BÀI TẬP VÀ ĐỀ THI MÔN KHAI PHÁ DỮ LIỆU

Contents

ĐỀ 1

ĐỀ THI MÔN DATAMINING

Thời gian: 120 phút (Được phép sử dụng tài liệu)

1. Cho bối cảnh khai thác dữ liệu như sau (4 điểm)

	i1	i2	i3	i4	i5
o1	1	0	1	1	0
o2	1	0	1	0	0
o3	0	0	1	0	1
o4	1	1	0	1	1
o5	0	1	0	1	0
o6	1	1	0	1	1

   1.1 Tìm các tập phổ biến tối đại theo ngưỡng minsupp=0.3
   1.2 Tìm các luật kết hợp từ tập phổ biến tối đại với ngưỡng minconf=1.0
2. Cho bảng quyết định sau (4 điểm)

	Vóc dáng	Quốc tịch	Gia cảnh	Nhóm
O1	Nhỏ	Đức	Độc thân	A
O2	Lớn	Pháp	Độc thân	A
O3	Lớn	Đức	Độc thân	A
O4	Nhỏ	Ý	Độc thân	B
O5	Lớn	Đức	Có gia đình	B
O6	Lớn	Ý	Độc thân	B
O7	Lớn	Ý	Có gia đình	B
O8	Nhỏ	Đức	Có gia đình	B

   2.1 Tìm các luật phân lớp của bảng quyết định trên với
   - Tập thuộc tính điều kiện là {Vóc dáng, Quốc tịch, Gia cảnh}
   - Thuộc tính phân lớp là {Nhóm}

   2.2 Tìm các reducts bảng quyết định trên và liệt kê các luật phân lớp có số thuộc tính vế trái nhỏ nhất
3. Trình bày một ứng dụng cụ thể của CSDL dạng khối 3 chiều và nêu lên một số thao tác trên CSDL dạng khối mà CSDL quan hệ khó thực hiện (2 điểm)

ĐỀ THI MÔN DATAMINING

Thời gian: 120 phút (Được phép sử dụng tài liệu)

4. Cho bối cảnh khai thác dữ liệu như sau (4 điểm)

	i1	i2	i3	i4	i5
o1	1	0	1	1	0
o2	1	0	1	0	0
o3	0	0	1	0	1
o4	1	1	0	1	1
o5	0	1	0	1	0
o6	1	1	0	1	1

4.1 Tìm các tập phổ biến tối đại theo ngưỡng minsupp=0.3
4.2 Tìm các luật kết hợp từ tập phổ biến tối đại với ngưỡng minconf=1.0

5. Cho bảng quyết định sau (4 điểm)

	Vóc dáng	Quốc tịch	Gia cảnh	Nhóm
O1	Nhỏ	Đức	Độc thân	A
O2	Lớn	Pháp	Độc thân	A
O3	Lớn	Đức	Độc thân	A
O4	Nhỏ	Ý	Độc thân	B
O5	Lớn	Đức	Có gia đình	B
O6	Lớn	Ý	Độc thân	B
O7	Lớn	Ý	Có gia đình	B
O8	Nhỏ	Đức	Có gia đình	B

5.1 Tìm các luật phân lớp của bảng quyết định trên với
- Tập thuộc tính điều kiện là {Vóc dáng, Quốc tịch, Gia cảnh}
- Thuộc tính phân lớp là {Nhóm}

5.2 Tìm các reducts bảng quyết định trên và liệt kê các luật phân lớp có số thuộc tính vế trái nhỏ nhất

6. Trình bày một ứng dụng cụ thể của CSDL dạng khối 3 chiều và nêu lên một số thao tác trên CSDL dạng khối mà CSDL quan hệ khó thực hiện (2 điểm)

BÀI GIẢI

Câu 1:

1.1 Tìm các tập phổ biến tối đại theo ngưỡng minsupp=0.3

Tính F1:

Supp({i1}) = 4/6 = 0.66

Supp({i2}) = 3/6 = 0.5

Supp({i3}) = 3/6 =0.5

Supp({i4}) = 4/6 = 0.66

Supp({i5}) = 3/6 =0.5

Vậy: F1 = {{i1},{i2},{i3},{i4},{i5}}

Tính C2 từ F1:

	i1	i2	i3	i4	i5
i1					
i2	i1,i2				
i3	i1,i3	i2,i3			
i4	i1,i4	i2,i4	i3,i4		
i5	i1,i5	i2,i5	i3,i5	i4,i5	

C2 = {{i1,i2},{i1,i3},{i1,i4},{i1,i5},{i2,i3},{i2,i4},{i2,i5},{i3,i4},{i3,i5},{i4,i5}}

Từ C2 tính F2:

Supp({i1,i2}) = 2/6 = 0.3

Supp({i1,i3}) = 2/6 = 0.3

Supp({i1,i4}) = 3/6 = 0.5

Supp({i1,i5}) = 2/6 = 0.3

Supp({i2,i3}) = 0/6 = 0 < minsupp : **loại**

Supp({i2,i4}) = 3/6 = 0.5

Supp({i2,i5}) = 2/6 = 0.3

Supp({i3,i4}) = 1/6 = 0.17 < minsupp: **loại**

Supp({i3,i5}) = 1/6 = 0.17 < minsupp: **loại**

Supp({i4,i5}) = 2/6 = 0.3

Vậy: F2 = { {i1,i2},{i1,i3},{i1,i4},{i1,i5},{i2,i4},{i2,i5},{i4,i5}}

Tính C3 từ F2:

	{i1,i2}	{i1,i3}	{i1,i4}	{i1,i5}	{i2,i4}	{i2,i5}	{i4,i5}
{i1,i2}							
{i1,i3}	{i1,i2,i3}						
{i1,i4}	{i1,i2,i4}	{i1,i3,i4}					
{i1,i5}	{i1,i2,i5}	{i1,i3,i5}	{i1,i4,i5}				
{i2,i4}	{i1,i2,i4}	{i1,i2,i3,i4}	{i1,i2,i4}	{i1,i2,i4,i5}			
{i2,i5}	{i1,i2,i5}	{i1,i2,i3,i5}	{i1,i2,i4.i5}	{i1,i2,i5}	{i2,i4,i5}		
{i4,i5}	{i1,i2,i4,i5}	{i1,i3,i4,i5}	{i1,i4,i5}	{i1,i4,i5}	{i2,i4,i5}	{i2,i4,i5}	

C3 = {{i1,i2,i3},{i1,i2,i4},{i1,i2,i5},{i1,i3,i4},{i1,i3,i5},{i1,i4,i5},{i2,i4,i5}}

Từ C3 tính F3:

Theo nguyên lý Apriori, ta loại các tập sau:

Loại {i1,i2,i3} vì {i2,i3} không có trong F2

Loại {i1,i3,i4} vì {i3,i4} không có trong F2

Loại {i1,i3,i5} vì {i3,i5} không có trong F2

Supp({i1,i2,i4}} = 2/6 = 0.3

Supp({i1,i2,i5}} = 2/6 = 0.3

Supp({i1,i4,i5}} = 2/6 = 0.3

Supp({i2,i4,i5}} = 2/6 = 0.3

Vậy: F3 = {{i1,i2,i4},{i1,i2,i5},{i1,i4,i5},{i2,i4,i5}}

Tính C4 từ F3:

	{i1,i2,i4}	{i1,i2,i5}	{i1,i4,i5}	{i2,i4,i5}
{i1,i2,i4}				

{i1,i2,i5}	{i1,i2,i4,i5}			
{i1,i4,i5}	{i1,i2,i4,i5}	{i1,i2,i4,i5}		
{i2,i4,i5}	{i1,i2,i4,i5}	{i1,i2,i4,i5}	{i1,i2,i4,i5}	

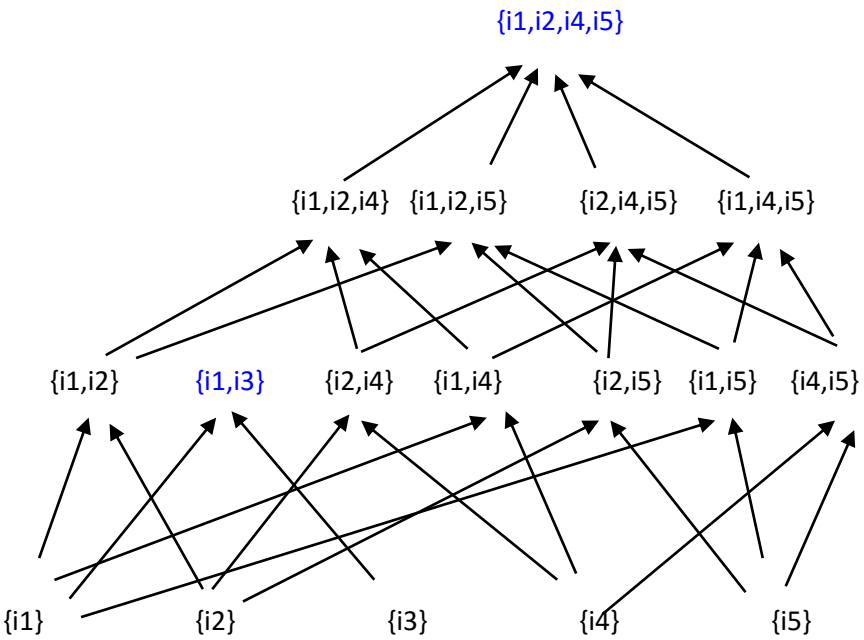
C4 = {{i1,i2,i4,i5}}

Tính F4:

Supp({i1,i2,i4,i5}) = 2/6 =0.3

Vậy: F4 = {{i1,i2,i4,i5}}

Tập phổ biến tối đại: {i1,i3}, {i1,i2,i4,i5}



1.2 Tìm các luật kết hợp từ tập phổ biến tối đại với ngưỡng minconf=1.0

Định nghĩa : cho I: tập các item, O: tập các giao tác. Ta định nghĩa ánh xạ ρ: I $\rightarrow$ O, khi đó S $\subseteq$ I thì: $\rho(S)$ = {o | $\forall i \in S, \rho(i) = o$}, ie. tập các giao tác có chứa S.

Cho luật kết hợp S1=>S2. Conf(S1=>S2) = | $\rho(S1) \cap \rho(S2)$|/| $\rho(S1)$|.

Conf(S1=>S2) = 1.0 khi và chỉ khi $\rho(S1) \subseteq \rho(S2)$ vì $\rho(S1) \cap \rho(S2) = \rho(S1)$.

- Xét tập phổ biến tối đại{i1,i3}, các luật kết hợp khả dĩ là : i1=>i2 và i2=>i1.

Ta có: $\rho(i1) = \{o1,o2,o4,o6\}$, $\rho(i2) = \{o4,o5,o6\}$, nên: $\rho(i1) \not\subset \rho(i2)$ và $\rho(i2) \not\subset \rho(i1)$

Do đó i1=>i2 và i2=>i1 không là luật kết hợp.

- Xét tập phổ biến tối đại{i1,i2,i4,i5}:

Luật : S1=>S2	**ρ(S1)**	**ρ(S2)**	**ρ(S1)⊆ ρ(S2)**
{i1}=>{i2,i4,i5}	{o1,o2,o4,o6}	{o4,o6}	
{i2,i4,i5}=>{i1}	{o4,o6}	{o1,o2,o4,o6}	x
{i1,i2} => {i4,i5}	{o4,o6)}	{o4,o6}	x
{i4,i5}=>{i1,i2}	{o4,o6}	{o4,o6}	x
{i1,i4}=> {i2,i5}	{o1,o4,o6}	{o4,o5,o6}	
{i2,i5}=>{i1,i4}	{o4,o5,o6}	{o1,o4,o6}	
{i1,i5}=>{i2,i4}	{o4,o6}	{o4,o5,o6}	x
{i2,i4}=>{i1,i5}	{o4,o5,o6}	{o4,o6}	
{i1,i2,i4}=>{i5}	{o4,o6}	{o3,o4,o6}	x
{i5}=>{i1,i2,i4}	{o3,o4,o6}	{o4,o6}	
{i1,i2,i5}=>{i4}	{o4,o6}	{o1,o4,o5,o6}	x
{i4}=>{i1,i2,i5}	{o1,o4,o5,o6}	{o4,o6}	
{i1,i4,i5}=>{i2}	{o4,o6}	{o4,o5,o6}	x
{i2}=>{i1,i4,i5}	{o4,o5,o6}	{o4,o6}	

Ta có các luật kết hợp:

L1: {i2,i4,i5}=>{i1}

L2: {i1,i2} => {i4,i5}

L3: {i4,i5}=>{i1,i2}

L4: {i1,i5}=>{i2,i4}

L5: {i1,i2,i4}=>{i5}

L6: {i1,i2,i5}=>{i4}

L7: {i1,i4,i5}=>{i2}

Câu 2:

2.1 Tìm các luật phân lớp của bảng quyết định :

	Vóc dáng	Quốc tịch	Gia cảnh	Nhóm
O1	Nhỏ	Đức	Độc thân	A
O2	Lớn	Pháp	Độc thân	A
O3	Lớn	Đức	Độc thân	A
O4	Nhỏ	Ý	Độc thân	B
O5	Lớn	Đức	Có gia đình	B
O6	Lớn	Ý	Độc thân	B
O7	Lớn	Ý	Có gia đình	B
O8	Nhỏ	Đức	Có gia đình	B

Đặt :

P=A, N= B;

p: số phần tử thuộc lớp P, p = 3;

n: số phần tử thuộc lớp N, n = 5;

Ta có: $I(p,n) = I(3,5) = -3/8*\log_2 3/8-5/8*\log_2 5/8 = 0.954$

Tính độ lợi thông tin cho các thuộc tính điều kiện:

Vóc dáng	pi	ni	I(pi,ni)
Nhỏ	1	2	0.92
Lớn	2	3	0.97

E(Vóc dáng) = 3/8*I(1,2)+5/8*I(2,3) = 3/8*0.92+5/8*0.97 = 0.951

G(Vóc dáng) = I(p,n) – E(Vóc dáng) = 0.954 – 0.951 = 0.003

Quốc tịch	pi	ni	I(pi,ni)
Đức	2	2	1
Pháp	1	0	0
Ý	0	3	0

E(Quốc tịch) = 4/8*I(2,2)+1/8*I(1,0)+3/8*I(0,3) = 4/8*1 = 0.5

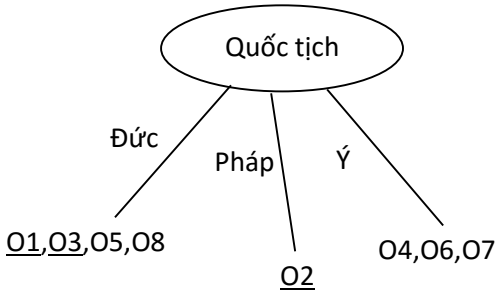
G(Quốc tịch) = I(p,n) – E(Quốc tịch) = 0.954 – 0.5 = **0.454**

Gia cảnh	pi	ni	I(pi,ni)
Độc thân	3	2	0.97
Có gia đình	0	3	0

E(Gia cảnh) = 5/8*I(3,2)+3/8*I(0,3) = 5/8*0.97 = 0.606

G(Gia cảnh) = I(p,n) – E(Gia cảnh) = 0.954-0.606 = 0.348

Thuộc tính Quốc tịch có độ lợi thông tin lớn nhất, nên được chọn để phân lớp:



(**Gạch dưới**: thuộc lớp A,
Không gạch dưới: thuộc lớp B)

Phân lớp nhóm Quốc tịch - Đức:

Bảng dữ liệu còn lại:

	Vóc dáng	Gia cảnh	Nhóm
O1	Nhỏ	Độc thân	A
O3	Lớn	Độc thân	A
O5	Lớn	Có gia đình	B
O8	Nhỏ	Có gia đình	B

Ta có: $I(p,n) = -2/4*\log_2 2/4 - 2/4*\log_2 2/4 = 1$

Vóc dáng	pi	ni	I(pi,ni)
Nhỏ	1	1	1
Lớn	1	1	1

E(Vóc dáng) = 2/4*I(1,1)+2/4*I(1,1) = 2/4*1+2/4*1 = 1

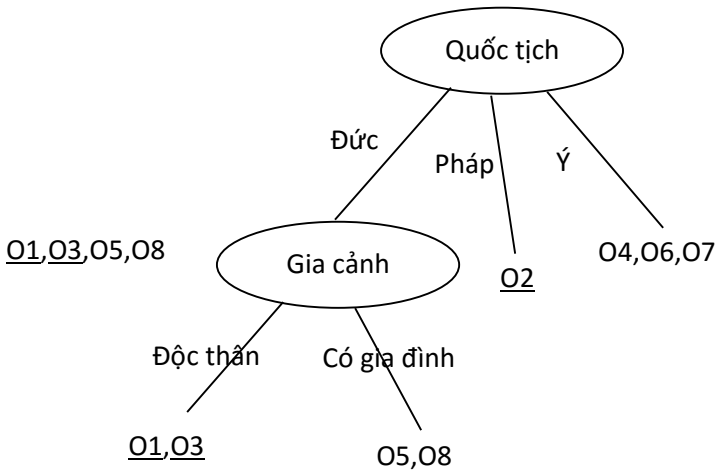
G(Vóc dáng) = I(p,n) – E(Vóc dáng) = 1 -1 = 0

Gia cảnh	pi	ni	I(pi,ni)
Độc thân	2	0	0
Có gia đình	0	2	0

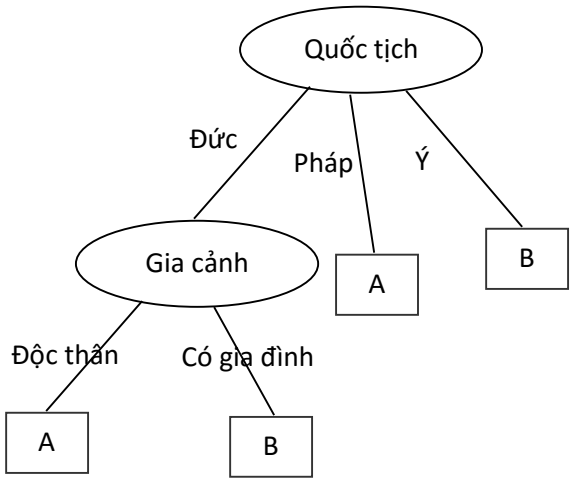
E(Gia cảnh) = 0

G(Gia cảnh) = **1**

Thuộc tính Gia cảnh được chọn để phân lớp:



Cây quyết định:



Các luật phân lớp:

L1: Nếu có Quốc tịch Đức và Gia cảnh Độc thân thì thuộc về nhóm A

L2: Nếu có Quốc tịch Đức và Gia cảnh Có gia đình thì thuộc về nhóm B

L3: Nếu có Quốc tịch Pháp thì thuộc về nhóm A

L4: Nếu có Quốc tịch Ý thì thuộc về nhóm B

Rút gọn luật:

Các luật trên đều không dư thừa.

2.2 Tìm các reducts bảng quyết định trên và liệt kê các luật phân lớp có số thuộc tính vế trái nhỏ nhất

	Vóc dáng	Quốc tịch	Gia cảnh	Nhóm
O1	Nhỏ	Đức	Độc thân	A
O2	Lớn	Pháp	Độc thân	A
O3	Lớn	Đức	Độc thân	A
O4	Nhỏ	Ý	Độc thân	B
O5	Lớn	Đức	Có gia đình	B
O6	Lớn	Ý	Độc thân	B
O7	Lớn	Ý	Có gia đình	B
O8	Nhỏ	Đức	Có gia đình	B

Ký hiệu : Q: Quốc tịch, V: Vóc dáng, G: Gia cảnh

Ma trận phân biệt:

	O1	O2	O3	O4	O5	O6	O7	O8
O1								
O2	λ							
O3	λ	λ						
O4	Q	V,Q	V,Q					
O5	V,G	Q,G	G	λ				
O6	V,Q	Q	Q	λ	λ			
O7	V,Q,G	Q,G	Q,G	λ	λ	λ		
O8	G	V,Q,G	V,G	λ	λ	λ	λ	

Từ ma trận phân biệt, ta có hàm phân biệt:

$F(V,Q,G) = Q \wedge (V \vee G) \wedge (V \vee Q) \wedge (V \vee Q \vee G) \wedge G \wedge (Q \vee G)$

Sử dụng luật hút: $p \wedge (p \vee q) = p$, ta có:

Q∧(V∨Q) = Q

Q∧(V∨Q∨G) = Q

G∧(V∨G) = G

G∧(Q∨G) = G

Vậy: F(V,Q,G) = Q∧G

Reduct: {Q,G}

Tìm luật từ reduct: {Q,G}

Đặt: O = {O1,O2,…,O8}

X1 = {Oi : Nhóm = A, i=1..8} = {O1,O2,O3}

X2 = {Oi : Nhóm = B, i=1..8} = {O4,O5,O6,O7,O8}

Xét phân hoạch O/Q = { {O1,O3,O5,O8},{O2},{O4,O6,O7}}

Lower(X1,Q) = {O2}

Lower(X2,Q) ={O4,O6,O7}

k= (|Lower(X1,Q)| + |Lower(X2,Q)|)/|O| = 4/8 < 1

Nên ta có luật phân lớp không đúng chính xác 100%: Q=>D (với D ={Nhóm})

Xét phân hoạch O/G = {{O1,O2,O3,O4,O6},{O5,O7,O8}}:

Lower(X1,G) = {O2}= {}

Lower(X2,G) ={O5,O7,O8}

k= (|Lower(X1,G)| + |Lower(X2,G)|)/|O| = 3/8 < 1

Nên ta có luật phân lớp không đúng chính xác 100%: G =>D

Xét phân hoạch O/QG = {{O1,O3},{O5,O8},{O2},{O4,O6},{O7}}: {O4,O5,O6,O7,O8}

Lower(X1,QG) = {O1,O2,O3}

Lower(X2,QG) = {O4,O5,O6,O7,O8}

k= (|Lower(X1,QG)| + |Lower(X2,QG)|)/|O| = 8/8 = 1

Nên ta có luật phân lớp đúng chính xác 100%: QG =>D

Các luật phân lớp có số thuộc tính vế trái nhỏ nhất:

Từ Q=>D , ta có các luật phân lớp:

L1: Nếu có Quốc tịch Pháp thì thuộc về nhóm A ({O2})

L2: Nếu có Quốc tịch Ý thì thuộc về nhóm B ({O4,O6,O7})

Từ G=>D, ta có các luật phân lớp:

L3: Nếu Có gia đình thì thuộc nhóm B ({O5,O7,O8})

ĐỀ 2

ĐỀ THI DATA MINING KHÓA 2

Câu 1:

	Kích thước	Màu sắc	Hình dạng	Lớp
1	Vừa	Xanh	Viên gạch	A
2	Nhỏ	Đỏ	Hình nêm	B
3	Nhỏ	Đỏ	Hình cầu	A
4	Lớn	Đỏ	Hình nêm	B
5	Lớn	Lục	Hình trụ	A
6	Lớn	Đỏ	Hình trụ	B
7	Lớn	Lục	Hình cầu	A

a) Tính các reduce tương đối của bảng quyết định trên.
b) Tìm các luật phân lớp được tạo lập dựa trên các reduce tương đối tìm được trong câu a)

Câu 2:

Bài tập về tập mặt hàng và tập giao tác I = {i1,…….i8}, O = {o1,…..o6}

o1={i1,i7,8}
o2={i1,i2,i6,i7,i8}
o3={i1,i2,i6,i7}
o4={i1,i7,i8}
o5={i3,i4,i5,i6,i8}
o6={i1,i4,i5}

a. Tìm ngữ cảnh khai thác dữ liệu được tạo từ I, O.
b. Tìm tất cả các tập phổ biến theo ngưỡng minsupp=0,3
c. Tìm tất cả các tập phổ biến tối đại theo ngưỡng minsupp=0,3
d. Tìm tất cả các luật kết hợp hợp lệ theo ngưỡng minsupp=0,3 và ngưỡng minconf=1 được tạo từ các tập phổ biến tối đại của câu 2c.
e. Anh chị có suy nghĩ gì về một thuật toán tìm tập phổ biến tối đại.

GIẢI ĐỀ THI KHOA 2

CÂU 1:

a).Tính các Reduct tương đối của bảng quyết định trên

Ký hiệu:

a: kích thước

b: màu sắc
c: hình dáng
Ta được ma trận phân biệt như sau:

	1	2	3	4	5	6	7
1							
2	abc						
3	λ	c					
4	abc	λ	ac				
5	λ	abc	λ	bc			
6	abc	λ	ac	λ	b		
7	λ	abc	λ	bc	λ	bc	

Từ ma trận phan biệt ta có hàm phân:
F(a,b,c) = (a∨b∨c)∧c∧(a∨c)∧(b∨c)∧b
Sử dụng luật hút : (a∨b)∧a = a
(a∨b∨c)∧c = c
Ta được:
F(a,b,c) = (b ∧ c)

Vậy Reduct: {b, c} hay {Màu sắc, Hình dạng}

b).Tìm các luật phân lớp được tạo lập dựa trên Reduct tương đối tìm được trong câu a.

1 Tính R dương của D
- Tính U/D

U/D = {X1, X2} với X1={1, 3, 5, 7}; X2={2, 4, 6}
- Tính U/R

U/R={{1}, {2, 4, 6}, {3}, {5}, {7}}
- Tính R dương của D

R duong của D = $\underline{R}$X1 ∪ $\underline{R}$X2 = {1, 3, 5, 7} ∪ {2, 4, 6} = U

1 Lấy một phần tử A của U/R ghép với 1 phần tử B thuộc U/D

	A, B	A∩B	A∩B<>∅	A⊆B	Kết quả
1	{1}, {1, 3, 5, 7}	{1}	yes	yes	có luật conf=1
2	{1}, {2, 4, 6}	{}	no	no	không có luật
3	{2, 4, 6}, {1,3,5,7}	{}	no	no	không có luật
4	{2,4,6}, {2,4,6}	{2,4,6}	yes	yes	có luật conf=1
5	{3}, {1,3,5,7}	{3}	yes	yes	có luật conf=1
6	{3}, {2,4,6}	{}	no	no	không có luật
7	{5}, {1,3,5,7}	{5}	yes	yes	có luật conf=1
8	{5}, {2,4,6}	{}	no	no	không có luật

9	{7}, {1,3,5,7}	{7}	yes	yes	có luật cong=1
10	{7}, {2,4,6}	{}	no	no	không có luật

2 Vậy ta có các luật phân lớp như sau:

1.Nếu Màu sắc = Xanh và Hình dạng = Viên gạch ⇒ Lớp A
2.Nếu Màu sắc = Đỏ và Hình dạng = Hình nêm ⇒ Lớp B
3.Nếu Màu sắc = Đỏ và Hình dạng = Hình cầu ⇒ Lớp A
4.Nếu Màu sắc = Lục và Hình dạng = Trụ ⇒ Lớp A
5.Nếu Màu sắc = Lục và Hình dạng = Hình cầu ⇒ Lớp B

CÂU 2:

a.Tìm ngữ cảnh khai thác dữ liệu được tạo từ I, O

Ta có bối cảnh nhị phân

	i1	i2	i3	i4	i5	i6	i7	i8
O1	1	0	0	0	0	0	1	1
O2	1	1	0	0	0	1	1	1
O3	1	1	0	0	0	1	1	0
O4	1	0	0	0	0	0	1	1
O5	0	0	1	1	1	1	0	1
O6	1	0	0	1	1	0	0	0

b.Tìm các tập phổ biến theo ngưỡng minsupp=0.3

Với minsupp=0.3 số dòng là 6*0.3=1.8 hay 2 dòng

Suy ra F1 = {{i1}, {i2}, {i4}, {i5}, {i6}, {i7}, {i8}}

Tính C1

	i1	i2	i4	i5	i6	i7	i8
i1							
i2	i1,i2						
i4	i1,i4	i2,i4					
i5	i1,i5	i2,i5	i4,i5				
i6	i1,i6	i2,i6	i4,i6	i5,i6			
i7	i1.i7	i2,i7	i4,i7	i5,i7	i6,i7		
i8	i1,i8	i2,i8	i4,i8	i5,i8	i6,i8	i7,i8	

Tu C1 tinh F2

C1 = {{i1,i2}, {i1,i4}, {i1,i5}, {i1,i6}, {i1,i7}, {i1,8}, {i2,i4}, {i2,i5}, {i2,i6}, {i2, i7}, {i2,i8}, {i4,i5}, {i4,i6}, {i4,i7}, {i4,i8}, {i5,i6}, {i5,i7}, {i5,i8}, {i6,i7}, {i6,i8}, {i7,i8}}

F2 = {{i1,i2}, {i1,i6}, {i1,i7}, {i1,i8}, {i2,i6}, {i2,i7}, {i4,i5}, {i6,i7}, {i6,i8}, {i7,i8}}

	{i1,i2}	{i1,i6}	{i1,i7}	{i1,i8}	{i2,i6}	{i2,i7}	{i4,i5}	{i6,i7}	{i6,i8}	{i7,i8}
{i1,i2}										
{i1,i6}	{i1,i2,i6}									
{i1,i7}	{i1,i2,i7}	{i1,i6,i7}								
{i1,i8}	{i1,i2,i8}	{i1,i6,i8}	{i1,i7,i8}							
{i2,i6}	{i1,i2,i6}	{i1,i2,i6}	{i1,i2,i6,i7}	{i1,i2,i6,i8}						
{i2,i7}	{i1.i2,i7}	{i1,i2,i6,i7}	{i1,i2,i7}	{i1,i2,i7,i8}	{i2,i6,i7}					
{i4,i5}	{i1,i2,i4,i5}	{i1,i4,i5,i6}	{i1,i4,i5,i7}	{i1,i4,i5,i8}	{i2,i4,i5,i6}	{i2,i4,i5,i7}				
{i6,i7}	{i1,i2,i6,i7}	{i1,i6,i7}	{i1,i6,i7}	{i1,i6,i7,i8}	{i2,i6,i7}	{i2,i6,i7}	{i4,i5,i6,i7}			
{i6,i8}	{i1,i2,i6,i8}	{i1,i6,i8}	{i1,i6,i7,i8}	{i1,i6,i8}	{i2,i6,i8}	{i2,i6,i7,i8}	{i4,i5,i6,i8}	{i6,i7,i8}		
{i7,i8}	{i1,i2,i7,i8}	{i1,i6,i7,i8}	{i1,i7,i8}	{i1,i7,i8}	{i2,i6,i7,i8}	{i2,i7,i8}	{i4,i5,i7,i8}	{i6,i7,i8}	{i6,i7,i8}	

Tinh F3 tu C2
C2 = {{nguyen ban tren}}

F3 = {{i1,i2,i6}, {i1,i2,i7}, {i1,i6,i7}, {i1,i2,i6,i7}, {i2,i6,i7}}

c.Tìm tất cả tập phổ biến tối đại theo ngưỡng minsupp=0.3
Ta nhận thấy tập phổ cực đại chính là **F3={i1,i2,i6,i7}**

d.(Đến đây các bạn làm giống bài mẫu)

ĐỀ 3

Câu 1: Cho tập mặt hàng : {i1,i2,i3,i4,i5,i6} và 6 giao tác

T1={i1,i2} ; T2={i1,i2,i3}, T3={i1,i2,i5} ;

T4={t1,t2,t5,t6} ; T5 = {i3,i4,i5,i6}

1.1 Tìm tất cả các tập phổ biến có minsupp=0.3
1.2 Tìm tất cả các tập phổ biến tối đại có minsupp=0.3
1.3 Tìm tất cả các lụật kết hợp có mincof-=1.0 từ các tập phổ biến tối đại ở câu 1.2

Giải:

1.1 Tìm tất cả các tập phổ biến có minsupp=0.3

Bối cảnh nhị phân

	i1	i2	i3	i4	i5	i6
T1	1	1	0	0	0	0
T2	1	1	1	0	0	0
T3	1	1	0	0	1	0
T4	1	1	0	0	1	1
T5	0	0	1	1	1	1

T6	1	1	1	0	0	0

Với minsupp =0,3 , số dòng là 6*0,3 =1,8 hay 2 dòng

F1={{i1},{i2},{i3},{d5},{d6}}

Tính C1

	i1	i2	i3	i5	i6
i1					
i2	i1,i2				
i3	i1,i3	i2,i3			
i5	i1,i5	i2,i5	i3,i5		
i6	i1,i6	i2,i6	i3,i6	i5,i6	

Từ C1 tính F2:

C1={{ i1,i2}, {i1,i3}, {i1,i5},{i1,i6},{i2,i3},{i2,i5},{i2,i6},{i3,i5},{i3,i6},{i5,i6}}

F2={ {i1,i2}, {i1,i3}, {i2,i3},{i1,i5},{i2,i5},{i5,i6} }

Tính C2

	i1i2	i1i3	i2i3	i1i5	i2i5	i5i6
i1i2						
i1i3	i1,i2,i3					
i2i3	i1,i2,i3	i1,i2,i3				
i1i5	i1,i2,i5	i1,i3,i5	i1,i2,i3,i5			
i2i5	i1,i2,i5	i1,i2,i3,i5	i2,i3,i5	i1,i2,i5		
i5i6	i1,i2,i5,i6	i1,i3,i5,i6	i2,i3,i5,i6	i1,i5,i6	i2,i5,i6	

C2={{i1,i2,i3} ,{i1,i2,i5} , { i1,i3,i5 },{ i2,i3,i5 }{ i1,i5,i6 }{ i2,i5,i6 }}

F3={{i1,i2,i3}, {i1,i2,i5}}

Tính C3

	i1i2i3	i1i2i5
i1i2i3		
i1i2i5	i1,i2,i3,i5	

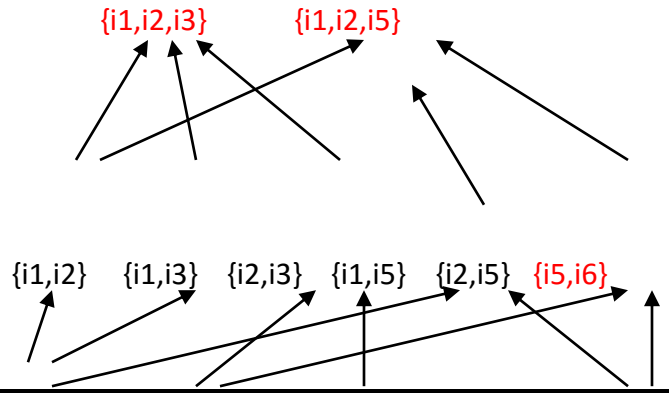
C3={{i1,i2,i3,i5}}

F4={∅}

Tập phổ biến là F1, F2, F3

1.2 Tìm tất cả các tập phổ biến tối đại có minsupp=0.3

Tập phổ biến tối đại: {i1,i2,i3}, {i1,i2,i5},{i5,i6}



{i1} {i2} {i3} {i4} {i5} {i6}

1.3 Tìm tất cả các luật kết hợp có mincof-=1.0 từ các tập phổ biến tối đại ở câu 1.2

Tạo luật kết hợp từ các tập tối đại:

Định nghĩa $\rho : I \rightarrow O$ với I : tập mặt hàng và O tập giao tác

Cho $S \subseteq O$, $\rho(S) = \{ o \in O \mid \forall i \in S$, giao tác o có mặt hàng i $\}$

Ý nghĩa $\rho(S)$ là tập các giao tác có chứa tất cả các mặt hàng trong S.

Cho luật kết hợp $S1 \rightarrow S2$,

$CF(S1 \rightarrow S2) = |\rho(S1) \cap \rho(S2)| / |\rho(S1)|$

Ta nhận thấy $CF(S1 \rightarrow S2) = 1.0$ khi và chỉ khi $\rho(S1) \subseteq \rho(S2)$

Lúc đó $\rho(S1) \cap \rho(S2) = \rho(S1)$

- Với tập phổ biến tối đại : {i1,i2,i3}

Các luật khả dĩ:

{i1} $\rightarrow$ {i2,i3}

{i2} $\rightarrow$ {i1,i3}

{i3} $\rightarrow$ {i1,i2}

{i2,i3} $\rightarrow$ {i1}

{i1,i3} $\rightarrow$ {i2}

{i1,i2} $\rightarrow$ {i3}

$\rho(\{i1\}) = \{T1,T2,T3,T4,T6\}$

$\rho(\{i2\}) = \{T1,T2,T3,T4,T6\}$

ρ({i3}) = {T2,T5,T6}

ρ({i1,i2}) = {T1,T2,T3,T4,T6}

ρ({i1,i3}) = {T2,T6}

ρ({i2,i3}) = {T2,T6}

Vậy ta có 2 luật thoả:

{i2,i3}→ {i1}

{i1,i3} → {i2}

- Với tập phổ biến tối đại : {i1,i2,i5}

Các luật khả dĩ:

{i1} →{i2,i5}

{i2} →{i1,i5}

{i5} →{i1,i2}

{i2,i5}→ {i1}

{i1,i5} → {i2}

{i1,i2} → {i5}

ρ({i5}) = {T3,T4,T5}

ρ({i1,i5}) = {T3,T4}

ρ({i2,i5}) = {T3,T4}

Vậy ta có 2 luật thoả:

{i2,i5}→ {i1}

{i1,i5} → {i2}

- Với tập phổ biến tối đại : {i5,i6}

Các luật khả dĩ:

{i5} →{i6}

{i6} →{i5}

ρ({i6}) = {T4,T5}

Vậy ta có 1 luật thoả: {i6} →{i5}

Tóm lại, có 5 luật:

{i2,i3}→ {i1}

{i1,i3} → {i2}

{i2,i5}→ {i1}

{i1,i5} → {i2}

{i6} →{i5}

LUẬT KẾT HỢP

Bài 2

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 minsup=0,4 hãy:

Câu1:

Tìm các tập phổ biến tối đại theo ngưỡng minsupp=0,4

Câu2:

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

Lý thuyết dựa trên thuật toán tập phổ biến và luật kết hợp

Bài giải:

-Hoá đơn O={o1, o2, o3, o4, o5} : 5 giao tác hoá đơn

-Mặt hàng {i1, i2, i3, i4, i5} : 5 mặt hàng

Ta có cơ sở dữ liệu nhị phân

	i1	i2	i3	i4	i5
o1	1		1	1	
o2	1		1	1	
o3			1		1
o4				1	1
o5		1	1		1

Câu 1:

Tìm các tập phổ biến tối đại theo ngưỡng minsup=0,4

Ta có: Độ phổ biến của từng mặt hàng :

SP(s)= Số giao tác của S /tổng số giao tác, Với SP(S) thuộc [0,1]

1/ Tập phổ biến 1 mặt hàng: F1=?

Ta có

SP({i1}) =2/5=0.4 = minsupp

SP({i2}) =1/5 (loại).

SP({i3}) =4/5 >0.4

SP({i4}) =3/5 >0.4

SP({i5}) =3/5 >0.4

==>F1={ {i1} , {i3} , {i4} , {i5} }

2/ Tập phổ biến 2 mặt hàng: F2=?

	{i1}	{i3}	{i4}	{i5}
{i1}		{i1,i3}	{i1,i4}	{i1,i5}
{i3}			{i3,i4}	{i3,i5}
{i4}				{i4,i5}

{i5}				

== > CF(F1)={ {i1,i3} ,{i1,i4},{i1,i5},{i3,i4},{i3,i5},{i4,i5} }

Vậy Tính tiếp F2=?

SP({i1,i3}) =2/5=0.4 = minsupp

SP({i1,i4}) =2/5

SP({i1,i5}) =0/5 (loại)

SP({i3,i4}) =2/5

SP({i3,i5}) =2/5

SP({i4,i5}) =1/5 (loại)

==>F2={ {i1,i3} , {i1,i4} , {i3,i4} , {i3,t5} }

3/ Tập phổ biến 3 mặt hàng: F3=?

	{i1,i3}	{i1,i4}	{i3,i4}	{i3,i5}
{i1,i3}		{i1,i3,i4}	{i1,i3,i4}	{i1,i3,i5}
{i1,i4}			{i1,i3,i4}	{i1,i3,i4,i5}
{i3,i4}				{i3,i4,i5}
{i3,i5}				

== > CF(F2)={ {i1,i3,i4}, {i1,i3,i5} , {i3,i4,i5} }

Vậy Tính tiếp F 3=?

SP({i1,i3,i4}) =2/5=0.4 = minsupp

SP({i1,i3,i5}) =0/5 (loại)

SP({i3,i4,i5}) =0/ 5 (loại)

==>F3={ {i1,i3,i4} }

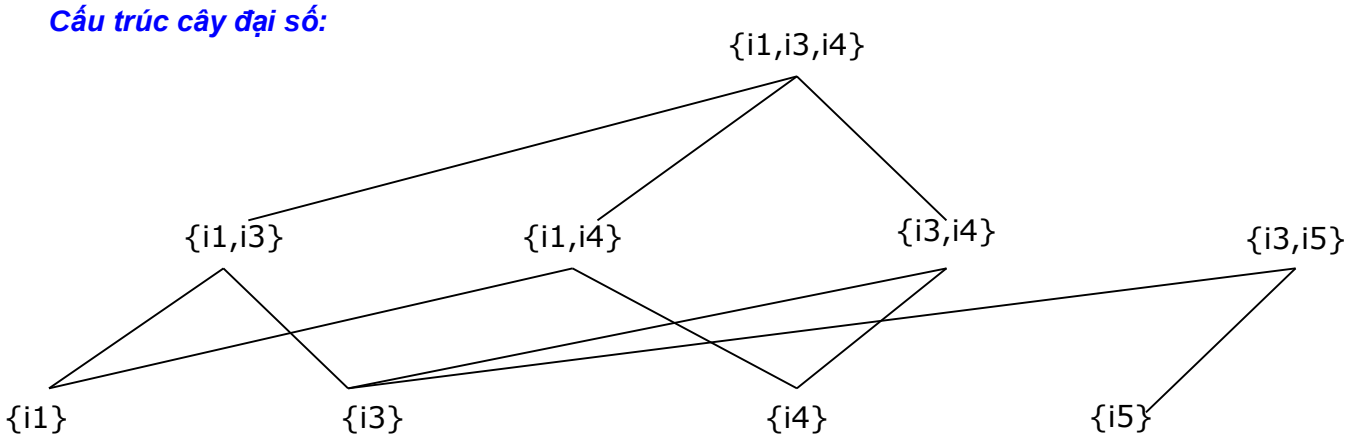
==> CF(F3)= rỗng.

Vậy tập phổ biến :

F=F1 U F2 U F3=

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

Cấu trúc cây đại số:



Kết luận Tập phổ biến tối đại : {i3,i5} , {i1,i3,i4}

Câu2:

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

R : X-> Y LÀ LUẬT KẾT HỢP < == > SP(XUY)>= minsupp và CF(X-> Y)>=minconf

Ta có : CF(X->Y)= SP(X UY)/ SP(X)

Với S1={i1,i3} =>SP(s1)=0.4

R11 : {i1} -> {i3} ==> CF(R11) = SP(S1)/ SP({i1})= 2/5 /2/5=1 > 0 .8

R12 : {i3} -> {i1} ==> CF(R12) = SP(S1)/ SP({i3})= 2/5 /4/5=1/2 < 0 .8 loại

Với S2={i1,i4} SP(s2)=0.4

R21 : {i1} -> {i4} ==> CF(R21) = SP({S2)/ SP({i1})= 2/5 /2/5=1 > 0 .8

R22 : {i4} -> {i1} ==> CF(R22) = SP(S2)/ SP({i4})= 2/5 /3/5=2/3 < 0 .8 loại

Với S3={i3,i4} SP(s3)=0.4

R31 : {i3} -> {i4} ==> CF(R31) = SP(S3)/ SP({i3})= 2/5 /4/5=1/2 < 0 .8 loại

R32 : {i4} -> {i3} ==> CF(R32) = SP(S3)/ SP({i4})= 2/5 /3/5=2/3 < 0 .8 loại

Với S4={i3,i5} SP(s4)=0.4

R41 : {i3} -> {i5} ==> CF(R41) = SP(S4)/ SP({i3})= 2/5 /4/5=1/2 < 0 .8 loại

R42 : {i5} -> {i3} ==> CF(R42) = SP(S4)/ SP({i5})= 2/5 /3/5=2/3 < 0 .8 loại

Với S5={i1,i3,i4} SP(s5)=0.4

R51 : {i1} -> {i3,i4} ==> CF(R51) = SP(S5)/ SP({i1})= 2/5 /2/5=1> 0 .8

R52 : {i3} -> {i1,i4} ==> CF(R52) = SP(S5)/ SP({i5})= 2/5 /4/5=1/2 < 0 .8 loại

R53 : {i4} -> {i1,i3} ==> CF(R53) = 2/5 /3/5=2/3 < 0 .8 loại

R54 : {i3,i4} -> {i1} ==> CF(R54) = 2/5 /2/5=1 > 0 .8

R55 : {i1,i4} -> {i3} ==> CF(R55) = 2/5 /2/5=1 > 0 .8

R56 : {i1,i3} -> {i4} ==> CF(R56) = 2/5 /2/5=1 > 0 .8

Vậy có 6 luật kết hợp R11, R21, R51, R54, R55, R56

TẬP PHỔ BIẾN

Bài tập: **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

Bài làm:

a) Ta có: Độ phổ biến của từng mặt hàng :

SP(s)= Số giao tác của S /tổng số giao tác, Với SP(S) thuộc [0,1]

1/ Tập phổ biến 1 mặt hàng: F1=?

Ta có

SP({i1}) =2/5=0.4 = minsupp

SP({i2}) =1/5 (loại).

SP({i3}) =4/5 >0.4

SP({i4}) =3/5 >0.4

SP({i5}) =3/5 >0.4

==>F1={ {i1} , {i3} , {i4} , {i5} }

2/ Tập phổ biến 2 mặt hàng: F2=?

	{i1}	{i3}	{i4}	{i5}
{i1}		{i1,i3}	{i1,i4}	{i1,i5}
{i3}			{i3,i4}	{i3,i5}
{i4}				{i4,i5}
{i5}				

== > CF(F1)={ {i1,i3} ,{i1,i4},{i1,i5},{i3,i4},{i3,i5},{i4,i5} }

Vậy Tính tiếp F2=?

SP({i1,i3}) =2/5=0.4 = minsupp

SP({i1,i4}) =2/5

SP({i1,i5}) =0/5 (loại)

SP({i3,i4}) =2/5

SP({i3,i5}) =2/5

SP({i4,i5}) =1/5 (loại)

==>F2={ {i1,i3} , {i1,i4} , {i3,i4} , {i3,t5} }

3/ Tập phổ biến 3 mặt hàng: F3=?

	{i1,i3}	{i1,i4}	{i3,i4}	{i3,i5}
{i1,i3}		{i1,i3,i4}	{i1,i3,i4}	{i1,i3,i5}
{i1,i4}			{i1,i3,i4}	{i1,i3,i4,i5}
{i3,i4}				{i3,i4,i5}
{i3,i5}				

== > CF(F2)={ {i1,i3,i4}, {i1,i3,i5} , {i3,i4,i5} }

Vậy Tính tiếp F 3=?

SP({i1,i3,i4}) =2/5=0.4 = minsupp

SP({i1,i3,i5}) =0/5 (loại)

SP({i3,i4,i5}) =0/ 5 (loại)

==>F3={ {i1,i3,i4} }

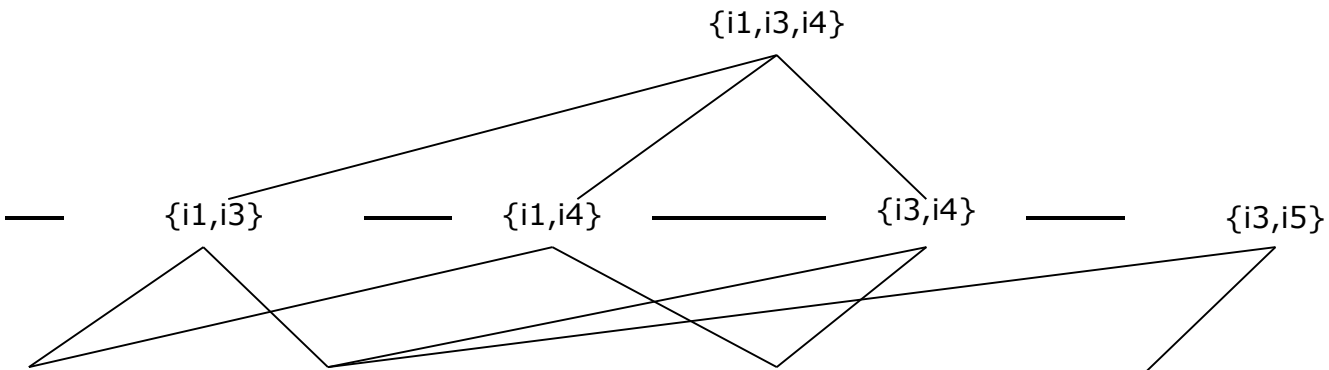
==> CF(F3)= rỗng.

Vậy tập phổ biến :

F=F1 U F2 U F3=

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

Cấu trúc cây đại số:



Kết luận Tập phổ biến tối đại : {i3,i5} , {i1,i3,i4}

Câu 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

R : X-> Y LÀ LUẬT KẾT HỢP < == > SP(XUY)>= minsupp và

CF(X-> Y)>=minconf

Ta có : CF(X->Y)= SP(X UY)/ SP(X)

Với S1={i1,i3} =>SP(s1)=0.4

R11 : {i1} -> {i3} ==> CF(R11) = SP(S1)/ SP({i1})= 2/5 /2/5=1 > 0 .8

R12 : {i3} -> {i1} ==> CF(R12) = SP(S1)/ SP({i3})= 2/5 /4/5=1/2 < 0 .8 loại

Với S2={i1,i4} SP(s2)=0.4

R21 : {i1} -> {i4} ==> CF(R21) = SP({S2)/ SP({i1})= 2/5 /2/5=1 > 0 .8

R22 : {i4} -> {i1} ==> CF(R22) = SP(S2)/ SP({i4})= 2/5 /3/5=2/3 < 0 .8 loại

Với S3={i3,i4} SP(s3)=0.4

R31 : {i3} -> {i4} ==> CF(R31) = SP(S3)/ SP({i3})= 2/5 /4/5=1/2 < 0 .8 loại

R32 : {i4} -> {i3} ==> CF(R32) = SP(S3)/ SP({i4})= 2/5 /3/5=2/3 < 0 .8 loại

Với S4={i3,i5} SP(s4)=0.4

R41 : {i3} -> {i5} ==> CF(R41) = SP(S4)/ SP({i3})= 2/5 /4/5=1/2 < 0 .8 loại

R42 : {i5} -> {i3} ==> CF(R42) = SP(S4)/ SP({i5})= 2/5 /3/5=2/3 < 0 .8 loại

Với S5={i1,i3,i4} SP(s5)=0.4

R51 : {i1} -> {i3,i4} ==> CF(R51) = SP(S5)/ SP({i1})= 2/5 /2/5=1> 0 .8

R52 : {i3} -> {i1,i4} ==> CF(R52) = SP(S5)/ SP({i5})= 2/5 /4/5=1/2 < 0 .8 loại

R53 : {i4} -> {i1,i3} ==> CF(R53) = 2/5 /3/5=2/3 < 0 .8 loại

R54 : {i3,i4} -> {i1} ==> CF(R54) = 2/5 /2/5=1 > 0 .8

R55 : {i1,i4} -> {i3} ==> CF(R55) = 2/5 /2/5=1 > 0 .8

R56 : {i1,i3} -> {i4} ==> CF(R56) = 2/5 /2/5=1 > 0 .8

Vậy có 6 luật kết hợp R11, R21, R51, R54, R55, R56

TẬP THÔ VÀ CÂY QUYẾT ĐỊNH

***Bài tập*: Dùng thuật tóan ID 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

Thuật toán Naïve Bayes

Xác suất (rám)= 3/8

Xác suất (không rám) =5/8

Ước lượng:

Màu tóc	
P(Đen/rám)=2/3	P(Đen/không rám)=2/5
P(Râm/rám)=0/3	P(Râm/không rám)=3/5
P(bạc/ rám)=1/3	***P(bạc/không rám)=0/5***
Chiều cao	
P(Tầm thước/rám)=2/3	P(Tầm thước/không rám)=1/5
P(Cao/rám)=0/3	P(Cao/không rám)=2/5

P(Thấp/rám)=1/3	P(Thấp/không rám)=2/5
Cân nặng	
P(nhẹ/rám)=1/3	P(nhẹ/không rám)=1/5
P(vừa phải/rám)=1/3	P(vừa phải/không rám)=2/5
P(nặng/rám)=1/3	P(nặng/không)=2/5
Dùng thuốc	
P(không/rám)=3/3	P(không/không)=2/5
P(có/rám)=0/3	P(có/không)=3/5

Ta có các luật được rút ra ngẫu nhiên :

-> Rules1: If màutóc=bạc then bị rám

-> Rules2: If màutóc=râm then khôngrám

-> Rules3: If dùngthuốc=có then khôngrám

-> Rules4: If chiềucao=cao then khôngrám

Áp dụng định lý bayes , tính xác suất có điều kiện và lấy tổng các trở ngại:

Các mẫu:

M1= <DDen, tầm thước, nhẹ, không dùng thuốc)

P(M1/rám).P(p)=P(đen/rám).P(tầm thước/rám).P(nhẹ/rám).P(không dùng thuốc/rám).P(rám)= 2/3x 2/3 x1/3x 3/3x3/8 =0.55

P(đen/không rám).P(tầm thước/ không rám).P(nhẹ/không rám).P(không dùng thuốc/không rám).P(không rám)= 2/5 x1/5x1/5x 2/5x 5/8 = 0.004

→ Mẫu M1 được đưa vào rám

M2= <DDen, tầm thước, nặng vừa, không dùng thuốc)

P(M1/rám).P(p)=P(đen/rám).P(tầm thước/rám).P(nặng vừa/rám).P(không dùng thuốc/rám).P(rám)= 2/3 x2/3 x1/3 x3/3 x3/8 =0.55

P(đen/không rám).P(tầm thước/ không rám).P(nặng vừa/không rám).P(không dùng thuốc/không rám).P(không rám)=

=2/5 x1/5 x2/5x2/5x5/8 =0.008

→ Mẫu M2 được đưa vào rám

M3= <DDen, tầm thước, nặng, không dùng thuốc)

P(M1/rám).P(p)=P(đen/rám).P(tầm thước/rám).P(nặng/rám).P(không dùng thuốc/rám).P(rám)=

=2/3x2/3 x1/3x 3/3x3/8 =0.55

P(đen/không rám).P(tầm thước/ không rám).P(nặng /không rám).P(không dùng thuốc/không rám).P(không rám)=

=2/5.1/5.2/5.2/5.5/8 =0.008

→ Mẫu M3 được đưa vào rám

M4= <DDen, thấp, nhẹ, không dùng thuốc)

P(M1/rám).P(p)=P(đen/rám).P(thấp/rám).P(nhẹ/rám).P(không dùng thuốc/rám).P(rám)=

=2/3 x1/3 x1/3 x3/3 x3/8 =0.027

P(đen/không rám).P(thấp/ không rám).P(nhẹ/không rám).P(không dùng thuốc/không rám).P(không rám)=

=2/5.2/5.1/5.2/5.5/8 =0.008

→ Mẫu M4 được đưa vào rám

M5= <DDen, thấp, nặng vừa, không dùng thuốc)

P(M1/rám).P(p)=

P(đen/rám).P(thấp/rám).P(nặng vừa/rám).P(không dùng thuốc/rám).P(rám)=

=2/3.1/3.1/3.3/3.3/8 =0.02777

P(đen/không rám).P(thấp/ không rám).P(nặng vừa /không rám).P(không dùng thuốc/không rám).P(không rám)=

=2/5.2/5.2/5.2/5.5/8 =0.016

→ Mẫu M5 được đưa vào rám

-> Ta rút ra được các luật như sau:

-> Rules1: If màutóc=bạc then bị rám

-> Rules2: If màutóc=râm then khôngrám

-> Rules3: If dùngthuốc=có then khôngrám

-> Rules4: If chiềucao=cao then khôngrám

-> Rules5: If màutóc=đen và chiềucao=tầmthước và khôngdùngthuốc then bị rám

-> Rules6: If màutóc=đen và chiềucao=thấp vàkhôngdùngthuốc then bị rám

GÔM CỤM K MEANS

1. Gom cụm theo k-means

Cho tập điểm

x1={1,3} ={x11,x12}

x2={1.5 , 3.2 }={x21,x22}

x3 ={1.3 ,2.8}={x31,x32}

x4={3, 1}={x41,x42}

Dùng k-means để gom cụm với k = 2

Bước 1 : Khởi tạo ma trận phân hoạch U có 4 cột ứng với 4 điểm và 2 dòng ứng với 2 cụm,

Bước 2: U=(m_{ij}) , $1 \le i \le 2$ và $1 \le j \le 4$

Cho n= 0 (số lần lặp), tạo U0

		x1	x2	x3	x4
U0=	c1	1	0	0	0
	c2	0	1	1	1

Lưu ý mỗi cột chỉ có 01 bit 1

Bước 3: Tính vector trọng tâm:

Do có hai cụm C1,C2 nên có hai vector trọng tâm v1,v2

Các tính vector trọng tâm:

Với vector v1 cho cụm 1:

$$v11 = \frac{m11*x11 + m12*x21 + m13*x31 + m14*x41}{m11 + m12 + m13 + m14}$$

$$= \frac{1*1 + 0*1.5 + 0*1.3 + 0*3}{1 + 0 + 0 + 0} = 1$$

$$v12 = \frac{m11*x12 + m12*x22 + m13*x32 + m14*x42}{m11 + m12 + m13 + m14}$$

$$= \frac{1*3 + 0*3.2 + 0*2.8 + 0*1}{1 + 0 + 0 + 0} = 3$$

Vậy v1 = (1,3)

Với vector v2 cho cụm 2:

$$v21 = \frac{m21*x11 + m22*x21 + m23*x31 + m24*x41}{m21 + m22 + m23 + m24}$$

$$= \frac{0*1 + 1*1.5 + 1*1.3 + 1*3}{0 + 1 + 1 + 1} = \frac{5.8}{3} = 1.93$$

$$v22 = \frac{m21*x12 + m22*x22 + m23*x32 + m24*x42}{m21 + m22 + m23 + m24}$$

$$= \frac{0*3 + 1*3.2 + 1*2.8 + 1*1}{0 + 1 + 1 + 1} = \frac{7}{3} = 2.33$$

Vậy v1 = (1.93,2.33)

Gom các đối tượng vào cụm

a) Tính khoảng cách Euclide từ từng điểm đến cụm c1, c2 chọn cụm có khoảng cách gần nhất để đưa đối tượng vào cụm

$$d(x1, v1) = \sqrt{(x11 - v11)^2 + (x12 - v12)^2} = \sqrt{(1-1)^2 + (3-3)^2} = 0$$

$$d(x1, v2) = \sqrt{(x11 - v21)^2 + (x12 - v22)^2} = \sqrt{(1-1.93)^2 + (3-2.33)^2} = 1.14$$

Gộp x1 vào cụm c1 vì d(x1,v1) < d(x1,v2)

Tính toán tương tự ta có:

d(x2,v1) = 0.54 < d(x2,v2) =0.97 xếp x2 vào cụm c1

d(x3,v1) = 0.36 < d(x3,v2) =0.78 xếp x3 vào cụm c1

d(x4,v1) = 2.83 > d(x4,v2) =1.70 xếp x4 vào cụm c2

Tăng n lên 1

Ma trận phân hoạch Un sẽ là :

		x1	x2	x3	X4
U1=	c1	1	1	1	0
	c2	0	0	0	1

Lặp cho đến khi | Un – Un-1| < epsilon thì dừng , nếu sai thì quay về bước 3.

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

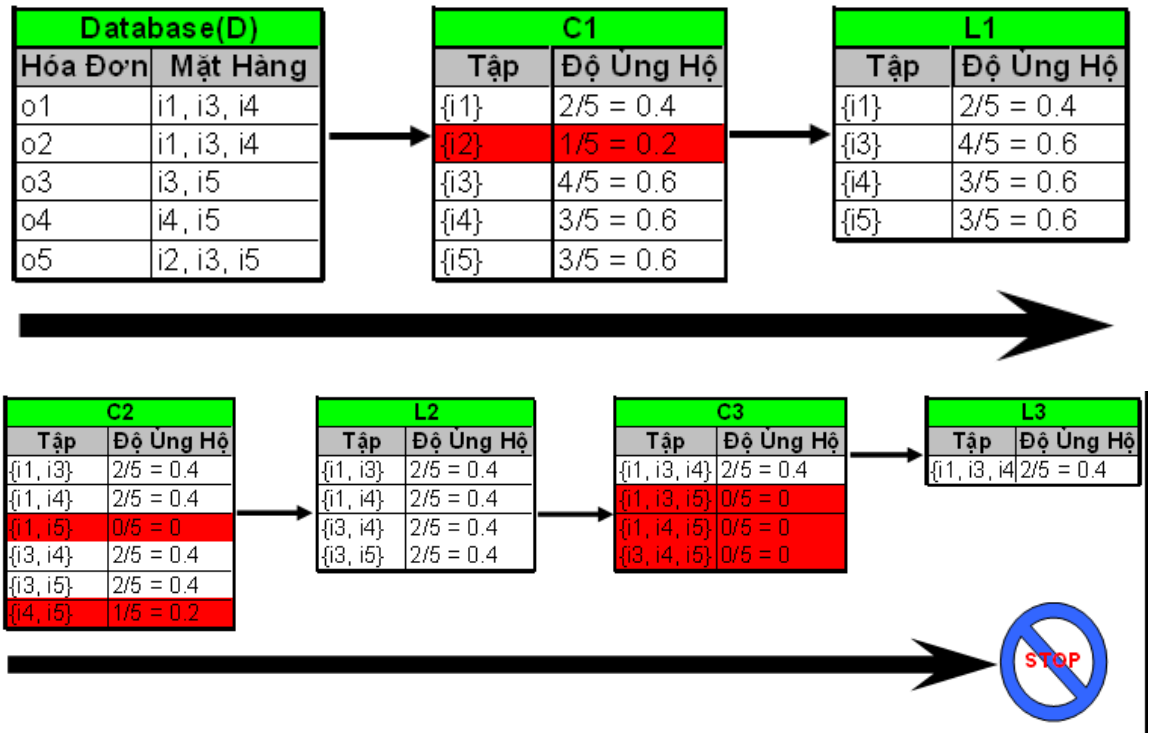
Bạn có suy nghĩ gì về việc dùng luật kết hợp để làm luật phân lớp.

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



Vậy các tập phổ biến thu được là :

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

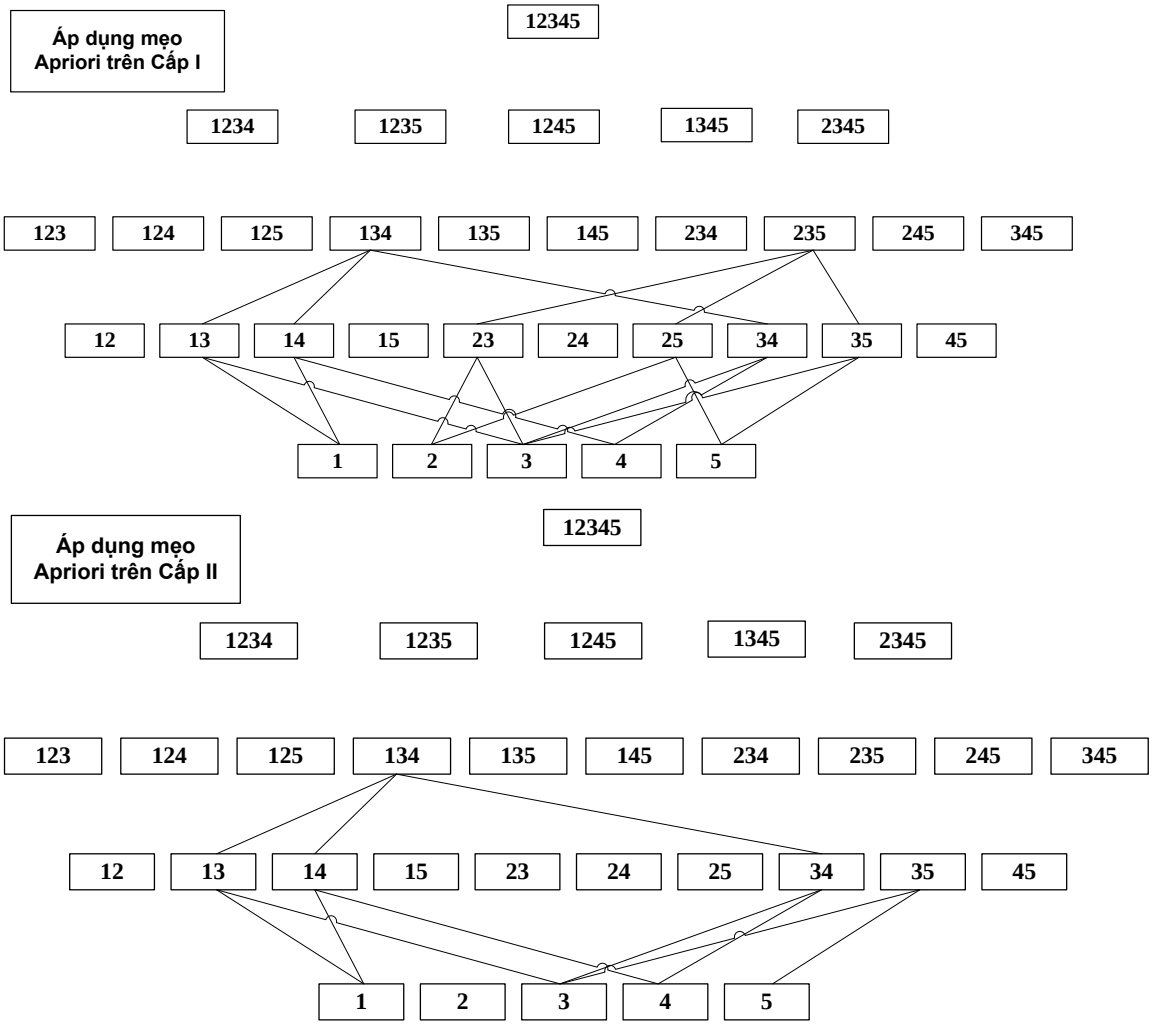
L2 = {{i1, i3}, {i1, i4}, {i3, i4}, {i3, i5}}

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			

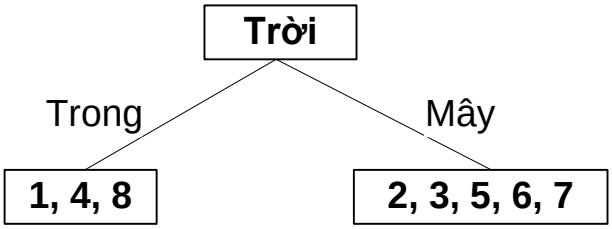
- Tính entropy cho thuộc tính [**Áp Suất**]:

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]:

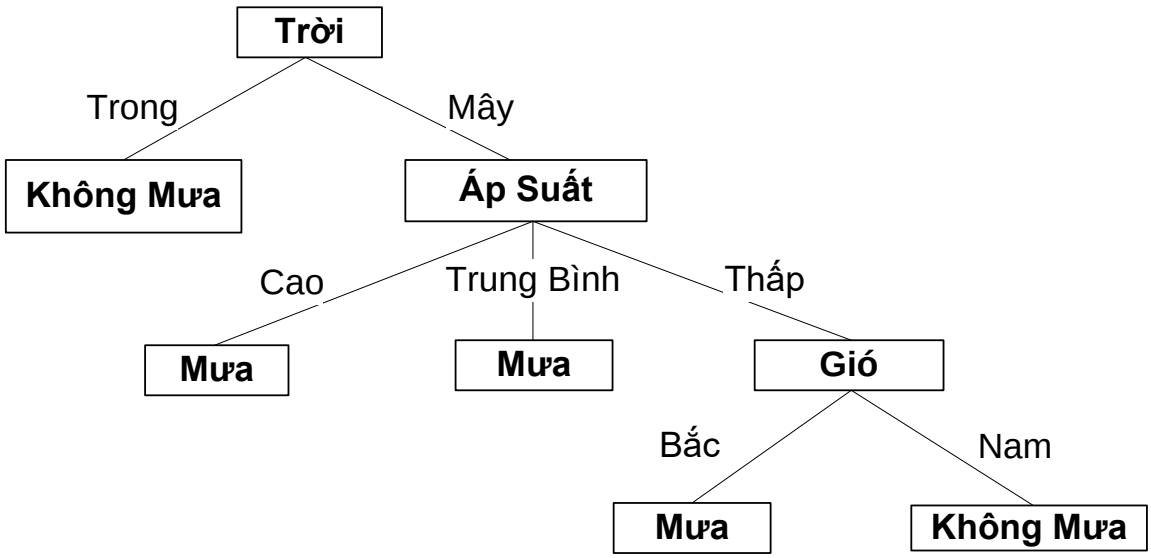
Áp Suất	Pi	ni	I(pi, ni)
Cao	2	0	0.00
Trung Bình	1	0	0.00
Thấp	1	1	1.0

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			

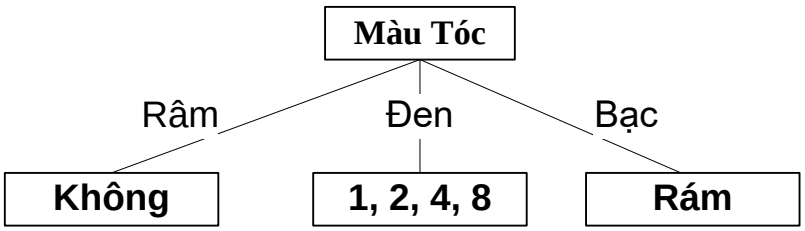
- Tính entropy cho thuộc tính [**Cân Nặng**]:

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**]:

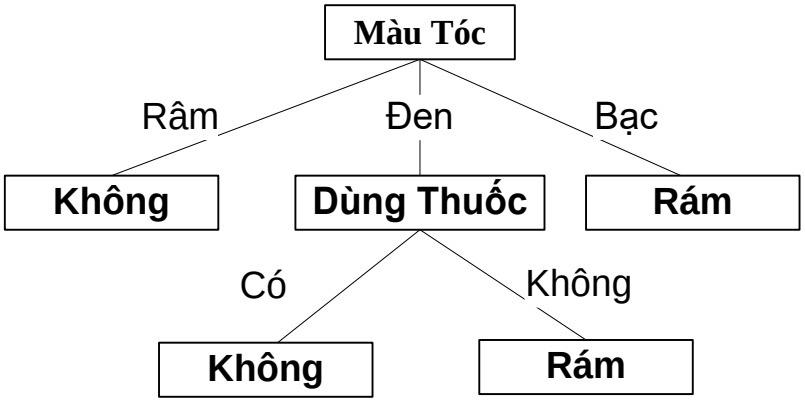
Cân Nặng	**pi**	**Ni**	**I(pi, ni)**
Nặng	0	0	0.00
Vừa phải	1	1	1.00
Nhẹ	1	1	1.00
E (Cân Nặng) = 0/4*0 + 2/4*1 + 2/4*1 = 1			
Gain(Cân Nặng) = I(2,2) - E (Cân Nặng) = 0			

  - 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 toán 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∨N∨T) ∧ (M∨C∨N∨T) ∧ (M∨N∨C) ∧ (M∨N) ∧ (C∨T) ∧ (M∨T) ∧ (N∨T) ∧ (M∧C) ∧ (M) = M ∧ (C∨N∨T) ∧ (C∨T) ∧ (N∨T) = (M∧T) ∨ (M∧C∧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 $\cap$ Z1 = {1} # $\varnothing$ và Z1 $\subseteq$ 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 $\cap$ Z2 = $\varnothing$ => không có luật phân lớp.
  - X1 $\cap$ Z3 = $\varnothing$ => không có luật phân lớp.
  - X1 $\cap$ Z4 = {4} # $\varnothing$ và Z4 $\subseteq$ 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 $\cap$ Z5 = {5} # $\varnothing$ và Z5 $\subseteq$ 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 $\cap$ Z6 = $\varnothing$ => không có luật phân lớp.
  - X1 $\cap$ Z7 = $\varnothing$ => không có luật phân lớp.
  - X1 $\cap$ Z8 = $\varnothing$ => 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 $\cap$ Z1 = $\varnothing$ => không có luật phân lớp.
  - X1 $\cap$ Z2 = {2} # $\varnothing$ và Z2 $\subseteq$ 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 $\cap$ Z3 = {3} # $\varnothing$ và Z3 $\subseteq$ 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 $\cap$ Z4 = $\varnothing$ => không có luật phân lớp.
  - X1 $\cap$ Z5 = $\varnothing$ => 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

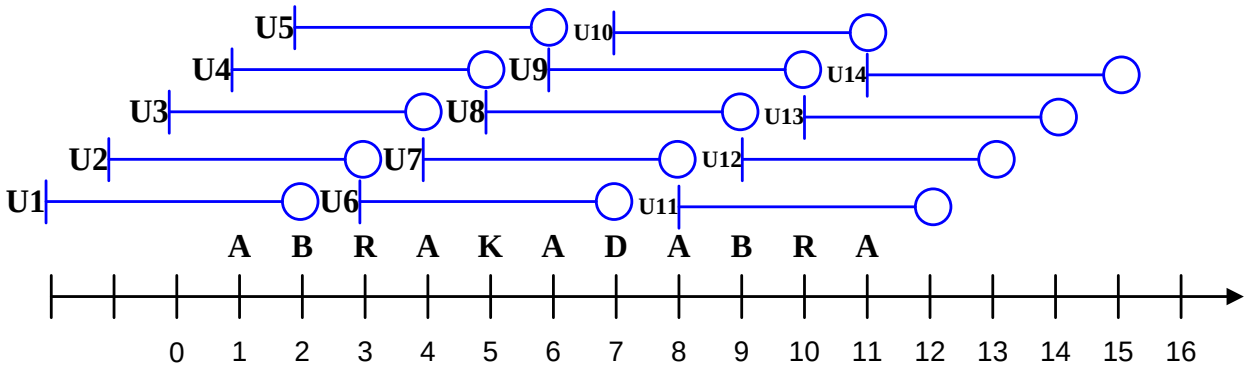
1) Có bao nhiêu cửa sổ có bề rộng là 4 sự kiện được xử lý để tìm các episodes phổ biến theo tiếp cận WINEPI ?

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 }

Bài tập 1 Data Mining

Bài tập 1: môn data mining

3. 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

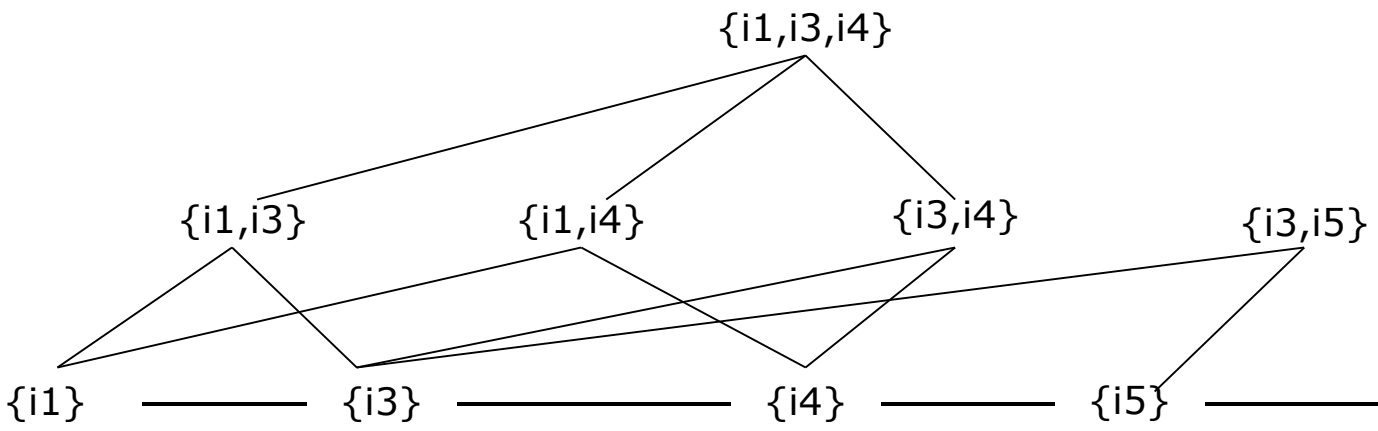
GIẢI

1. Tìm tập phổ biến tối đại:

Scan	Candidates	Large Itemsets
1	{i1}, {i2}, {i3}, {i4}, {i5}	{i1}, {i3}, {i4}, {i5}
2	{i1,i3}, {i1,i4}, {i1,i5}, {i3,i4}, {i3,i5}, {i4,i5)	{i1,i3}, {i1,i4}, {i3,i4}, {i3,i5}
3	{i1,i3,i4}, {i1,i3,i5}, {i3,i4,i5}	{i1,i3,i4}
4	{i1,i3,i4}	{i1,i3,i4}
5	{**Empty**}	{**Empty**}

Ta có tập phổ biến của bài toán là

{{i1}, {i3}, {i4}, {i5}, {i1,i3 } ,{i1,i4 } , {i3,i4 } ,{i3,i5 } , {i1,i3,i4}}



Vậy tập phổ biến tối đại của bài toán đã cho là: {i3,i5}, {i1,i3,i4}

2. Tìm tất cả các AR có MinSupp=0,4 và MinConf = 0,8

R : X-> Y LÀ LUẬT KẾT HỢP ⇔ SP(XUY)>= minsupp và CF(X-> Y)>=minconf

trong đó : ***CF(X->Y)= SP(X UY)/ SP(X)***

Xét S1={i1,i3} → *SP(S1)=0.4*

AR1 : {i1} -> {i3} → CF(R11) = SP(S1)/ SP({i1})= 2/5 /2/5=1 > 0 .8

AR2 : {i3} -> {i1} → CF(R12) = SP(S1)/ SP({i3})= 2/5 /4/5=1/2 < 0 .8 loại

Xét S2={i1,i4} → *SP(S2)=0.4*

AR3 : {i1} -> {i4} → CF(R21) = SP({S2)/ SP({i1})= 2/5 /2/5=1 > 0 .8

AR4 : {i4} -> {i1} → CF(R22) = SP(S2)/ SP({i4})= 2/5 /3/5=2/3 < 0 .8 loại

Xét S3={i3,i4} → *SP(S3)=0.4*

AR4 : {i3} -> {i4} → CF(R31) = SP(S3)/ SP({i3})= 2/5 /4/5=1/2 < 0 .8 loại

AR5 : {i4} -> {i3} → CF(R32) = SP(S3)/ SP({i4})= 2/5 /3/5=2/3 < 0 .8 loại

Xét S4={i3,i5} → *SP(S4)=0.4*

AR6 : {i3} -> {i5} → CF(R41) = SP(S4)/ SP({i3})= 2/5 /4/5=1/2 < 0 .8 loại

AR7 : {i5} -> {i3} → CF(R42) = SP(S4)/ SP({i5})= 2/5 /3/5=2/3 < 0 .8 loại

Với S5={i1,i3,i4} → *SP(S5)=0.4*

AR8 : {i1} -> {i3,i4} → CF(R51) = SP(S5)/ SP({i1})= 2/5 /2/5=1> 0 .8

AR9 : {i3} -> {i1,i4} → CF(R52) = SP(S5)/ SP({i5})= 2/5 /4/5=1/2 < 0 .8 loại

AR10 : {i4} -> {i1,i3} → CF(R53) = 2/5 /3/5=2/3 < 0 .8 loại

AR11 : {i3,i4} -> {i1} → CF(R54) = 2/5 /2/5=1 > 0 .8

AR12 : {i1,i4} -> {i3} → CF(R55) = 2/5 /2/5=1 > 0 .8

AR13 : {i1,i3} -> {i4} → CF(R56) = 2/5 /2/5=1 > 0 .8

Vậy có 6 luật kết hợp **AR1, AR3, AR8, AR11, AR12, AR13**

hay: *{i1} → {i3}*

{i1} → {i4}

{i1} → {i3,i4}

{i3,i4} → {i1}

{i1,i4} → {i3}

{i1,i3} → {i4}

4. 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:

Trời:

Trời	Mưa	Không mưa	I(mưa,không mưa)
Trong	0	3	0
Mây	4	1	0.722

E(Trời)=3/8*0 + 0.722*5/8=0.45

Áp suất:

Áp xuất	Mưa	Không mưa	I(Mưa, không mưa)
Cao	2	2	1
Trung bình	1	0	0
Thấp	1	2	0.9183

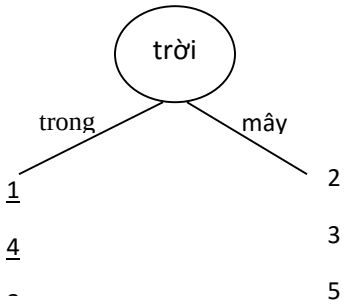
E(Áp suất) = 4/8*1+1/8*0+3/8*0.9183=0.844

Gió:

Gió	Mưa	Không mưa	I(Mưq, khôngmưa)
Bắc	3	2	0.971
Nam	1	2	0.9183

E(Gió) = 5/8*0.917+3/2*0.9183=0.951

Thuộc tính **Trời** có độ **hỗn loạn** nhỏ nhất nên ta chọn thuộc tính này để phân lớp.



Gạch chân: không mưa

Không gạch chân : mưa

+ Dữ liệu sau khi phân hoạch theo thuộc tính trời

Đối tượng	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

Tính Entropy cho các thuộc tính của các đối tượng còn lại:

Áp suất:

Áp suất	Mưa	Không mưa	I(mưa,khôngmưa)
Cao	2	0	0
TrungBình	1	0	0
thấp	1	1	1

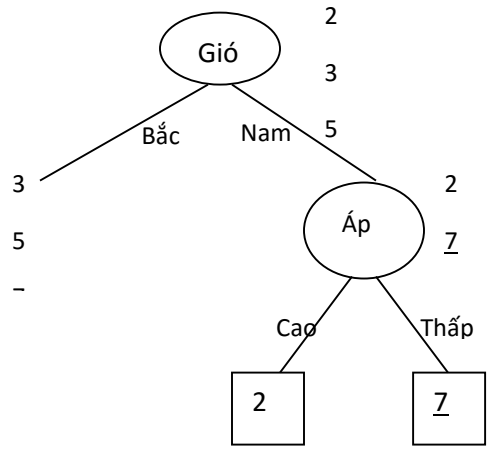
E(Áp suất)=2/5*0+1/5*0+2/5*1=0.4

Gió:

Gió	Mưa	Không mưa	I(mưa,khôngmưa)
Bắc	3	0	0
Nam	1	1	1

E(Gió)=3/5*0+2/5*1=0.4

Độ hỗn loạn của thuộc 2 thuộc tính Gió và Áp suất bằng nhau, nhưng số phân hoạch của thuộc tính gió nhỏ hơn nên ta chọn thuộc tính gió để phân hoạch.



Phân hoạch theo cây trên ta rút được các luật sau:

L1: Nếu **Trời trong** thì **không mưa**

L2:Nếu **trời có mây** và gió có **hướng bắc** thì **mưa**

L3:Nếu **trời có mây**, gió có **hướng Nam** và áp **xuất thấp** thì **không mưa**.

L4:nếu trời **có mây**, gió có **hướng Nam** và **áp xuất cao** thì **mưa**.

Lời giải 2

Bài tập 1: môn data mining

5. 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.

Bài giải:

a. Tìm các tập phổ biến tối đại theo ngưỡng minsupp=0,4

Ta có:

C1		L1
{i1} = 0,4		{i1} = 0,4
{i2} = 0,2 < minsup	⇨	{i3} = 0,4
{i3} = 0,8		{i4} = 0,4
{i4} = 0,6		{i5} = 0,4
{i5} = 0,6		

C2		L2
{i1,i3} = 0,4		{i1,i3} = 0,4
{i1,i4} = 0,4	⇨	{i1,i4} = 0,4
{i1,i5} = 0 < minsup		{i3,i4} = 0,4
{i3,i4} = 0,4		{i3,i5} = 0,4
{i3,i5} = 0,4		

{i4,i5} = 0,2 < minsup

C3		L3
{i1, i3, i4} = 0,4	⇨	{i1, i3, i4} = 0,4

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

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.

Xét tập phổ biến {i3, i5} ta có:

{i3, i5}

supp (i3,i5) / sup (i3) = 0.5 < minconf = 0,8

supp (i3,i5) /sup (i5)= 0,66

=> không tạo được luật thỏa minsup = 0,4 và minconf = 0.8

Xét tập phổ biến {i1, i3, i4} ta có:

{i1, i3}

supp (i1,i3) / sup (i1) = 1

supp (i1,i3) /sup (i3)= 0,5

=> i1 -> i3

{i1,i4}

supp (i1,i4) / sup (i1) = 1

supp (i1,i4) / sup (i4) = 0,66

=>i1->i4

{i3,i4}

supp (i3,i4) / sup (i3) = 0,5

supp (i3,i4) / sup (i4) = 0,66

=> không tạo được luật thỏa minsup = 0,4 và minconf = 0.8

{i1,i3,i4}

supp (i1,i3,i4) / supp(i1) = 1

supp (i1,i3,i4) / supp(i3) = 0,5

supp (i1,i3,i4) / supp(i4) = 0,66

supp (i1,i3,i4) / supp(i1,i3) = 1

supp (i1,i3,i4) / supp(i1,i4) = 1

supp (i1,i3,i4) / supp(i3,i4) = 1

=>i1,i3 -> i4

=>i1,i4 -> i3

=>i1 -> i3, i4

=>i3,i4 -> i1

Các luật kết hợp tìm được là:

i1->i3

i1->i4

i1 -> i3, i4

i1,i3 -> i4

i1,i4 -> i3

i3,i4 -> i1

2. Sử dụng cây định danh để tìm các luật phân lớp từ bảng quyết định sau đây:

#	TRỜI	ÁPSUẤT	GIÓ	KẾTQỦA
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

Bài giải:

Ta có:

Trời

P(trong|không mưa) = 1 P(trong|mưa) = 0

P(mây|không mưa) = 1/5 P(mây|mưa) = 4/5

Áp Suất

P(thấp|không mưa) = 2/3 P(thấp|mưa) = 1/3

P(TB|không mưa) = 0 P(TB|mưa) = 1

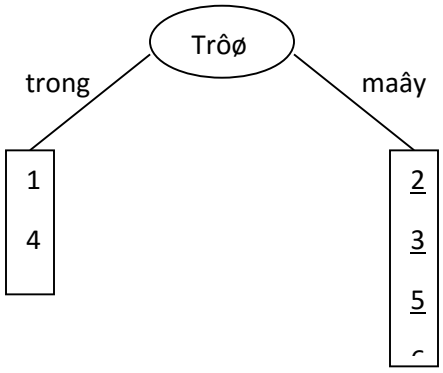
P(cao | không mưa) = 2/4 P(cao | mưa) = 2/4

Gió

P(nam | không mưa) = 2/3 P(nam | mưa)=1/3

P(bắc | không mưa) = 2/5 P(bắc | mưa)=3/5

Ta chọn thuộc tính Trời để phân hoạch:



Trong phân hoạch trời mây còn các lẫn lộn giữa mưa và không mưa. Ta có tập dữ liệu sau:

#	ÁPSUẤT	GIÓ	KẾTQỦA
2	**Cao**	**Nam**	**Mưa**
3	**Trung bình**	**Bắc**	**Mưa**
5	**Thấp**	**Bắc**	**Mưa**
6	**Cao**	**Bắc**	**Mưa**
7	**Thấp**	**Nam**	**Không mưa**

Ta có:

Áp Suất

P(thấp | không mưa) = 1/2 P(thấp | mưa) = 1/2

P(TB | không mưa) = 0 P(TB | mưa) = 1

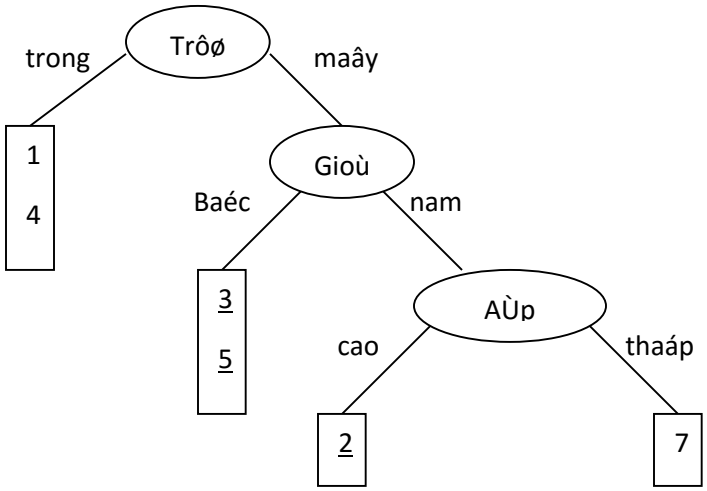
P(cao | không mưa) = 0 P(cao | mưa) = 1

Gió

P(nam | không mưa) = 1/2 P(nam | mưa)=1/2

P(bắc | không mưa) = 0 P(bắc | mưa)=1

Ta chọn thuộc tính Gió để phân hoạch vì Gió có thuộc tính phân họach ít hơn:



Ta có các luật phân lớp từ cây quyết định là:

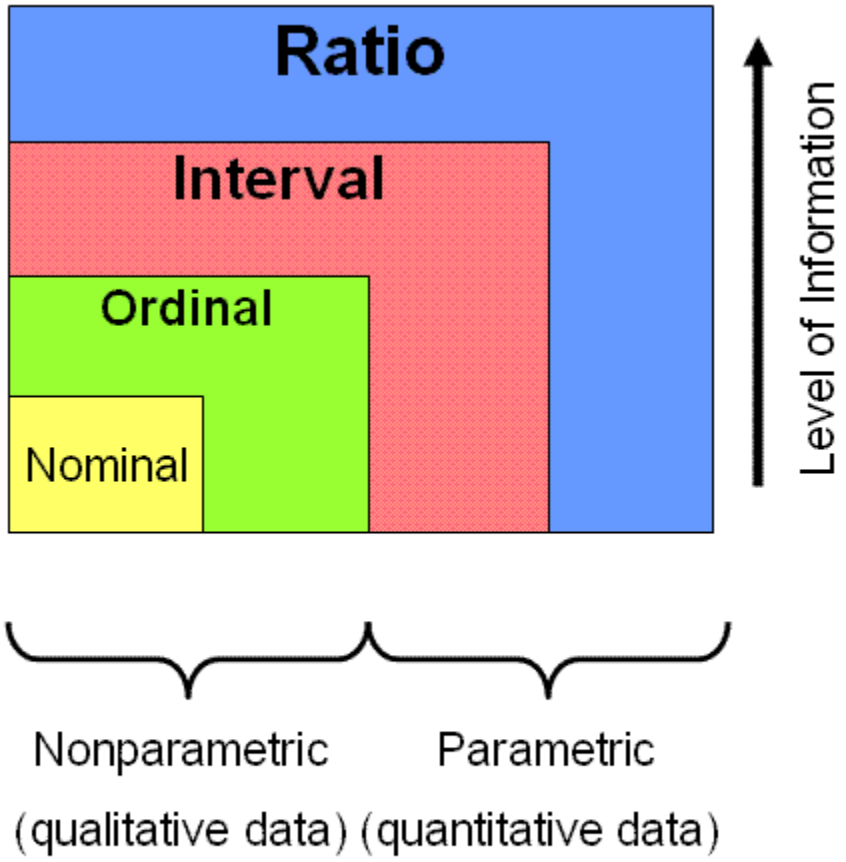
L1: Nếu trời**=trong** thì kết quả=**không mưa**

L2:Nếu trời**=mây** và gió=b**ắc** thì kết quả=**mưa**

L3:Nếu trời**=mây** và gio=**nam** và áp **suất=thấp** thì kết quả=**mưa**.

L4: Nếu trời=**mây** và gió=n**am** và áp s**uất=cao** thì kết quả=**mưa**

HÌNH ẢNH KIỂU DỮ LIỆU LIÊN TỤC VÀ RỜI RẠC



	Continuous	Discrete		
	Quantitative data	Qualitative / Categorical / Attribute data		
Measurement	Units (example)	Ordinal (example)	Nominal (example)	Binary (example)
Time of day	Hours, minutes, seconds	1, 2, 3, etc.	N/A	a.m./p.m.
Date	Month, date, year	Jan., Feb., Mar., etc.	N/A	Before / After
Cycle time	Hours, minutes, seconds, month, date, year	10, 20, 30, etc.	N/A	Before / After
Speed	Miles per hour/centimeters per second	10, 20, 30, etc.	N/A	Fast / Slow
Brightness	Lumens	Light, medium, dark	N/A	On / Off
Temperature	Degrees C or F	10, 20, 30, etc.	N/A	Hot / Cold
<Count data>	Number of things	10, 20, 30, etc.	N/A	Large / Small
Test scores	Percent, number correct	F, D, C, B, A	N/A	Pass / Fail
Defects	N/A	Number of cracks	N/A	Good / Bad
Defects	N/A	N/A	Oversized, missing	Good / Bad
Color	N/A	N/A	Red, blue, green	N/A
Location	N/A	N/A	East, West, South	Domestic / International
Groups	N/A	N/A	HR, Legal, IT	Exempt / Non-exempt
Anything	Percent	10, 20, 30, etc.	N/A	Above / Below

Phân lớp (Classfication)

Ứng dụng lý thuyết Bayes trong phân lớp (Using Bayes Theorem for Classification)

Nguyễn Văn Chức – chucnv@ud.edu.vn

1. Giới thiệu Bayes Theorem

Trong lĩnh vực Data Mining, Bayes Theorem (hay Bayes' Rule) là kỹ thuật phân lớp dựa vào việc tính xác suất có điều kiện. Bayes' Rule được ứng dụng rất rộng rãi bởi tính dễ hiểu và dễ triển khai.

Bayes' Rule (CT1):

$$P(h|D) = P(h) \cdot \frac{P(D|h)}{P(D)}$$

Trong đó:

D : Data

h : Hypothesis (giả thuyết)

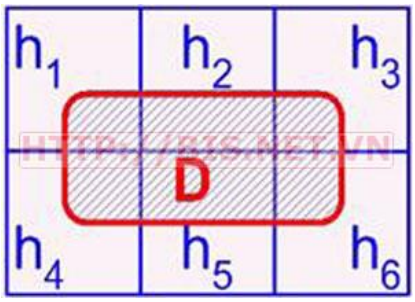
P(h) : Xác suất giả thuyết h (tri thức có được về giả thuyết h trước khi có dữ liệu D) và gọi là ***prior*** probability của giả thuyết h.

P(D| h): Xác suất có điều kiện D khi biết giả thuyết h (gọi là ***likelihood*** probability).

P(D): xác suất của dữ liệu quan sát D không quan tâm đến bất kỳ giả thuyết h nào.(gọi là**prior** probability của dữ liệu D**)**

Tỷ số $\frac{P(D|h)}{P(D)}$: Chỉ số liên quan (*irrelevance index) dùng để đo lường sự liên quan giữa 2 biến A và B. Nếu irrelevance index =1*, có nghĩa A và B không liên quan nhau.

P(h|D) :Xác suất có điều kiện h khi biết D (gọi là ***posterior*** probability của giả thuyết h)



Trong rất nhiều ứng dụng, các giả thuyết h_i có thể loại trừ nhau và vì dữ liệu quan sát D là tập con của tập giả thuyết cho nên chúng ta có thể phân rã P(D) như sau (CT2):

$$P(D) = P(D \cap h_1) \cup P(D \cap h_2) \cup \ldots \cup P(D \cap h_k) = \bigcup_j P(D \cap h_j)$$

Vì $P(D \cap h_j) = P(D \mid h_j) \cdot P(h_j)$ nên (CT1) có thể viết lại như sau (CT3)

$$P(D) = \sum_j P(D \mid h_j) \cdot P(h_j)$$

Thay P(D) trong (CT2) vào (CT1) ta được (CT4)

$$P(h_i \mid D) = \frac{P(D \mid h_i) \cdot P(h_i)}{\sum_j P(D \mid h_j) \cdot P(h_j)}$$

(CT4) gọi là Bayes's Theorem

2. Ứng dụng Bayes Theorem trong phân lớp dữ liệu (Naïve Bayes Classifier)

Các ví dụ sau đây minh họa việc sử dụng Bayes Theorem trong việc phân lớp dữ liệu. Bộ phân lớp dữ liệu dựa trên Bayes theorem còn gọi là **Naïve Bayes Classifier**.

Ví dụ 1: Có training data về thời tiết như sau (xem mô tả chi tiết về dữ liệu weather trong bài *Cây quyết định (Decision Tree) tại http://bis.net.vn/forums/t/378.aspx)*

	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Overcast	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	No
7	Overcast	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Overcast	Mild	High	TRUE	Yes
13	Overcast	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Sử dụng **Naïve Bayes Classifier** để xác định khả năng đến chới thể thao (Play = "yes" hay "no") với thời tiết của ngày quan sát được như sau:

Outlook	Temp	Humidity	Windy	Play
Sunny	Cool	High	True	?

Từ Training data ta có dữ liệu như sau:

Outlook			Temp			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Vì thuộc tính phân lớp Play chỉ có 2 giá trị là "yes" (nghĩa là có đến chơi thể thao) và "no"(không đến chơi thể thao) nên ta phải tính Pr(yes|E) và Pr(no|E) như sau. Trong đó E là dữ liệu cần phân lớp (dự đoán)

Outlook	Temp	Humidity	Windy	Play
Sunny	Cool	High	True	?

← *Evidence E*

$$\Pr[yes \mid E] = \Pr[Outlook = Sunny \mid yes] \times \Pr[Temperature = Cool \mid yes] \times \Pr[Humidity = High \mid yes] \times \Pr[Windy = True \mid yes] \times \frac{\Pr[yes]}{\Pr[E]}$$

Probability of class "yes"

$$= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr[E]}$$

Outlook	Temp	Humidity	Windy	Play
Sunny	Cool	High	True	?

← *Evidence E*

$$\Pr[No \mid E] = \Pr[Outlook = Sunny \mid No] \times \Pr[Temperature = Cool \mid No] \times \Pr[Humidity = High \mid No] \times \Pr[Windy = True \mid No] \times \frac{\Pr[No]}{\Pr[E]}$$

Probability of class "No"

$$= \frac{\frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14}}{\Pr[E]}$$

Likelihood of the two classes
For "yes" = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053
For "no" = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206
Conversion into a probability by normalization:
P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205
P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

Vì P("no") > P("yes") nên kết quả dự đoán Play ="no"

Outlook	Temp	Humidity	Windy	Play
Sunny	Cool	High	True	no ← Result

Ví dụ 2: Có Training Data và Unseen data như sau

	age	income	student	credit_rating	buys_computer	
1	age	income	student	credit_rating	buys_computer	
2	<=30	high	no	fair	no	
3	<=30	high	no	excellent	no	
4	31…40	high	no	fair	yes	
5	>40	medium	no	fair	yes	
6	>40	low	yes	fair	yes	← Training Data
7	>40	low	yes	excellent	no	
8	31…40	low	yes	excellent	yes	
9	<=30	medium	no	fair	no	
10	<=30	low	yes	fair	yes	
11	>40	medium	yes	fair	yes	
12	<=30	medium	yes	excellent	yes	
13	31…40	medium	no	excellent	yes	
14	31…40	high	yes	fair	yes	
15	>40	medium	no	excellent	no	
16						
17	age	income	student	credit_rating	buys_computer	Unseen Data
18	<=30	medium	yes	fair	?	← Predictive Class
19						Buys_computer = yes or no

Sử dụng ***Naïve Bayes Classifier*** để phân lớp cho Unseen data (X)

Class: C1:buys_computer =”yes”, C2:buys_computer =”no”

Tính P(X|Ci) cho mỗi class

X=(age<=30, income =”medium”, student=”yes”,credit_rating=”fair”)

P(age=“<=30” | buys_computer=“yes”) = 2/9=0.222

P(age=“<=30” | buys_computer=“no”) = 3/5 =0.6

P(income=“medium” | buys_computer=“yes”)= 4/9 =0.444

P(income=“medium” | buys_computer=“no”) = 2/5 = 0.4

P(student=“yes” | buys_computer=“yes”)= 6/9 =0.667

P(student=“yes” | buys_computer=“no”)= 1/5=0.2

P(credit_rating="fair" | buys_computer="yes")=6/9=0.667

P(credit_rating="fair" | buys_computer="no")=2/5=0.4

Tính P(X|Ci) :

P(X|buys_computer="yes")= 0.222 x 0.444 x 0.667 x 0.667 =0.044

P(X|buys_computer="no")= 0.6 x 0.4 x 0.2 x 0.4 =0.019

P(X|Ci)*P(Ci):

P(X|buys_computer="yes") * P(buys_computer="yes")=0.044*9/14= **0.028**

P(X|buys_computer="no") * P(buys_computer="no")=0.019*5/14 = 0.007

Do đó ta có X thuộc lớp **buys_computer="yes"**

age	income	student	credit_rating	buys_computer
<=30	medium	yes	fair	yes

Unseen Data
Predictive Class
Buys_computer = yes

Bayes Thoerem *được triển khai rất rộng rãi trong Data Mining vì dễ hiểu, dễ triển khai . Tuy nhiên, Bayes Theorem giả thiết rằng các biến trong mô hình là độc lập nhau. Nếu các biến không độc lập nhau thì Bayes Theorem cho kết quả thiếu chính xác.*

Dùng thuật tóan ID3 và Naïve Bayes để tìm 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:

1. Thuật toán ID3

Bước 1:

Các thuộc tính và miền giá trị tương ứng baogồm:

- Thuộc tính *Màu tóc* có miền giá trị {Đen, Râm, Bạc}
- Thuộc tính *Chiều cao* có miền giá trị {Cao, Tầm thước, Thấp}
- Thuộc tính *Cân nặng* có miền giá trị {Nặng, Vừa phải, Nhẹ}
- Thuộc tính *Dùng thuốc* có miền giá trị {Có, Không}
- Thuộc tính Lớp có miền giá trị {P, N} (P ứng với không bị rám và N là ngược lại)

Khối lượng thông tin cần thiết để quyết định một mẫu tùy ý có thuộc về lớp P hay N hay không là:

$$I(p,n) = -(p/(p+n))*\log_2(p/(p+n))-(n/(p+n))*\log_2(n/(p+n))$$

$$I(5,3) = -(5/8)* \log_2(5/8)-(3/8)* \log_2(3/8) = 0,954$$

Tính Entropy cho thuộc tính *Màu tóc*

Màu tóc	p_i	n_i	$I(p_i,n_i)$
Đen	2	2	1
Râm	3	0	0
Bạc	0	1	0

Ta có:

$$E(\text{Màu tóc}) = (4/8)*I(2,2) + (3/8)*I(3,0) + (1/8)*I(0,1) = 0,5$$

Do đó:

$$\text{Gain}(\text{Màu tóc}) = I(5,3) – E(\text{Màu tóc}) = 0,954 – 0,5 = 0,454$$

Tương tự:

Tính Entropy cho thuộc tính *Chiều cao*

Chiều cao	p_i	n_i	$I(p_i,n_i)$
Cao	2	0	0
Tầm thước	1	2	0,918
Thấp	2	1	0,918

E(Chiều cao) = (2/8)*I(2,0) + (3/8)*I(1,2) + (3/8)*I(2,1) = 0,689

Gain(Chiều cao) = 0,954 – 0,689 = 0,265

Tính Entropy cho thuộc tính *Cân nặng*

Cân nặng	p_i	n_i	$I(p_i,n_i)$
Nặng	2	1	0,918
Vừa phải	2	1	0,918
Nhẹ	1	1	1

E(Cân nặng) = (3/8)*I(2,1) + (3/8)*I(2,1) + (2/8)*I(1,1) = 0,939

Gain(Cân nặng) = 0,954 – 0,939 = 0,015

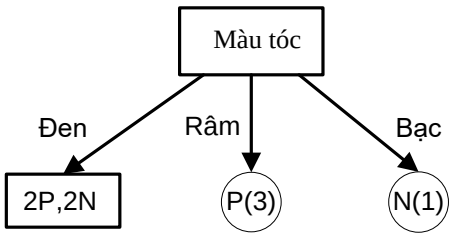
Tính Entropy cho thuộc tính *Dùng thuốc*

Dùng thuốc	p_i	n_i	$I(p_i,n_i)$
Có	3	0	0
Không	2	3	0,970

E(Dùng thuốc) = (3/8)*I(3,0) + (5/8)*I(2,3) = 0,607

Gain(Dùng thuốc) = 0,954 – 0,607 = 0,347

Chọn thuộc tính có độ lợi thông tin lớn nhất là thuộc tính "Màu tóc", ta có cây có dạng:



Bước 2:

Trong cây này ta thấy ứng với màu tóc đen còn 2 phần tử có trị P và 2 phần tử có trị N. Tiếp tục áp dụng ID3 cho nút con này cho đến khi đạt đến nút lá hoặc nút có entropy=0. Ta có tập dữ liệu (con) ứng với màu tóc đen như sau:

Chiều cao	Cân nặng	Dùng thuốc?	Kết quả
Tầm thước	Nhẹ	Không	Bị rám
Cao	Vừa phải	Có	Không
Thấp	Vừa phải	Không	Bị rám
Thấp	Nhẹ	Có	Không

Các thuộc tính và miền giá trị tương ứng baogồm:

- Thuộc tính *Chiều cao* có miền giá trị {Cao, Tầm thước, Thấp}
- Thuộc tính *Cân nặng* có miền giá trị {Vừa phải, Nhẹ}
- Thuộc tính *Dùng thuốc* có miền giá trị {Có, Không}
- Thuộc tính Lớp có miền giá trị {P, N} (P ứng với không bị rám và N là ngược lại)

Khối lượng thông tin cần thiết để quyết định một mẫu tùy ý có thuộc về lớp P hay N hay không là:

$$I(p,n) = -(p/(p+n))*\log_2(p/(p+n))-(n/(p+n))*\log_2(n/(p+n))$$

$$I(2,2) = 1$$

Tính Entropy cho thuộc tính *Chiều cao*

Chiều cao	p_i	n_i	$I(p_i,n_i)$
Cao	1	0	0
Tầm thước	0	1	0
Thấp	1	1	1

$$E(\text{Chiều cao}) = (1/4)*I(1,0) + (1/4)*I(0,1) + (2/4)*I(1,1) = 0,5$$

$$\text{Gain(Chiều cao)} = 1 – 0,5 = 0,5$$

Tính Entropy cho thuộc tính *Cân nặng*

Cân nặng	p_i	n_i	$I(p_i,n_i)$
Vừa phải	1	1	1
Nhẹ	1	1	1

E(Cân nặng) = (2/4)*I(1,1) + (2/4)*I(1,1) = 1

Gain(Cân nặng) = 0,954 – 1 = -0,046

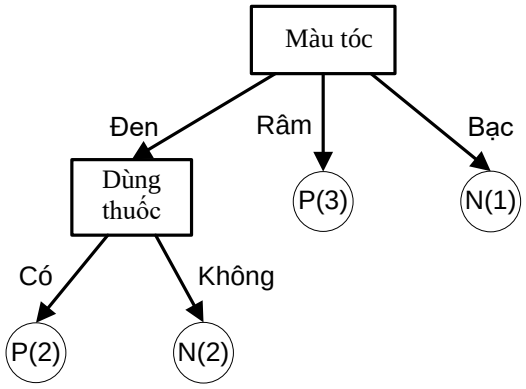
Tính Entropy cho thuộc tính *Dùng thuốc*

Dùng thuốc	p_i	n_i	$I(p_i,n_i)$
Có	2	0	0
Không	0	2	0

E(Dùng thuốc) = (2/4)*I(2,0) + (2/8)*I(0,2) = 0

Gain(Dùng thuốc) = 0,954 – 0 = 0,954

Như vậy thuộc tính "Dùng thuốc" có độ lợi thông tin lớn nhất được dùng để phân lớp, ta có cây quyết định do thuật toán ID3 tạo ra như sau:



Như vậy các luật được tạo ra như sau:

IF (Màu tóc = ***Râm***) THEN ***Không rám nắng***

ELSE IF Màu tóc = ***Đen*** AND Dùng thuốc = ***Có*** THEN ***Không rám nắng***

ELSE IF (Màu tóc = ***Bạc***) THEN ***Rám nắng***

ELSE ***Rám nắng***

Hay rút gọn luật như sau:

IF (Màu tóc = ***Râm***) OR (Màu tóc = ***Đen*** AND Dùng thuốc = ***Có***) THEN ***Không rám nắng***

ELSE ***Rám nắng***

2. ***Thuật toán Naïve Bayes***

Dựa vào bảng dữ liệu ta có thể tính các xác suất sau:

- Xác suất lớp dương (không rám nắng): P(p) = 5/8

- Xác suất lớp âm (rám nắng): P(n) = 3/8

Màu tóc	
P(Đen | p) = 2/5	P(Đen | n) = 2/3
P(Râm | p) = 3/5	P(Râm | n) = 0
P(Bạc | p) = 0	P(Bạc | n) = 1/3
Chiều cao	
P(Cao | p) = 2/5	P(Cao | n) = 0
P(Tầm thước | p) = 1/5	P(Tầm thước | n) = 2/3
P(Thấp | p) = 2/5	P(Thấp | n) = 1/3
Cân nặng	
P(Nặng | p) = 2/5	P(Nặng | n) = 1/3
P(Vừa phải | p) = 2/5	P(Vừa phải | n) = 1/3
P(Nhẹ | p) = 1/5	P(Nhẹ | n) = 1/3
Dùng thuốc	
P(Có | p) = 3/5	P(Có | n) = 0
P(Không | p) = 2/5	P(Không | n) = 3/3

- Xét một mẫu X có màu tóc Râm (không quan tâm các thuộc tính khác)
- P(Râm | n) = 0 => P(X|n)*P(n) = 0
- $P(x_i | p) > 0$ => P(X | p) > 0
  Suy ra mẫu X thuộc lớp P (không rám nắng)

- Xét một mẫu X có màu tóc Đen và có Dùng thuốc (không quan tâm các thuộc tính còn lại)
- P(Có | n) = 0 => P(X|n)*P(n) = 0
- $P(x_i | p) > 0$ => P(X | p) > 0
  Suy ra mẫu X thuộc lớp P (không rám nắng)

Hai phân lớp trên phù hợp với luật được suy ra từ giải thuật ID3. Tuy nhiên, xét mẫu X = <Bạc,Cao,Vừa phải,Có>, ta có:

- P(Bạc | p) = 0 => P(X|p)*P(p) = 0
- P(Cao | n) = 0 => P(X|n)*P(n) = 0
  => Không thể xác định X thuộc lớp nào!

Kết hợp (association rules)

Một số ví dụ về "luật kết hợp" (associate rule)

- "98% khách hàng mà mua tạp chí thể thao thì đều mua các tạp chí về ôtô" ⇨ sự kết hợp giữa "tạp chí thể thao" với "tạp chí về ôtô"
- "60% khách hàng mà mua bia tại siêu thị thì đều mua bỉm trẻ em" ⇨ sự kết hợp giữa "bia" với "bỉm trẻ em"
- "Có tới 70% người truy nhập Web vào địa chỉ Url 1 thì cũng vào địa chỉ Url 2 trong một phiên truy nhập web" ⇨ sự kết hợp giữa "Url 1" với "Url 2". Khai phá dữ liệu sử dụng Web (Dữ liệu từ file log của các site, chẳng hạn được MS cung cấp).
- Các Url có gắn với nhãn "lớp" là các đặc trưng thì có luật kết hợp liên quan giữa các lớp Url này.

Thuật toán Apriori khai phá luật kết hợp

1. Luật kết hợp trong khai phá dữ liệu (Association Rule in Data Mining)

Trong lĩnh vực Data Mining, mục đích của luật kết hợp (Association Rule - AR) là tìm ra các mối quan hệ giữa các đối tượng trong khối lượng lớn dữ liệu. Nội dung cơ bản của luật kết hợp được tóm tắt như dưới đây.

Cho cơ sở dữ liệu gồm các giao dịch T là tập các giao dịch t1, t2, …, tn.

T = {t1, t2, …, tn}. T gọi là cơ sở dữ liệu giao dịch (Transaction Database)

Mỗi giao dịch ti bao gồm tập các đối tượng I (gọi là itemset)

I = {i1, i2, …, im}. *Một itemset gồm k items gọi là k-itemset*

Mục đích của luật kết hợp là tìm ra sự kết hợp (association) hay tương quan (correlation) giữa các items. Những luật kết hợp này có dạng ***X =>Y***

Trong Basket Analysis, luật kết hợp X =>Y có thể hiểu rằng những người mua các mặt hàng trong tập X cũng thường mua các mặt hàng trong tập Y. (X và Y gọi là itemset).

Ví dụ, nếu X = {Apple, Banana} và Y = {Cherry, Durian} và ta có luật kết hợp X =>Y thì chúng ta có thể nói rằng những người mua Apple và Banana thì cũng thường mua Cherry và Durian.

Theo quan điểm thống kê, X được xem là biến độc lập (Independent variable) còn Y được xem là biến phụ thuộc (Dependent variable)

Độ hỗ trợ (Support) và độ tin cây (Confidence) là 2 tham số dùng để đo lường luật kết hợp.

Độ hỗ trợ (Support) của luật kết hợp X =>Y là tần suất của giao dịch chứa tất cả các items trong cả hai tập X và Y. Ví dụ*, support của luật* X =>Y là 5% có nghĩa là *5% các giao dịch X và Y được mua cùng nhau.*

Công thức để tính support của luật X =>Y như sau:

$$support\ (X \to Y) = P(X \cup Y) = \frac{n(X \cup Y)}{N}$$

Trong đó: N là tổng số giao dịch.

Độ tin cậy (*Confidence) của luật kết hợp* X =>Y là xác suất xảy ra Y khi đã biết X. Ví dụ độ tin cậy của luật kết hợp {Apple} =>Banana} là *80% có nghĩa là 80% khách hàng mua Apple cũng mua Banana.*

Công thức để tính độ tin cậy của luật kết hợp X =>là xác suất có điều kiện Y khi đã biết X như sau :

$$confidence\ (X \to Y) = P(Y|X) = \frac{n(X \cup Y)}{n(X)}$$

Trong đó: n(X) là số giao dịch chứa X

Để thu được các luật kết hợp, ta thường áp dụng 2 tiêu chí: *minimum support (min_sup) và minimum confidence (min_conf)*

Các luật thỏa mãn có support và confidence thỏa mãn (lớn hơn hoặc bằng) cả Minimum support và Minimum confidence gọi là các luật mạnh (Strong Rle)

Minimum support và Minimum confidence gọi là các giá trị ngưỡng (threshold) và phải xác định trước khi sinh các luật kết hợp.

Một itemsets mà tần suất xuất hiện của nó >= min_sup goi là frequent itemsets

Một số loại luật kết hợp

Binary association rules (luật kết hợp nhị phân): Apple => Banana

Quantitative association rules (luật kết hợp định lượng):

weight in [70kg – 90kg] => height in [170cm – 190cm]

Fuzzy association rules (Luật kết hợp mờ): weight in HEAVY => height in TALL

Thuật toán phổ biến nhất tìm các luật kết hợp là Apriori sử dụng Binary association rules.

2.Thuật toán sinh các luật kết hợp Apriori *(by Agrawal and Srikant 1994)*

Tư tưởng chính của thuật toán Apriori là:

- Tìm tất cả frequent itemsets:

k-itemset (itemsets gồm k items) được dùng để tìm (k+1)- itemset.

Đầu tiên tìm 1-itemset (ký hiệu L_1). L_1 được dùng để tìm L_2 (2-itemsets). L_2 được dùng để tìm L_3 (3-itemset) và tiếp tục cho đến khi không có k-itemset được tìm thấy.

- Từ frequent itemsets sinh ra các luật kết hợp mạnh (các luật kết hợp thỏa mãn 2 tham số min_sup và min_conf)

Apriori Algorithm

1. Duyệt (Scan) toàn bộ transaction database để có được support S của 1-itemset, so sánh S với min_sup, để có được 1-itemset (L_1)

2. Sử dụng L_{k-1} nối (join) L_{k-1} để sinh ra candidate k-itemset. Loại bỏ các itemsets không phải là frequent itemsets thu được k-itemset

3. Scan transaction database để có được support của mỗi candidate k-itemset, so sánh S với min_sup để thu được frequent k –itemset (L_k)

4. Lặp lại từ bước 2 cho đến khi Candidate set (C) trống (không tìm thấy frequent itemsets)

5. Với mỗi frequent itemset I, sinh tất cả các tập con s không rỗng của I

6. Với mỗi tập con s không rỗng của I, sinh ra các luật **s => (I-s)** nếu độ tin cậy (Confidence) của nó > =min_conf

Chẳn hạn với I= {A1,A2,A5},các tập con của I:

{A1}, {A2}, {A5}, {A1,A2},{A1,A5},{A2,A5}

sẽ có các luật sau

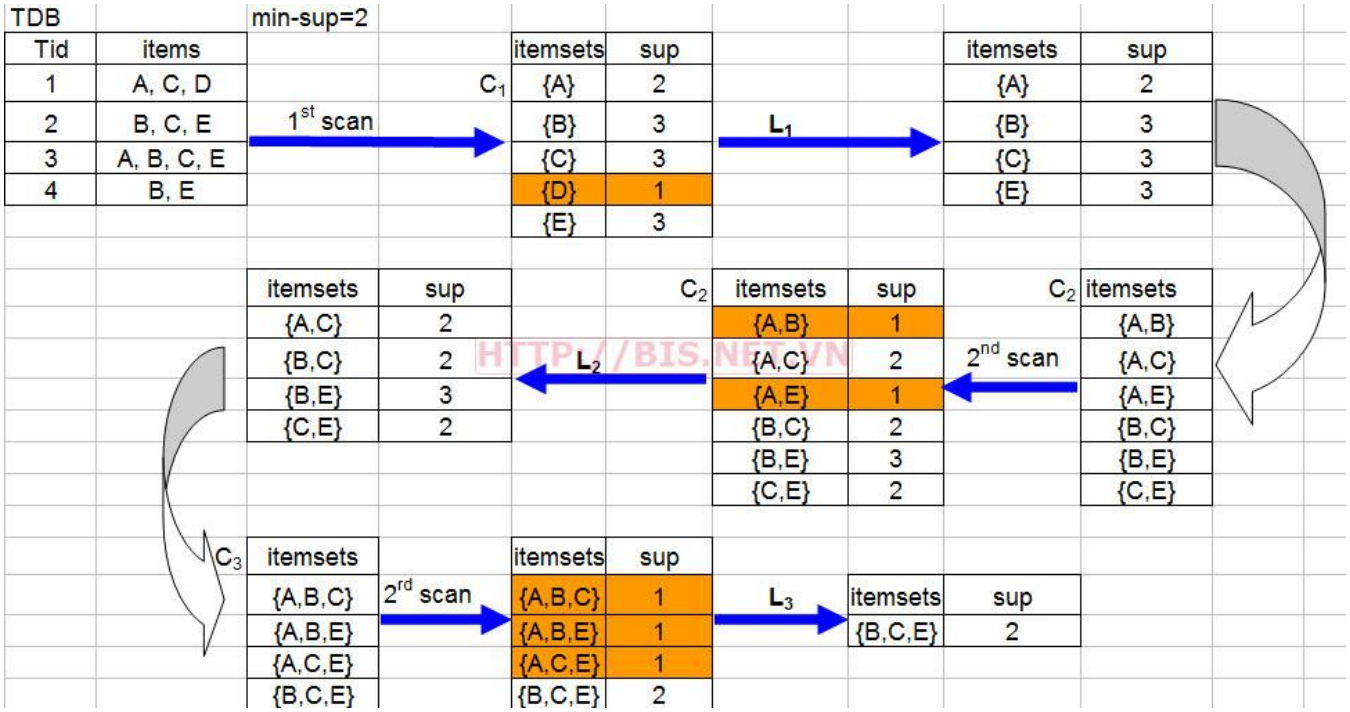
{A1} => {A2,A5},{A2} =>{A1,A5},{A5} =>{A1,A2}

{A1,A2} =>{A5},{A1,A5} =>{A2},{A2,A5} => {A1}

Ví dụ: Giả sử ta có có sở dữ liệu giao dịch (Transaction Database -TDB) như sau :

Tid	items
1	A, C, D
2	B, C, E
3	A, B, C, E
4	B, E

Thuật toán Apriori khai phá luật kết hợp được mô tả qua các bước sau



Ta có frequent itemsets I ={B,C,E}, với min_conf =80% ta có 2 luật kết hợp là {B,C} => {E} và {C,E} => {B}

	min-conf = 80%
Association Rule	Confidence
{B} => {C,E}	67%
{C}=>{B,E}	67%
{E}=>{B,C}	67%
{B,C}=>{E}	100%
{B,E}=>{C}	67%
{C,E}=>{B}	100%

Giả sử có cơ sở dữ liệu giao dịch bán hàng gồm 5 giao dịch như sau:

Tid	List of Items
1	Beer,Diaper,Baby Powder,Bread,Umbrella
2	Diaper,Baby Powder
3	Beer,Diaper,Milk
4	Diaper,Beer,Detergent
5	Beer,Milk,Coca-Cola

Thuật toán Apriori tìm các luật kết hợp trong giao dịch bán hàng trên như sau:

Step 1 min-sup =40% (2/5)

C_1

itemsets	support
Beer	4/5
Diaper	4/5
Baby Powder	2/5
Bread	1/5
Umbrella	1/5
Milk	2/5
Detergent	1/5
Coca-Cola	1/5

→

L_1

items	support
Beer	4/5
Diaper	4/5
Baby Powder	2/5
Milk	2/5

Step 2 and Step 3

C_2

items	support
Beer,Diaper	3/5
Beer, Baby Powder	0
Beer, Milk	2/5
Diaper, Baby Powder	2/5
Diaper, Milk	0
Baby Powder, Milk	0

→

L_2

items	support
Beer,Diaper	3/5
Beer, Milk	2/5
Diaper, Baby Powder	2/5

Step4 C_3 min-sup =40%

items	support
Beer,Diaper, Milk	1/5
Beer,Diaper,Baby Powder	1/5
Diaper, Milk, Baby Powder	0
Beer, Milk, Bayby Powder	0

→ L_3 = Empty (Stop)

Step 5
min_sup=40% min_conf=70%

itemsets	Support (A,B)	Support(A)	Confidence
Beer, Diaper	60%	80%	75.00%
Diaper, Beer	60%	80%	75.00%
Beer, Milk	40%	80%	50.00%
Milk, Beer	40%	40%	100.00%
Diaper, Baby Powder	40%	80%	50.00%
Baby Powder, Diaper	40%	40%	100.00%

Kết quả ta có các luật kết hợp sau (với min_sup= 40%, min_conf=70%)

R1: Beer => Diaper (support =60%, confidence = 75%)

R2: Diaper =>Beer (support =60%,confidence = 75%)

R3: Milk =>Beer (support =40%, confidence = 100%)

R4: Baby Powder => Diaper (support =40%,confidence = 100%)

*Từ kết quả các luật được sinh ra bởi giao dịch bán hàng trên, ta thấy rằng có luật có thể tin được (hợp lý) như **Baby Powder => Diaper**, có luật cần phải phân tích thêm như **Milk =>Beer** và có luật có vẻ khó tin như **Diaper =>Beer.**Ví dụ này sinh ra các luật có thể không thực tế vì dữ liệu dùng để phân tích (transaction database) hay còn gọi là tranining data rất nhỏ.*

Thuật toán Apriori được dùng để phát hiện các luật kết hợp dạng khẳng định (Positive Rule X=>Y) nhị phân (Binary Association Rules) chứ không thể phát hiện các luật kết hợp ở dạng phủ định (Negative Association Rule) chẳn hạn như các kết hợp dạng "Khách hàng mua mặt hàng A thường KHÔNG mua mặt hàng B" hoặc "Nếu ủng hộ quan điểm A thường KHÔNG ủng hộ quan điểm B". Khai phá các luật kết hợp dạng phủ định (Mining Negative Association Rules) có phạm vi ứng dụng rất rộng và thú vị nhất là trong Marketing, Health Care và Social Network Analysis.

Hồi qui (Regression)

Phương trình hồi qui tuyến tính một chiều

Ví dụ: Chúng ta có thể quan sát số tiền chi tiêu (yi) và thu nhập (xi) của 22 hộ gia đình trong một tháng có mối quan hệ với nhau như thế nào (1.000đ). Số liệu thu thập được trình bày ở bảng 11.3.

Từ bảng bên ta có:

n = 22

$$\sum x_i = 237.579 \qquad \sum y_i = 132.933$$
$$\sum x_i y_i = 1.448.555.000 \qquad \sum x_i^2 = 2.599.715.000$$

Bảng 6.3: Thu nhập và chi tiêu của 22 hộ gia đình trong một tháng

Thu nhập (x_i)	Chi tiêu (y_i)	Thu nhập (x_i)	Chi tiêu (y_i)	Thu nhập (x_i)	Chi tiêu (y_i)
9.098	5.492	10.282	5.871	12.018	6.718
9.138	5.540	10.662	6.157	12.523	6.921
9.094	5.305	11.019	6.342	12.053	6.471
9.282	5.507	11.307	5.907	12.088	6.394
9.229	5.418	11.432	6.124	12.215	6.555
9.347	5.320	11.449	6.186	12.494	6.755
9.525	5.538	11.697	6.224		
9.756	5.692	11.871	6.496		

$$\bar{x} = \frac{\sum x_i}{n} = \frac{237579}{22} = 10799$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{132.933}{22} = 6042,4$$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}.\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{1.448.555.000 - (22)(10799)(6042,4)}{2.599.715.000 - (22)(10799)^2}$$

$$b = 0,3815$$

$$a = \bar{y} - b\bar{x} = 6042,4 - (0,3815)(10799) = 1923$$

Đường hồi qui quan sát như sau: $y = 1923 + 0,3815x$

Phương trình này hàm ý rằng nếu thu nhập của hộ gia đình tăng lên 1.000đ thì trung bình chi tiêu tăng thêm là 381,5 đ. Còn 1923đ là phần chi tiêu do các nguồn khác.

b) Khoảng tin cậy và kiểm định giả thuyết trong hồi qui một chiều

Ví dụ: Trong sự liên hệ giữa chi tiêu và thu nhập mỗi hộ. Chúng ta có những thông tin như sau:

$$n = 22;\ b = 0{,}3815; \qquad S_b = 0{,}0253$$

Khoảng tin cậy 99% cho (được tính như sau:

Tra bảng phân phối t ta có: Ġ

Suy ra: $0{,}3815 - (2{,}845)(0{,}0253) < \beta < 0{,}3815 + (2{,}845)(0{,}0253)$

$$0{,}3095 < \beta < 0{,}4535$$

Vì vậy, với khoảng tin cậy 99%, cứ 1000 đồng tăng lên trong thu nhập/hộ thì chi tiêu tăng thêm nằm trong khoảng 309,5 đồng đến 453,5 đồng.

Ví dụ: Giả sử rằng chúng ta quan tâm đến dự đoán doanh thu bán lẻ trên hộ trong một năm mà trong đó thu nhập trên hộ/năm là 12 triệu đồng.

Ta có: $x_{n+1} = 12.000$, $a = 1.923$ và $b = 0{,}3815$

Suy ra:

$$\begin{aligned}\hat{y}_{n+1} &= a + bx_{n+1} \\ &= 1923 + (0{,}3815)(12.000) \\ &= 6501\end{aligned}$$

Vì vậy, khi thu nhập/năm là 12 triệu đồng thì doanh thu bán lẻ thu được trên hộ khoảng 6,5 triệu đồng. Dựa trên các đại lượng:

$n = 22,$ $\bar{x} = 10.799$ $\sum x_i^2 = 2.599.715.000$ ĉ và Ġ

Thế các đại lượng trên vào công thức (6.6) và (6.7) ta có kết luận sau:

- Dự báo giá trị thật cho doanh thu bán lẻ là 6,501 (321. Có nghĩa là với khoảng tin cậy 95% của doanh thu trong một năm tại mức thu nhập là 12 triệu nằm trong khoảng 6,18 triệu đến 6,82 triệu.

- Và dự báo cho giá trị mong đợi của doanh thu bán lẽ là 6,501 (91. Như vậy, rõ ràng rằng trong cùng khoảng tin cậy nhưng sự không chắc chắn trong việc dự báo cho giá trị thật thì lớn hơn giá trị mong đợi vì dự báo giá trị thật có khoảng ước lượng rộng hơn.

Chú ý: Nếu tất cả các yếu tố khác không đổi thì:

- Cỡ mẫu n càng lớn, càng hẹp khoảng tin cậy khi ước lượng, dự báo càng chính xác.

-Ġ càng lớn, khoảng tin cậy ước lượng càng lớn, dự báo càng kém chính xác.

- Phương saiĠ càng lớn, thì khoảng tin cậy ước lượng càng hẹp, dự báo càng chính xác.

-Ġcàng lớn, khoảng tin cậy ước lượng càng rộng, và dự báo càng kém chính xác.

Hồi qui nhiều chiều: (Multiple Regression)

a) Phương trình hồi qui nhiều chiều:

Mục tiêu của mô hình này giải thích biến phụ thuộc (y) bị ảnh hưởng bởi nhiều biến độc lập (xi). Ví dụ, trong kinh doanh ngành ngân hàng, lợi tức thu được từ việc chênh lệch giữa lãi suất tiền gởi và cho vay phụ thuộc ít nhất vào hai yếu tố: Phần trăm tăng lên trong lượng tiền gởi (x1) và số đơn vị đến gởi (x2). Để xét mối quan hệ này ta sử dụng tài liệu thu thập của ngân hàng qua 25 năm như sau:

Bảng 6.4: Lợi tức, % tăng của tiền gởi và số đơn vị gởi tiền qua 25 năm

Năm	x_1(%)	x_2	y (%)	Năm	x_1(%)	x_2	y (%)
1	3,92	7.298	0,75	14	3,78	6.672	0,84
2	3,61	6.855	0,71	15	3,82	9.890	0,79
3	3,32	6.636	0,66	16	3,97	7.115	0,70
4	3,07	6.506	0,61	17	4,07	7.327	0,68
5	3,06	6.450	0,70	18	4,25	7.546	0,72
6	3,11	6.402	0,72	19	4,41	7.931	0,55
7	3,21	6.368	0,77	20	4,49	8.097	0,63
8	3,26	6.340	0,74	21	4,70	8.468	0,56
9	3,42	6.349	0,90	22	4,58	8.717	0,41
10	3,42	6.352	0,82	23	4,69	8.991	0,51
11	3,45	6.361	0,75	24	4,71	9.179	0,47
12	3,58	6.369	0,77	25	4,78	9.318	0,32
13	3,66	6.546	0,78				

Phương trình hồi qui nhiều chiều cho ví dụ này có dạng:

$$y = a + b_1x_1 + b_2x_2$$

Một cách tổng quát, phương trình hồi qui tuyến tính nhiều chiều có dạng:

$$y = a + b_1x_1 + b_2x_2 + ... + b_kx_k \quad (6.8)$$

Các tham số a, b1, b2,....,bn có thể được ước lượng dễ dàng nhờ các phần mềm có sẵn trên máy tính. Phương trình này sẽ được suy rộng cho tổng thể có biến phụ thuộc Y và các biến độc lập X1,X2,...,Xk . Trở lại ví dụ trên các tham số có kết quả giải bằng phương pháp ma trận hoặc từ phần mềm Excel như sau:

$$a = 1,565 ; b_1 = 0,237 ; b_2 = - 0,000249$$

Vì vậy: $y = 1,565 + 0,237x1 - 0,000249x2$

Giải thích:

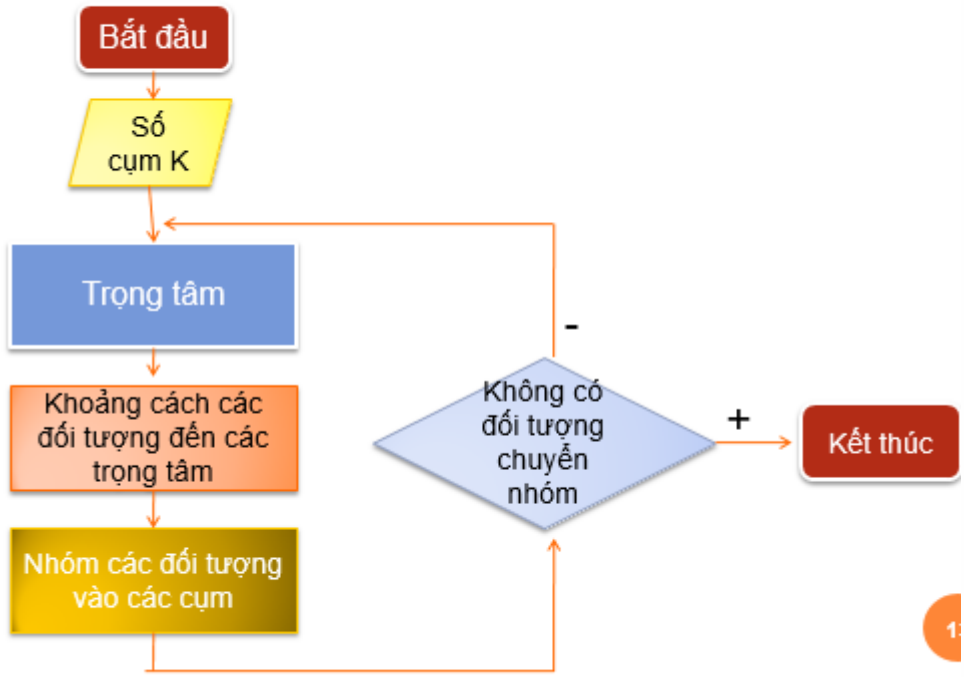
- Khi cố định số lượng đơn vị tiền gởi (x2), lươûng tiền gởi tăng 1% dẫn đến 0,237% tăng lên trong lợi tức.
- Khi cố định % tăng lên trong lượng tiền gởi (x1), cứ tăng lên 1000 đơn vị tiền gởi dẫn đến giảm trong lợi tức 0,249%.
- Ngoài hai nhân tố trên, các nhân tố khác làm tăng lợi tức 1,565% (các nguồn thu từ Nhà nước chẳng hạn).

Phân cụm (Clustering)

Phân cụm là gì?
Quá trình phân chia 1 tập dữ liệu ban đầu thành các cụm dữ liệu thỏa mãn:
Các đối tượng trong 1 cụm "tương tự" nhau.
Các đối tượng khác cụm thì "không tương tự" nhau.

Giải quyết vấn đề tìm kiếm, phát hiện các cụm, các mẫu dữ liệu trong 1 tập hợp ban đầu các dữ liệu không có nhãn.

II.2. CÁC BƯỚC CỦA THUẬT TOÁN

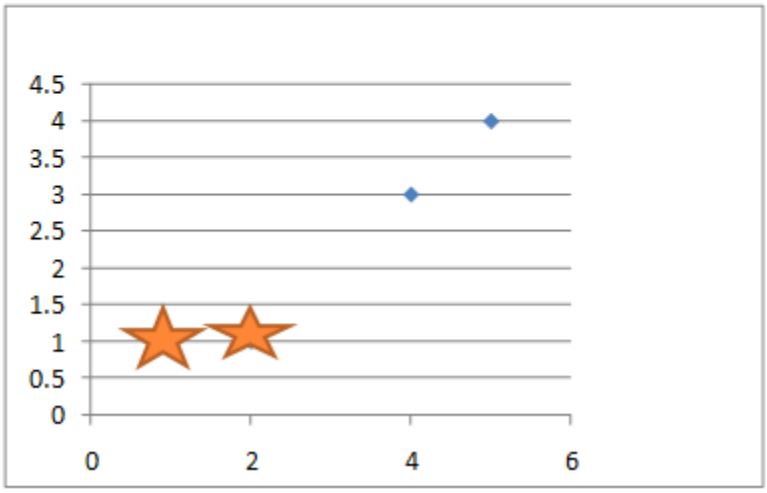


II.3 VÍ DỤ MINH HỌA

- **Bước 1:** Khởi tạo

  Chọn 2 trọng tâm ban đầu:

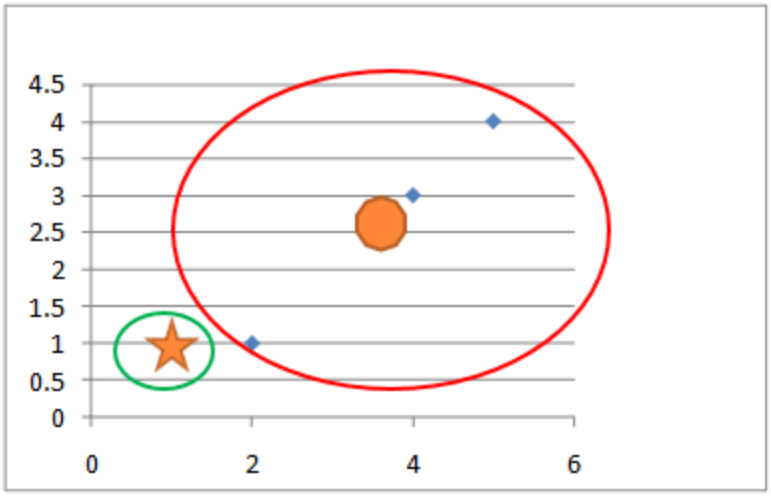
  $c_1(1,1) \equiv A$ và $c_2(2,1) \equiv B$, thuộc 2 cụm 1 và 2



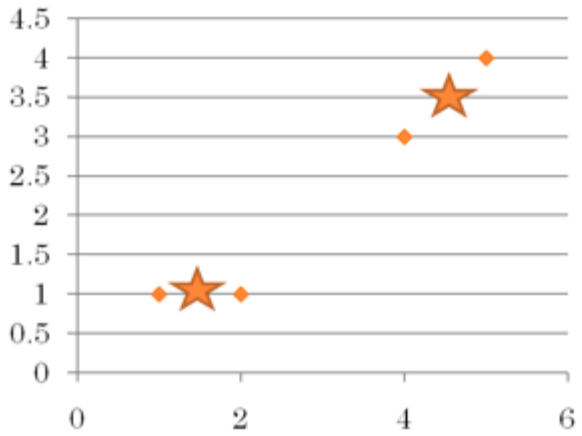
II.3 VÍ DỤ MINH HỌA

- **Bước 2:** Tính toán khoảng cách
  - d(C, c_1) = $(4-1)^2+(3-1)^2$
    = 13
    d(C, c_2) = $(4-2)^2+(3-1)^2$
    = 8
    d(C, c_1) > d(C, c_2) ⇨ C thuộc cụm 2
  - d(D, c_1) = $(5-1)^2+(4-1)^2$
    = 25
    d(D, c_2) = $(5-2)^2+(4-1)^2$
    = 18
    d(D,c_1) > d(D, c_2) ⇨ D thuộc cụm 2

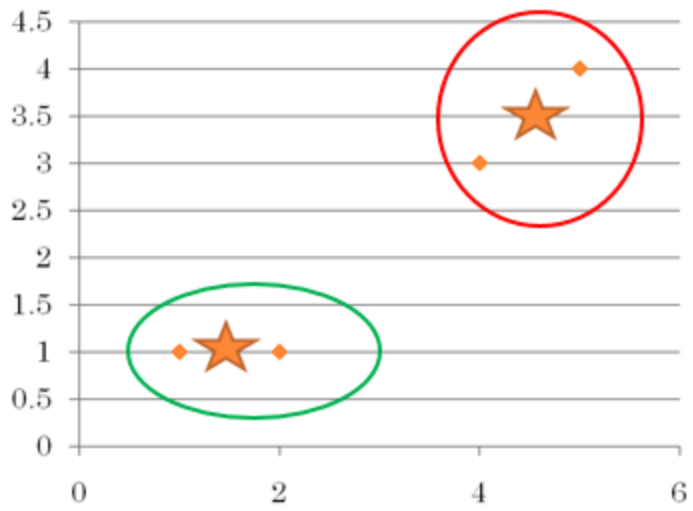
- **Bước 3:** Cập nhật lại vị trí trọng tâm
  - Trọng tâm cụm 1 c_1 ≡ A (1, 1)
  - Trọng tâm cụm 2 c_2 (x,y) = $(\frac{2+4+5}{3}, \frac{1+3+4}{3})$



- **Bước 4-1:** Lặp lại bước 2 – Tính toán khoảng cách
  - $d(A, c_1) = 0 < d(A, c_2) = 9.89$
    A thuộc cụm 1
  - $d(B, c_1) = 1 < d(B, c_2) = 5.56$
    B thuộc cụm 1
  - $d(C, c_1) = 13 > d(C, c_2) = 0.22$
    C thuộc cụm 2
  - $d(D, c_1) = 25 > d(D, c_2) = 3.56$
    D thuộc cụm 2
- **Bước 4-2:** Lặp lại bước 3-Cập nhật trọng tâm
  $c_1 = (3/2, 1)$ và $c_2 = (9/2, 7/2)$



- **Bước 4-3:** Lặp lại bước 2
  - $d(A, c_1) = 0.25 < d(A, c_2) = 18.5$
    A thuộc cụm 1
  - $d(B, c_1) = 0.25 < d(B, c_2) = 12.5$
    B thuộc cụm 1
  - $d(C, c_1) = 10.25 < d(C, c_2) = 0.5$
    C thuộc cụm 2
  - $d(D, c_1) = 21.25 > d(D, c_2) = 0.5$
    D thuộc cụm 2


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
2
4
6