BÀI 4

TẬP THÔ (REDUCT)

Nội dung

- Ứng dụng của tập thô
- Các khái niệm
- Các bài tập liên quan

Ứng dụng của tập thô (reduct)

- Dùng để khắc phục hiện tượng dữ liệu dùng để KPDL bị nhiễu
- Rút gọn dữ liệu (khử dữ liệu thừa)
- Tạo luật quyết định
- Nhận diện phụ thuộc riêng phần và toàn phần của các thuộc tính

Các khái niệm

- Hệ thông tin, hệ quyết định
- Quan hệ bất khả phân biệt(indiscernibility)
- Xấp xỉ tập hợp (set approximation)
- Rút gọn
- Phụ thuộc thuộc tính

Hệ thông tin (Information System)

- IS là cặp (U, A)
- U là tập khác rỗng các đối tượng.
- A là tập hữu hạn các thuộc tính sao cho với mọi $a \in A.$

$$a : U \rightarrow V_a$$

- V_a được gọi là tập trị của a.

	Độ tuổi	Số buổi
x1	16-30	50
x2	16-30	0
x3	31-45	1-25
x4	31-45	1-25
x5	46-60	26-49
x6	16-30	26-49
x7	46-60	26-49

Hệ quyết định (Decision system)

- DS: (U , A ∪ {d})
- $d \notin A$ là thuộc tính quyết định (có thể có nhiều thuộc tính quyết định).
- Các phần tử của A được gọi là thuộc tính điều kiện.

	Age	số buổi	thi đậu
x1	16-30	50	yes
x2	16-30	0	no
x3	31-45	1-25	no
x4	31-45	1-25	yes
x5	46-60	26-49	no
x6	16-30	26-49	yes
x7	46-60	26-49	no

Bảng 1

Một số nhận xét

- {x3,x4}, {x5,x7}: Có cùng thuộc tính điều kiện {độ tuổi, số buổi}
  - x3,x4: khác nhau về giá trị thuộc tính quyết định
  - x5,x7: Có cùng kết quả thi đậu
- Ví dụ 1 luật được rút ra:
  - "*Nếu Độ_tuổi là 16-30 và Số_buổi là 50 thì Thi_đậu là Có*".

Các vấn đề của bảng quyết định

- Có thể biểu diễn lặp lại các đối tượng giống nhau hay bất khả phân biệt
- Một số thuộc tính có thể thừa

Quan hệ bất khả phân biệt

- Cho IS = (U, A) là hệ thông tin,với tập $B \subseteq A$
- Có quan hệ tương đương tương ứng :

$$IND_{IS}(B) = \{(x, x') \in U^2 \mid \forall a \in B, a(x) = a(x')\}$$

- $IND_{IS}(B)$được gọi là quan hệ bất khả phân theo B (B-indiscernibility relation)

- Nếu $(x, x') \in IND_{IS}(B)$,thì các đối tượng x và x' là không thể phân biệt nhau qua tập thuộc tính B.
- Các lớp tương đương của quan hệ bất khả phân theo B được ký hiệu là

Ví dụ về quan hệ bất khả phân biệt

	Tuổi	số buổi	thi đậu
x1	16-30	50	yes
x2	16-30	0	no
x3	31-45	1-25	no
x4	31-45	1-25	yes
x5	46-60	26-49	no
x6	16-30	26-49	yes
x7	46-60	26-49	no

- IND({Tuổi})= {{x1,x2,x6},{x3,x4},{x5,x7};
- IND({sốbuổi})={{x1},{x2},{x3,x4},{x5,x6,x7}}
- IND({Tuổi, số buổi})={{x1},{x2},{x3,x4},{x5,x7},{x6}};

Các quan sát

- Quan hệ tương đương (bất khả phân biệt) dẫn đến một phân hoạch tập phổ quát.
- Có thể dùng các phân hoạch để tạo các tập con mới của tập phổ quát.
- Các tập con thường được quan tâm có cùng giá trị thuộc tính điều kiện.

Xấp xỉ tập hợp

- **Lý do**: Không thể định nghĩa rõ ràng tập các khách hàng có thuộc tính quyết định dương (thi đậu = có) từ các thuộc tính khác.
  - Trong **bảng 1(slide 6)**: Những khách hàng gặp khó khăn là các bộ x_3,x_4 -> không thể có 1 định nghĩa chính xác của những khách hàng như vậy từ bảng này.

    => Tập thô

Xấp xỉ tập hợp (tt)

- **Mục đích**:
  - Chỉ ra được khách hàng nào có thuộc tính quyết định có giá trị dương
  - Chỉ ra được khách hàng nào có thuộc tính quyết định không có giá trị dương.
  - Những khách hàng nào thuộc vào vùng biên giữa các trường hợp chắc chắn.

Xấp xỉ tập hợp (tt)

- **Định nghĩa**:
  - Gọi T = (U, A) và $B \subseteq A$ và $X \subseteq U$
    Chúng ta có thể xấp xỉ X dùng các thông tin chứa trong B bằng cách tạo các xấp xỉ B-dưới và B-trên của X, ký hiệu lần lượt là $\underline{B}X$ và $\overline{B}X$ với

$$\underline{B}X = \{x \mid [x]_B \subseteq X\}$$

$$\overline{B}X = \{x \mid [x]_B \cap X \neq \phi\}$$

Xấp xỉ tập hợp (tt)

- Các đối tượng trong $\underline{B}X$ chắc chắn được phân lớp như là các thành viên của tập X
- Các đối tượng trong $\overline{B}X$ chỉ có thể phân lớp là các đối tượng dương tính
- Vùng B-biên của X, $BN_B(X) = \overline{B}X - \underline{B}X,$
  - Chứa các đối tượng không thể phân lớp chắc chắn vào X theo B
- Vùng B-ngoài của X, $U - \overline{B}X$
  - Chứa các đối tượng chắn chắn được phân lớp không thuộc về X
- Một tập được gọi là thô (rough) nếu vùng biên của nó khác rỗng, ngược lại tập là rõ

Ví dụ về xấp xỉ tập hợp

	tuổi	số buổi	thi đậu
x1	16-30	50	yes
x2	16-30	0	no
x3	31-45	1-25	no
x4	31-45	1-25	yes
x5	46-60	26-49	no
x6	16-30	26-49	yes
x7	46-60	26-49	no

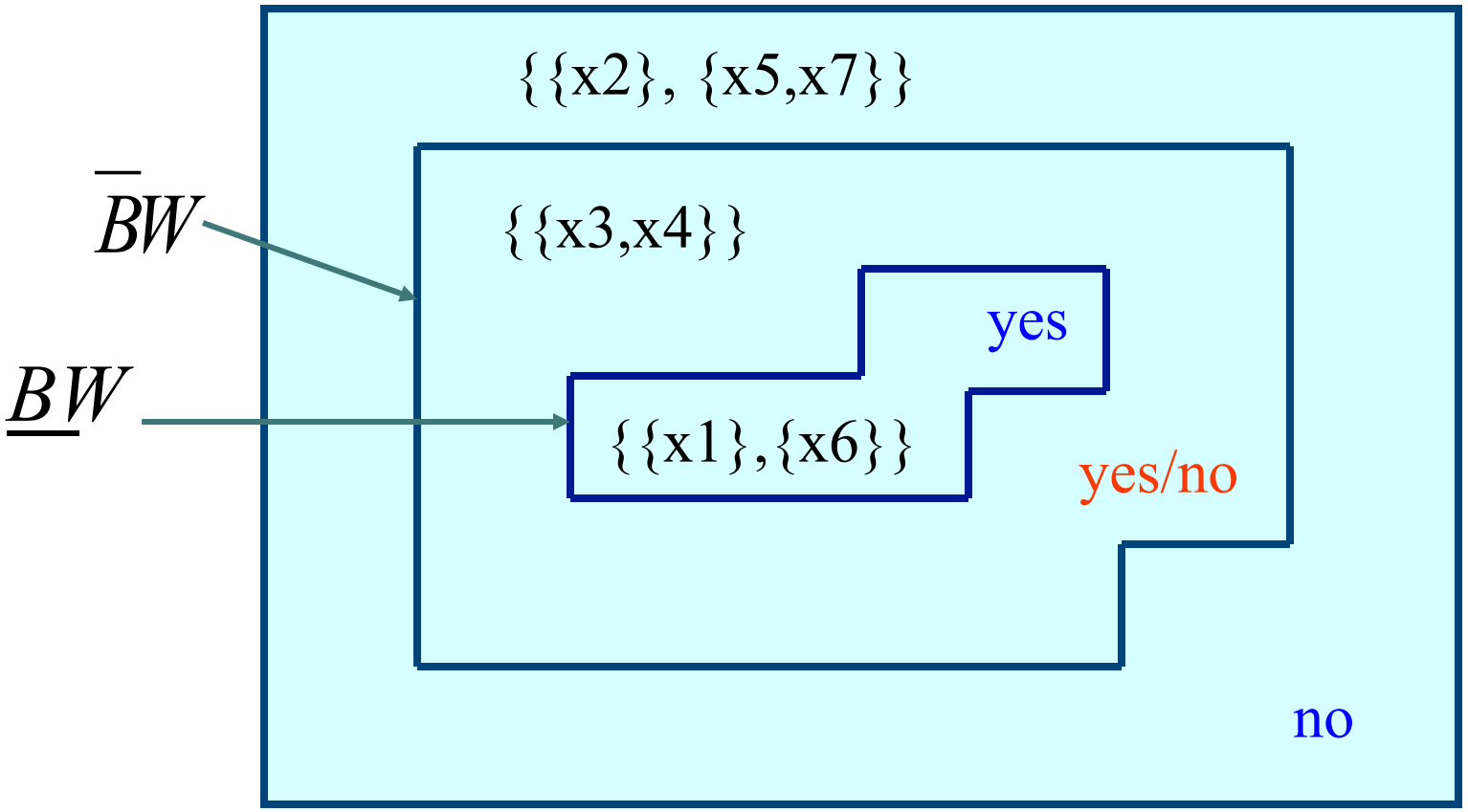
Ví dụ về xấp xỉ tập hợp (tt)

- Gọi tập các đối tượng W = {x |thi đậu(x) = yes} <=> W={x1,x4,x6}
- Và B={Độ tuổi, số buổi}
- IND({Độ_tuổi, Số_buổi})={{x_1};{x_2}; {x_3,x_4}; {x_5; x_7}; {x_6}}

$$\underline{B}W = \{x1, x6\},$$
$$\overline{B}W = \{x1, x3, x4, x6\},$$
$$BN_A(W) = \{x3, x4\},$$
$$U - \overline{B}W = \{x2, x5, x7\}.$$

- Như vậy lớp quyết định Thi_đậu là thô vì vùng biên khác rỗng (hình dưới)

Ví dụ về xấp xỉ (tt)



Độ chính xác của tập thô

$$\alpha_B(X) = \frac{|\underline{B}(X)|}{|\overline{B}(X)|}$$

- Với |X| là lực lượng của $X \neq \phi$.
- Rõ ràng $0 \leq \alpha_B \leq 1$.
- Nếu $\alpha_B(X) = 1$, X là rõ so với B.
- Nếu $\alpha_B(X) < 1$, X là thô so với B.

Các vấn đề của bảng quyết định

- Có thể biểu diễn nhiều lần các đối tượng giống nhau hay bất khả phân biệt.
- Một số thuộc tính có thể bị dư. Nghĩa là có thể loại bỏ chúng mà không làm xấu đi việc phân lớp.

Các rút gọn

- Chỉ giữ lại các thuộc tính bảo toàn quan hệ bất khả phân biệt và hệ quả là bảo toàn xấp xỉ tập hợp.

- Thường có nhiều tập con như thế và tập con nhỏ nhất được gọi là rút gọn(reducts).

Các bước thực hiện

- Xác định ma trận phân biệt
- Xác định hàm phân biệt và rút gọn hàm

Ma trận phân biệt

- Cho IS=(U,A) là 1 hệ thông tin ma trận phân biệt của S là 1 ma trận **n×n** (**n** là số đối tượng), với c_{ij} được tính bởi công thức:
  - $c_{ij} = \{ a \in A \mid a(x_i) \neq a(x_j) \}$ với i; j = 1,.., n

Hàm phân biệt

- $f_{IS}(a_1, ..., a_m)$ = trong đó $c_{ij} = \{ a \mid a \in c_{ij} \}$
- Tập các đơn thức của f_{IS} xác định tập các rút gọn của IS

Ví dụ về rút gọn

	Bằng_cấp d	Kinh_nghiệm e	Tiếng_Anh f	Giới_thiệu r	Tuyển_dụng
X_1	MBA	Vừa	Tốt	Xuất_sắc	Chấp_nhận
X_4	MSC	Nhiều	Tốt	Trung_bình	Chấp_nhận
X_6	MSC	Nhiều	Tốt	Xuất_sắc	Chấp_nhận
X_7	MBA	Nhiều	Không	Tốt	Chấp_nhận
X_2	MBA	Thấp	Tốt	Trung_bình	Từ_chối
X_3	MCE	Thấp	Tốt	Tốt	Từ_chối
X_5	MSC	Vừa	Tốt	Trung_bình	Từ_chối
X_8	MCE	Thấp	Không	Xuất_sắc	Từ_chối

Ví dụ về rút gọn (tt)

- Đặt:
  - Bằng cấp : d
  - Kinh nghiệm: e
  - Tiếng anh: f
  - Giới thiệu: r

Ma trận phân biệt

	[x1]	[x4]	[x6]	[x7]	[x2]	[x3]	[x5]	[x8]
[x1]	∅							
[x4]	∅	∅						
[x6]	∅	∅	∅					
[x7]	∅	∅	∅	∅				
[x2]	e,r	d,e	d,e,r	e,f,r	∅			
[x3]	d,e,r	d,e,r	d,e,r	d,e,f	∅	∅		
[x5]	d,r	e	e,r	d,e,f,r	∅	∅	∅	
[x8]	d,e,f	d,e,f,r	d,e,f	d,e,r	∅	∅	∅	∅

Ma trận phân biệt (tt)

- f=(evr)^(dve)^(dvevr)^(evfvr)^(dvevr)^
  (dvevr)^(dvevr)^(dvevf)^(dvr)^(e)^(evr)
  (dvevfvr)^(dvevf)^(dvevfvr)^(dvevf)^
  (dvevr)
- f được rút gọi lại như sau:
  - f = ed v er
- Vậy hệ quyết định có 2 rút gọn là
  - Kinh nghiệm,bằng cấp và kinh nghiệm, giới thiệu

Phụ thuộc thuộc tính

- Tập thuộc tính D phụ thuộc hoàn toàn vào tập thuộc tính C, ký hiệu là $C \Rightarrow D$, nếu tất cả các thuộc tính của D đều được xác định duy nhất bởi giá trị của các thuộc tính trong C.
- Công thức tính: $$k = \gamma(C, D) = \sum_{X \in U/D} \frac{|\underline{C}(X)|}{|U|}$$

Nếu k = 1 thì D phụ thuộc hoàn toàn vào C.
Nếu k < 1 thì D phụ thuộc một phần (theo mức độ k) vào C

Bài tập

- Cho hệ quyết định như sau:

	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

Yêu cầu

- a) Tính xấp xỉ tập X = {o1, o3 , o4} qua tập thuộc tính B={trời,gió}
- b) Khảo sát sự phụ thuộc tính của C={Ketqua} vào B={trời,gió}

Bài giải_câu 1a

Các lớp tương đương:

- IND(trời,gió)={{o1, o4},{o2, o7},{o3, o5, o6},{o8}}

Bài giải_câu 1a (tt)

- Với B = { Troi, Gio }, ta có :
  - Xấp xỉ dưới của X qua tập thuộc tính B là $\underline{B}(X)$ = {o1 , o4 }
  - Xấp xỉ trên của X qua tập thuộc tính B là Upper(B,X) = { o1, o4 , o3 , o5 , o6 }

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

Bài giải_câu 1b

- Với C = { Ketqua} , ta có:
  - X1={o1,o4,o7,o8 } và
  - X2={o2, o3 , o5 , o6 }
- Ta tính:
  - B(X1) = {o1 , o4 , o8 } và
  - B(X2) = { o3 , o5 , o6 }
- Do vậy:

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

Tìm reduct cho hệ quyết định sau

TT	Tên người	Màu tóc	Chiều cao	Cân nặng	Dùng thuốc	Kết quả
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

Bài giải

- Đặt tập thuộc tính điều kiện:
  - Màu tóc:T
  - Chiều cao: C
  - Cân nặng: N
  - Dùng thuốc:D

Ma trận phân biệt

\	O1	O2	O3	O4	O5	O6	O7
O2	C,N,D						
O3	T,C,N,D	λ					
O4	λ	C,D	T,D				
O5	λ	T,C,N,D	T,C,N,D	λ			
O6	T,C,N	λ	λ	T,C,N	T,C		
O7	T,N	λ	λ	T,C,N	T	λ	
O8	C,D	λ	λ	N,D	T,C,N,D	λ	λ

Các reduct

- Ta có hàm phân biệt:
- F(T,C,N,D)= (C∨N∨D) ∧ (T∨C∨N∨D) ∧ (T∨C∨N)∧ (T∨N) ∧ (C∨D) ∧ (T∨D) ∧ (N∨ D) ∧ (T∨ C) ∧ (T) = T ∧ (C∨ D) ∧ (N∨ D) = (T ∧ D) ∨ (T ∧ C ∧ N)
- Vậy các reducts của hệ thông tin trên là B_1={T,D} và B_2={T,C,N}

Liệt kê luật có độ chính xác 100%

- Ta có phân hoạch U trên B1: U/ B1={Z1={o1,o4}, Z2={o2,o8}, Z3={o3}, Z4={o5}, Z5={o6,o7}}
- Phân hoạch U trên B2: U/B2={Z6={o1}, Z7={o2}, Z8={o3}, Z9={o4}, Z10={o5}, Z11={o6}, Z12= {o7}, Z13={o8}}

Liệt kê luật có độ chính xác 100% (tt)

- **X là các đối tượng phân lớp theo kết quả thì ta có 2 phân lớp:**
  - **X1={o1, o4, o5} //các đối tượng có kết quả bị rám**
  - **X2={o2, o3, o6, o7,o8}//các đối tượng không bị rám**

Nhận xét: nếu các lớp tương đương Bi là tập con của Xj, luật dạng Bi →Xj có độ chính xác phân lớp là 100%. Do vậy, ta có luật sau :

- Xét các luật có dạng B1→X1 (với X1={o1,o4,o5})
- Vì Z1⊆X1 nên ta có luật "Màu tóc=Đen và DùngThuốc=Không → Kết quả = bị rám"
- Vì Z4⊆X1 nên ta có luật "Màu tóc=Bạc và D=Không → Kết quả=bị rám"

Liệt kê luật có độ chính xác 100%(tt)

Tương tự, xét các luật có dạng:

B1→X2(X2={o2,o3,o6,o7,o8})

- Vì Z2⊆X2 nên ta có luật "Màu tóc=Đen và DùngThuốc=Có → Kết quả=không"
- Vì Z3⊆X2 nên ta có luật "Màu tóc=Râm và DùngThuốc=Có → Kết quả=Không"
- Vì Z5⊆X2 nên ta có luật "Màu tóc=Râm và DùngThuốc=không → Kết quả=Không"

Liệt kê luật có độ chính xác 100%(tt)

Tương tự, xét các luật có dạng
B2→X1(X1={o1,o4,o5})

- Vì Z6⊆X1 nên ta có luật "Màu tóc=Đen và ChiềuCao=Tầm thước và CânNặng=nhẹ → Kết quả=bị rám"
- Vì Z9⊆X1 nên ta có luật "Màu tóc=Đen và ChiềuCao=Thấp và CânNặng=vừa → Kết quả=bị rám"
- Vì Z10⊆X1 nên ta có luật "Màu tóc=Bạc và ChiềuCao=Tầm thước và CânNặng=nặng => Kết quả=Bị rám"

Liệt kê luật có độ chính xác 100%(tt)

Tương tự, xét các luật có dạng B2→ X2

- (X2={o2,o3,o6,o7,o8})
- Vì Z7⊆X2 nên ta có luật "Màu tóc=Đen và ChiềuCao=Cao và CânNặng= vừa → Kết quả=không"
- Vì Z8⊆X2 nên ta có luật "Màu tóc=Râm và ChiềuCao=Thấp và CânNặng=vừa → Kết quả=không"
- Vì Z11⊆X2 nên ta có luật "Màu tóc=Râm và ChiềuCao=Cao và CânNặng=nặng → Kết quả=không"
- Vì Z12⊆X2 nên ta có luật "Màu tóc=Râm và ChiềuCao=Tầm thước và CânNặng=nặng → Kết quả=không"
- Vì Z13⊆X2 nên ta có luật "Màu tóc=Đen và ChiềuCao=thấp và CânNặng=nhẹ => Kết quả=không"

Bài tập

	Bằng cấp	Kinh nghiệm	Tiếng Anh	Lời giới thiệu	Tuyển dụng
O1	MBA	Trung bình	Biết	Xuất sắc	Chấp nhận
O2	MSC	Nhiều	Biết	Bình thường	Chấp nhận
O3	MSC	Nhiều	Biết	Xuất sắc	Chấp nhận
O4	MBA	Nhiều	Không	Bình thường	Từ chối
O5	MBA	Ít	Biết	Bình thường	Từ chối
O6	MCE	Ít	Biết	Tốt	Từ chối
O7	MSC	Trung bình	Biết	Bình thường	Từ chối
O8	MCE	Ít	Không	Xuất sắc	Từ chối

Bài tập

- 1) Khảo sát sự phụ thuộc thuộc tính giữa B = {Bằng cấp, Lời giới thiệu) và C = {Tuyển dụng} và đề xuất một số phân loại chính xác 100%
- 2) Tính xấp xỉ tập X = {o1, o2 , o3} qua tập thuộc tính B={bằng cấp,kinh nghiệm}
- 3) Tính các reducts