc Bạc, Tầm thước, Nặng thì Bị rám.
L'2: Nếu tóc Đen, Cao, Vừa phải thì Không bị rám.
L'3: Nếu tóc Râm, Thấp, Vừa phải thì Không bị rám.
L'6: Nếu tóc Râm, Cao, Nặng thì Không bị rám.
L'7: Nếu tóc Râm, Tầm thước, Nặng thì Không bị rám.
L'8: Nếu tóc Đen, Thấp, Nhẹ thì Không bị rám.

Kết hợp {M,T} => {K} và {M,C,N} => {K} , ta có tổng cộng các luật sau:

L1: Nếu tóc Đen, Không dùng thuốc thì Bị rám.
L5: Nếu tóc Bạc, Không dùng thuốc thì Bị rám.
L2: Nếu tóc Đen, Có dùng thuốc thì Không bị rám.
L3: Nếu tóc Râm thì Không bị rám.
L'1: Nếu tóc Đen, Tầm thước, Nhẹ thì Bị rám.
L'4: Nếu tóc Đen, Thấp, Vừa phải thì Bị rám
L'5: Nếu tóc Bạc, Tầm thước, Nặng thì Bị rám.
L'2: Nếu tóc Đen, Cao, Vừa phải thì Không bị rám.
L'3: Nếu tóc Râm, Thấp, Vừa phải thì Không bị rám.

L'6: Nếu tóc Râm, Cao, Nặng thì Không bị rám.
L'7: Nếu tóc Râm, Tầm thước, Nặng thì Không bị rám.
L'8: Nếu tóc Đen, Thấp, Nhẹ thì Không bị rám.

Loại bỏ các luật thừa , ta có :

L1: Nếu tóc Đen, Không dùng thuốc thì Bị rám.
L5: Nếu tóc Bạc, Không dùng thuốc thì Bị rám.
L2: Nếu tóc Đen, Có dùng thuốc thì Không bị rám.
L3: Nếu tóc Râm thì Không bị rám.
L'5: Nếu tóc Bạc, Tầm thước, Nặng thì Bị rám.

Kết hợp L5 và L'5 , ta có các luật sau cùng:

L1: Nếu tóc Đen, Không dùng thuốc thì Bị rám.
L5: Nếu tóc Bạc thì Bị rám.
L2: Nếu tóc Đen, Có dùng thuốc thì Không bị rám.
L3: Nếu tóc Râm thì Không bị rám.

<u>Kết luận</u>: Kết quả tạo luật từ cây quyết định và kết quả tạo luật từ rút gọn các reducts thì giống nhau.

Bài Tập 3: Episodes

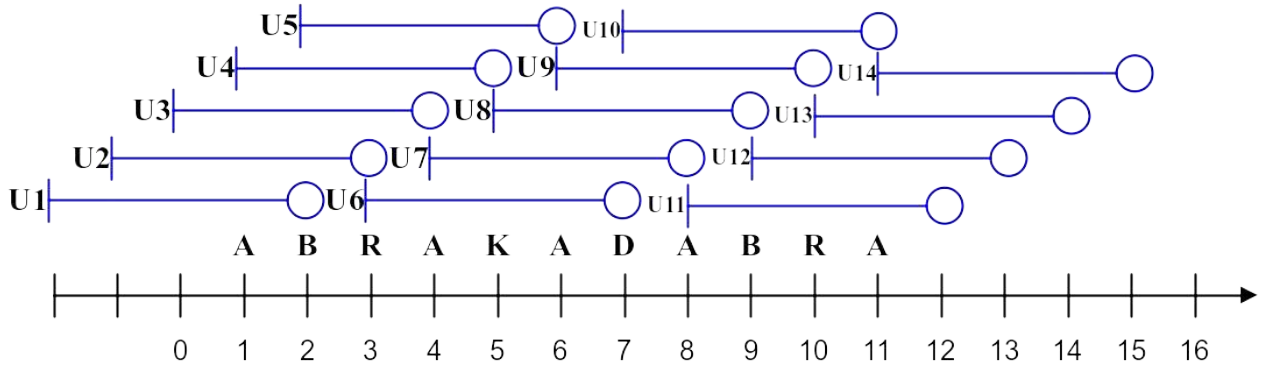
Cho chuỗi sự kiện sau đây:

A B R A K A D A B R A

1) Có bao nhiêu cửa sổ có bề rộng là 4 sự kiện được xử lý để tìm các episodes phổ biến theo tiếp cận WINEPI ?
2) Giả sử ngưỡng min_fr là 0.3. Tìm các episode phổ biến tuần tự và song song trong chuỗi sự kiện nêu trên trên ?
3) Tìm các epsiode tối đại ?

Bài Giải :

1\ Có bao nhiêu cửa sổ có bề rộng là 4 sự kiện được xử lý để tìm các episodes phổ biến theo tiếp cận WINEPI?



- Bằng cách trượt cửa sổ, chúng ta có 14 cửa sổ, có bề rộng là 4 sự kiện:

Cửa Sổ U_i	Nội dung của U_i	Episodes song song xảy ra trong U_i
$U_{1,[-2,2]}$	[-,-,-,A]	{A}
$U_{2,[-1,3]}$	[-,-,A,B]	{A,B},{AB}
$U_{3,[0,4]}$	[-,A,B,R]	{A,B,R},{AB,AR,BR},{ABR}
$U_{4,[1,5]}$	[A,B,R,A]	{A,B,R},{AB,AR,BR},{ABR}
$U_{5,[2,6]}$	[B,R,A,K]	{A,B,K,R},{AB,AK,AR,BK,BR,KR},{ABK,ABR,AKR,BKR},{ABKR}
$U_{6,[3,7]}$	[R,A,K,A]	{A,K,R},{AK,AR,KR},{AKR}
$U_{7,[4,8]}$	[A,K,A,D]	{A,D,K},{AD,AK,DK},{ADK}
$U_{8,[5,9]}$	[K,A,D,A]	{A,D,K},{AD,AK,DK},{ADK}
$U_{9,[6,10]}$	[A,D,A,B]	{A,B,D},{AB,AD,BD},{ABD}
$U_{10,[7,11]}$	[D,A,B,R]	{A,B,D,R},{AB,AD,AR,BD,BR,DR},{ABD,ABR,ADR,BDR},{ABDR}
$U_{11,[8,12]}$	[A,B,R,A]	{A,B,R},{AB,AR,BR},{ABR}
$U_{12,[9,13]}$	[B,R,A,-]	{A,B,R},{AB,AR,BR},{ABR}
$U_{13,[10,14]}$	[R,A,-,-]	{A,R},{AR}
$U_{14,[11,15]}$	[A,-,-,-]	{A}

2\ Giả sử ngưỡng min_fr là 0.3. Tìm các episode phổ biến tuần tự và song song trong chuỗi sự kiện nêu trên trên ?

a\ Tìm các Episode song song

- Xây dựng episode L1 chứa 1 phần tử:

Episode	Tần suất của Episode
A	fr(A , S ,W) = 14/14 = 1 > min_fr
B	fr(B , S ,W) = 8/14 = 0.57 > min_fr
D	fr(D , S ,W) = 4/14 = 0.29 < min_fr
K	fr(K , S ,W) = 4/14 = 0.29 < min_fr
R	fr(R , S ,W) = 8/14 = 0.57 > min_fr
Tập Episode phổ biến song song có 1 sự kiện là L1 = {A, B, R}	

- Xây dựng episode L2 chứa 2 phần tử:

Episode	Tần suất của Episode
AB	fr(AB , S ,W) = 8/14 = 0.57 > min_fr
AR	fr(AR , S ,W) = 8/14 = 0.57 > min_fr
BR	fr(BR , S ,W) = 6/14 = 0.43 > min_fr
Tập Episode phổ biến song song có 2 sự kiện là L2 = {AB, AR, BR}	

- Xây dựng episode L3 chứa 3 phần tử:

Episode	Tần suất của Episode
ABR	fr(ABR , S ,W) = 6/14 = 0.43 > min_fr
Tập Episode phổ biến song song có 3 sự kiện là L3 = {ABR}	

- Không có Episode phổ biến song song có 4 sự kiện từ L3

Tập Episode phổ biến song song

= L1 ∪ L2 ∪ L3

= {A, B, R, AB, AR, BR, ABR}

Tập phổ biến song song tối đại chính là tập phổ biến L3 = {ABR}

b\ Tìm các Episode tuần tự

- Xây dựng episode L1 chứa 1 phần tử:

Episode	Tần suất của Episode
A	fr(A , S ,W) = 14/14 = 1 > min_fr
B	fr(B , S ,W) = 8/14 = 0.57 > min_fr
D	fr(D , S ,W) = 4/14 = 0.29 < min_fr
K	fr(K , S ,W) = 4/14 = 0.29 < min_fr
R	fr(R , S ,W) = 8/14 = 0.57 > min_fr
Tập Episode phổ biến song song có 1 sự kiện là L1 = {A, B, R}	

- Xây dựng episode L2 chứa 2 phần tử:

Episode	Tần suất của Episode
AB	fr(AB , S ,W) = 6/14 = 0.43 > min_fr
BA	fr(BA , S ,W) = 4/14 = 0.29 < min_fr
AR	fr(AR , S ,W) = 4/14 = 0.29 < min_fr
RA	fr(RA , S ,W) = 6/14 = 0.43 > min_fr
BR	fr(BR , S ,W) = 6/14 = 0.43 > min_fr
RB	fr(RB , S ,W) = 0/14 = 0.00 < min_fr
Tập Episode phổ biến song song có 2 sự kiện là L2 = { AB, RA, BR}	

- Xây dựng episode L3 chứa 3 phần tử:

Episode	Tần suất của Episode
ABR	fr(ABR, S ,W) = 4/14 = 0.29 < min_fr
RAB	fr(RAB, S ,W) = 0/14 = 0.00 < min_fr
BRA	fr(BRA, S ,W) = 4/14 = 0.29 < min_fr
Tập Episode phổ biến song song có 3 sự kiện là L3 = {}	

- Không có Episode phổ biến song song có 4 sự kiện từ L3

Tập Episode phổ biến song song

= L1 ∪ L2

= {A, B, R, AB, RA, BR}

Tập phổ biến song song tối đại chính là tập phổ biến L2= {AB,RA,BR }