BÀI TẬP MẪU MÔN HỌC DATA MINING & DATA WAREHOUSE

Thang 10/2004

1. Tập phổ biến và luật kết hợp

Cho bối cảnh gồm các giao tác : o1={d1,d3,d4} ; o2={d1,d3,d4}, o3={d3,d5} ; o4={d4,d5} ; o5 = {d2,d3,d5}

Bối cảnh nhị phân

	d1	d2	d3	d4	d5
o1	1	0	1	1	0
o2	1	0	1	1	0
o3	0	0	1	0	1
o4	0	0	0	1	1
o5	0	1	1	0	1

a) Tìm tất cả các tập phổ biến với minsupp=0,3

Với minsupp =0,3 , số dòng là 5*0,3 =1,5 hay 2 dòng

F1={{d1},{d3},{d4},{d5}}

Tính C1

	d1	d3	d4	d5
d1				
d3	d1,d3			
d4	d1,d4	d3,d4		
d5	d1,d5	d3,d5	d4,d5	

Từ C1 tính F2:

C1={{ d1,d3}, {d1,d4}, {d1,d5},{d3,d4},{d3,d5},{d4,d5}}

F2={ {d1,d3}, {d1,d4}, {d3,d4},{d3,d5} }

Tính C2

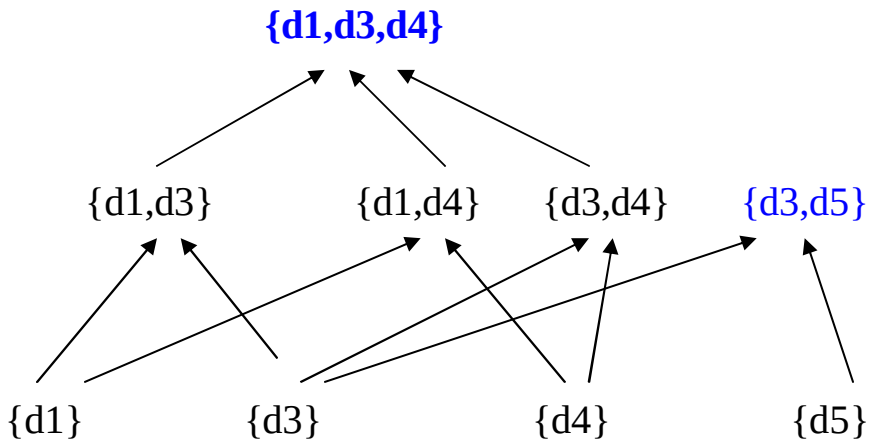
	d1d3	d1d4	d3d4	d3d5
d1d3				
d1d4	d1,d3,d4			
d3d4	d1,d3,d4	d1,d3,d4		
d3d5	d1,d3,d5	d1,d3,d4,d5	d3,d4,d5	

C2={{d1,d3,d4} ,{d1,d3,d5} , { d3,d4,d5 }}

F3={{d1,d3,d4}

Tập tối đại (maximal frequent sets)

{d1,d3,d4} ; {d3,d5}



Tạo luật kết hợp từ các tập tối đại:
Với tập phổ biến tối đại : {d1,d3,d4}
Các luật khả dĩ:
{d1} $\rightarrow${d3,d4}
{d3} $\rightarrow${d1,d4}
{d4} $\rightarrow${d1,d3}
{d3,d4}$\rightarrow$ {d1}
{d1,d4} $\rightarrow$ {d3}
{d1,d3} $\rightarrow$ {d4}

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 $\rho(S1) \cap \rho(S2) = \rho(S1)$
{d1} $\rightarrow${d3,d4} vì $\rho(\{d1\})=\{o1,o2\} \subseteq \rho(\{d3,d4\}) =\{o1, o2\}$
{d3,d4} $\rightarrow$ {d1} vì $\rho(\{d3,d4\}) =\{o1, o2\} \subseteq \rho(\{d1\})=\{o1,o2\}$
Xét {d4} $\rightarrow$ {d1,d3} và {d1,d3} $\rightarrow$ {d4}

ρ({d4})={o1, o2, o4 } và ρ({d1,d3})={o1, o2}
Chọn {d1,d3} → {d4} vì ρ({d1,d3}) ⊆ ρ({d4})=

2. 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.

3. Bài tập về tập thô

Hệ thông tin

	Troi	Gio	Apsuat	Ketqua
O1	Trong	Bac	Cao	Kmua
O2	May	Nam	Cao	Mua
O3	May	Bac	TB	Mua
O4	Trong	Bac	Thap	Kmua
O5	May	Bac	Thap	Mua
O6	May	Bac	Cao	Mua
O7	May	Nam	Thap	Kmua
O8	Trong	Nam	Cao	Kmua

1. Tính xấp xỉ :

X = {o1, o3 , o4 }

Các lớp tương đương:

Trong Bac {o1, o4 }

May Nam {o2 , o7 }

May Bac {o3 , o6 , o5 }

Trong Nam {o8 }

Với B = { Troi, Gio }

Lower(B,X) = {o1 , o4 }

Upper(B,X) = { o1, o4 , o3 , o5 , o6 }

$$\alpha = \frac{|Lower(B,X)|}{Upper(B,X)} = \frac{|\{o1,o4\}|}{|\{o1,o4,o3,o5,o6\}|} = \frac{2}{5} = 0.4$$

Khảo sát sự phụ thuộc của C={Ketqua} với B={Troi,Gio}

Với C = { Ketqua}

X1={o1, o4 , o7 , o8 }

X2={o2, o3 , o5 , o6 }

Tính

Lower(B,X1) = {o1 , o4 , o8 }

Lower(B,X2) = { o3 , o5 , o6 }

Vậy

$$k = \frac{|Lower(B,X1)| + |Lower(B,X2)|}{|O|} = \frac{6}{8} = 0.66$$

Đề tìm luật phân lớp
Giả sử xét các luật
{Troi,Gio} → {Ketqua}
Với B={Troi,Gio}, ta có các lớp tương đương
Z1= {o1, o4 } Trong Bac
Z2 = {o2 , o7 } May Nam
Z3 = {o3 , o5, o6 } May Bac
Z4= {o8 } Trong Nam
Với C={Ketqua}, ta có các lớp tương đương
X1={o1, o4 , o7 , o8 } Khong mua
X2={o2, o3 , o5 , o6 } Mua

Tìm luật

1. Xét các phần giao Z1= {o1, o4 } Trong Bac
Z1 ∩ X1 = {o1 , o4 } ≠∅ , ngoài ra Z1 ⊆ X1 , nên ta có luật phân lớp đúng chính xác 100%
Nếu trời = trong và gio = bac Thì Khong Mua
2. Xét các phần giao Z2 = {o2 , o7 } May Nam
Z2 ∩ X1 = {o7 } ≠∅ , ngoài ra Z2 ⊄ X1 , nên ta có luật phân lớp không đúng chính xác 100%
Nếu trời = may và gio = nam Thì Khong Mua
3. Xét các phần giao Z3 = {o3 , o6 } May Bac
Z1 ∩ X3 = {o1 , o4 } =∅ không có luật phân lớp
……..

Bài tập về reduct

Cho hệ thông tin

	a	b	c	d	e
o1	1	0	2	1	0
o2	0	0	1	2	1
o3	2	0	2	1	0
o4	0	0	2	2	2
o5	1	1	2	1	0

Ma trận phân biệt

	o1	o2	o3	o4	o5
o1					
o2	a,c,d,e				
o3	a	a,c,d,e			
o4	a,d,e	c,e	a,d,e		
o5	b	a,b,d,e	a,b	a,b,d,e	

Hàm phân biệt:

F(a,b,c,d,e) = (a ∨ c ∨ d ∨ e) ∧ (a) ∧ (a ∨ d ∨ e) ∧ (b) ∧(c ∨ e) ∧ (a ∨ b ∨ d ∨ e) ∧ (a ∨ b)

Rút gọn bằng luật hút :

(a ∧ b) ∨ a = a hay (a ∨ b) ∨ a = a

Ta có :

F(a,b,c,d,e)= (a) ∧ (b) ∧ (c ∨ e)

= (a ∧ b ∧ c) ∨ (a ∧ b ∧ e)

Các reducts là {a,b,c} và { a,b,e}