TẬP THÔ

Khai thác dữ liệu

ThS. Nguyễn Hồ Duy Trí
trinhd@uit.edu.vn

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN

2020

NỘI DUNG

- Khái niệm tập thô
- Quan hệ bất khả phân biệt
- Xấp xỉ
- Sự phụ thuộc thuộc tính
- Thành lập ma trận phân biệt
- Thiết lập hàm phân biệt
- Tìm các rút gọn từ hàm phân biệt

Khái niệm

Tập thô (Rough Set)

Zdzisław Pawlak, *Rough sets*, International Journal of Paralled Programming, Vol. 11, No. 5, pp. 341–356, 1982.

Ứng dụng của tập thô

- Khắc phục dữ liệu dư thừa, nhiễu
- Rút gọn dữ liệu (chọn thuộc tính đặc trưng)
- Nhận diện các phụ thuộc (riêng phần, toàn phần) giữa các thuộc tính
- Tạo luật quyết định

Các khái niệm

- Hệ thông tin, bảng quyết định (information system, information table, decision table)
- Quan hệ bất khả phân biệt (indiscernibility relation)
- Xấp xỉ (approximations): xấp xỉ dưới (lower approximation) và xấp xỉ trên (upper approximation)
- Phụ thuộc thuộc tính
- Rút gọn

Hệ thông tin
Information system

- Được biểu diễn ở dạng bảng 2 chiều
- IS là cặp (U, A), với
  - ✓ U: tập các đối tượng
  - ✓ A: tập các thuộc tính

	Headache	Vomiting	Temperature
#1	No	Yes	High
#2	Yes	No	High
#3	Yes	Yes	Very high
#4	No	Yes	Normal
#5	Yes	No	High
#6	No	Yes	Very high

Bảng thông tin
Information table

- Biểu diễn dạng bảng hai chiều có thêm thuộc tính quyết định
- Còn được gọi là hệ quyết định (Decision system)
- IT là cặp (U, A∪{d}), với
  - ✓ d∉A là thuộc tính quyết định
  - ✓ A: tập các thuộc tính (điều kiện)

	Headache	Vomitting	Temperature	*Viral illness*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

Hệ thông tin và bảng thông tin

Dữ liệu có cùng giá trị các thuộc tính điều kiện nhưng có thể khác về giá trị thuộc tính quyết định

	Headache	Vomitting	Temperature	*Viral illness*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

- Dữ liệu có thể trùng nhau, dữ liệu biểu diễn lặp lại các đối tượng giống nhau
- Một số thuộc tính có thể thừa

- Rút ra luật

  Ví dụ: "Nếu **Headache** là ***Yes***, **Vomitting** là ***Yes*** và **Temperature** là ***Very high*** thì **Viral illness** là ***Yes***"

	Headache	Vomitting	Temperature	*Viral illness*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

Quan hệ bất khả phân biệt
Indiscernibility relation

- Các đối tượng có giá trị giống nhau trên một tập thuộc tính được gọi là bất khả phân biệt (không phân biệt được).
- Định nghĩa
  - ✓ Cho IS(U, A) là hệ thông tin, cho $B \subseteq A$
  - ✓ $IND_{IS}(B)=\{(x,y) \in U^2 \mid \forall b \in B, b(x)=b(y)\}$
  - ✓ Khi đó $IND_{IS}(B)$ là quan hệ bất khả phân biệt theo B

- IND_{IS}(Headache)=({#1,#4,#6}, {#2,#3,#5})
- IND_{IS}(Teperature)=({#1,#2,#5}, {#3,#6}, {#4})
- IND_{IS}(Vomitting, Teperature)=({#1}, {#2,#5}, {#3,#6}, {#4})

	Headache	Vomitting	Temperature	*Viral illness*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

- Mỗi đối tượng thuộc về 1 lớp duy nhất
- Các đối tượng trong 1 lớp sẽ có giá trị trên tập thuộc tính bằng nhau

	Headache	Vomitting	Temperature	***Viral illness***
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

- IND_{IS}(Headache)=({#1,#4,#6}, {#2,#3,#5})
  - ✓ #1,#4,#6 có giá trị Headache = "No"
  - ✓ #2,#3,#5 có giá trị Headache = "Yes"

	Headache	Vomitting	Temperature	*Viral illness*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

Xấp xỉ
Approximations

Xét #2 và #5, giá trị giống nhau trên mọi thuộc tính, trừ thuộc tính quyết định. Tuy nhiên, giá trị thuộc tính quyết định lại khác nhau.

	Headache	Vomitting	Temperature	*Viral illness*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

- Dùng phân hoạch để tìm ra các tập con của lớp phổ quát.
- Các tập con có cùng giá trị thuộc tính quyết định.
- Đối với các phần tử trong quan hệ bất khả phân biệt: không thể có một định nghĩa chính xác → Khái niệm xấp xỉ:
  - ✓ Xấp xỉ trên (upper approximation)
  - ✓ Xấp xỉ dưới (lower approximation)
  - ✓ Vùng biên (boundary region)

- Cho IS = (U,A), B $\subseteq$ A và X $\subseteq$ U
- Xấp xỉ X theo B: dựa vào các thuộc tính trong B, xác định một đối tượng của U có thuộc tập X hay không?
- Ví dụ: Cho B = {Vomitting, Temperature}, X = {#1,#2,#3,#6}

  //các đối tượng nhiễm virus, Viral illness = yes

1. Phân lớp U theo B

$$\mathrm{U/B}$$

2. Xấp xỉ dưới

$$\underline{B}X = \{x \in O | [x]_B \in X\}$$

→ Gồm các phần tử **chắc chắn** thuộc X

3. Xấp xỉ trên

$$\overline{B}X = \{x \in O | [x]_B \cap X \neq \emptyset\}$$

→ Gồm các phần tử **có khả năng** được phân loại thuộc X

4. Vùng biên

$$\text{B\_Biên} = \overline{B}X - \underline{B}X$$

→ Gồm các phần tử không thể phân lớp chắc chắn

→ Một tập được gọi là tập thô nếu vùng biên khác rỗng, ngược lại gọi là tập rõ

5. Vùng ngoài

$$\text{B\_Ngoài} = U - \overline{B}X$$

→ Gồm các phần tử **chắc chắn không** thuộc X

Ví dụ

Cho B = {Vomitting, Temperature}, X = {#1,#2,#3,#6} *//Viral illness = yes*

Khi đó:

- **IND_{IS}(Vomitting, Temperature)=({#1}, {#2,#5}, {#3,#6}, {#4})**

$$\underline{B}X = \{\#1, \#3, \#6\}$$

$$\overline{B}X = \{\#1, \#2, \#5, \#3, \#6\}$$

$$B\_Biên = \{\#2, \#5\}$$

$$B\_Ngoài = \{\#4\}$$

Vì vùng biên khác rỗng
Lớp quyết định *Viral illness* là Thô

	Headache	Vomitting	Temperature	*Viral illness*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

Độ chính xác của tập thô

Hệ số

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

- $|X|$ là lực lượng của tập X
- Các xấp xỉ là tập khác trống
- Giá trị $0 \leq \alpha_B \leq 1$
- Nếu $\alpha_B(X) = 1$ X là xấp xỉ rõ theo B
- Nếu $\alpha_B(X) < 1$ X là xấp xỉ thô theo B

Ví dụ

Ở kết quả ví dụ trước, với B={Vomitting, Temperature}, X={#1,#2,#3,#6} // viral illness=yes, ta có

$$\underline{B}X = \{\#1, \#3, \#6\}$$

$$\overline{B}X = \{\#1, \#2, \#5, \#3, \#6\}$$

- Hệ số đo độ chính xác

$$\alpha_B(X) = \frac{|\underline{B}X|}{|\overline{B}X|} = \frac{|\{\#1, \#3, \#6\}|}{|\{\#1, \#2, \#5, \#3, \#6\}|} = \frac{3}{5} = 0.6$$

- Nhận xét

Bài tập nhóm

- Thời gian: 10 phút
- Cho hệ quyết định (slide sau), với
  - ✓ B = {Temp, Windy}
  - ✓ X = {3,4,5,7,9,10,11,12,13} *//tập các đối tượng có giá trị thuộc tính Play = yes*
- Hãy tìm
  - ✓ Xấp xỉ dưới theo B
  - ✓ Xấp xỉ trên theo B
  - ✓ Vùng biên theo B
  - ✓ Vùng ngoài của B

Day	Outlook	Temp	Humidity	Windy	Play
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D3	Overcast	Hot	High	Weak	*Yes*
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D7	Overcast	Cool	Normal	Strong	*Yes*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D10	Rain	Mild	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

Sự phụ thuộc thuộc tính
Dependency of attributes

- Tập thuộc tính D phụ thuộc hoàn toàn vào tập thuộc tính C khi mọi giá trị của các thuộc tính trong D là duy nhất được xác định bởi các giá trị của thuộc tính trong C.
- Kí hiệu C $\Rightarrow$ D

- Xét thuộc tính Temperature và Viral illness:
  - ✓ (Temperature, very high) xác định (Viral illness, Yes)
  - ✓ (Temperature, normal) xác định (Viral illness, No)

	Headache	Vomitting	Temperature	*Viral illness*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

- Xét thuộc tính Temperature và Viral illness:
  - ✓ (Temperature, very high) xác định (Viral illness, Yes)
  - ✓ (Temperature, normal) xác định (Viral illness, No)
  - ✓ (Temperature, high) không luôn luôn xác định (Viral illness, Yes)

	Headache	Vomitting	Temperature	*Viral illness*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

- Cho C, D là các tập con của A. Khi đó D phụ thuộc C với độ phụ thuộc k $(0 \leq k \leq 1)$
  - ✓ $k = 1$: phụ thuộc hoàn toàn
  - ✓ $k < 1$: phụ thuộc một phần
- Kí hiệu: C $\Rightarrow k$D

$$k = \gamma(C, D) = \sum_{X \in U/D} \frac{|\underline{C}(X)|}{|U|}$$

{Headache, Vomitting, Temperature} $\Rightarrow k$ {Viral illness}

$$k = \frac{4}{6} = 0.67$$

	Headache	Vomitting	Temperature	*Viral illness*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

- Tính hệ số k trong các trường hợp sau
  - ✓ {Temperature} $\Rightarrow k$ {Viral illness}
  - ✓ {Headache} $\Rightarrow k$ {Viral illness}
  - ✓ {Vomitting} $\Rightarrow k$ {Viral illness}

Nhận xét

	Headache	Vomitting	Temperature	*Viral illness*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

Các thuộc tính rút gọn
Reduction attributes

- Một số 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 thừa
- Chỉ giữ lại các thuộc tính điều kiện bảo toàn quan hệ bất khả phân biệt và hệ quả là bảo toàn xấp xỉ tập hợp
- Các tập con thuộc tính điều kiện lúc này được gọi là rút gọn (reduct)

Các bước thực hiện

1. Xác định ma trận phân biệt (discernibility matrix)
2. Xác định hàm phân biệt (discernibility function)
3. Rút gọn hàm
4. Luật phân lớp

Xác định ma trận phân biệt

- Với hệ quyết định $(U, C \cup D)$
- Ma trận n x n (n là số đối tượng trong U)
- C_{ij} là giá trị tại dòng i, cột j
  - ✓ Nếu $D(x_i) = D(x_j)$ thì $c_{ij} = \lambda$ (lambda)
  - ✓ Nếu $D(x_i) \neq D(x_j)$ thì $c_{ij} = \{c \in C \mid C(x_i) \neq C(x_j)\}$, với i,j = 1..n

	Headache (H)	Vomitting (V)	Temperature (T)	*Viral illness (VI)*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

	#1	#2	#3	#4	#5	#6
#1	λ					
#2	λ	λ				
#3	λ	λ	λ			
#4	T	H,V,T	H,T	λ		
#5	H,V	∅	V,T	λ	λ	
#6	λ	λ	λ	T	H,V,T	λ

Thiết lập hàm phân biệt

Quy ước

- λ: `True`
- Trong 1 ô: phép toán `or` (∨)
- Giữa các ô: phép toán `and` (∧)
- Hàm phân biệt f(H,V,T) =

`True` ∧ `True` ∧ `True` ∧ `True` ∧ `True` ∧ `True` ∧ //dòng 1, 2, 3

T ∧ (H ∨ V ∨ T) ∧ (H ∨ T) ∧ `True` ∧ //dòng 4

(H ∨ V) ∧ (V ∨ T) ∧ `True` ∧ `True` ∧ //dòng 5

`True` ∧ `True` ∧ `True` ∧ T ∧ (H ∨ V ∨ T) ∧ `True` //dòng 6

Tìm các rút gọn

- Rút gọn hàm phân biệt
  1. (A) ∧ (`True`) = (A)
  2. (A) ∧ (A ∨ B) = (A)
  3. (A) ∨ (A ∧ B) = (A)
- Hàm phân biệt f(H,V,T) =

`True` ∧ `True` ∧ `True` ∧ `True` ∧ `True` ∧ `True` ∧
T ∧ (H ∨ V ∨ T) ∧ (H ∨ T) ∧ `True` ∧
(H ∨ V) ∧ (V ∨ T) ∧ `True` ∧ `True` ∧
`True` ∧ `True` ∧ `True` ∧ T ∧ (H ∨ V ∨ T) ∧ `True`
= T ∧ (H ∨ V)

- Hàm phân biệt f(H,V,T) = T ∧ (H ∨ V)
- Vậy có hai thu gọn (reduct) là
  - ✓ R1 = T∧H //Temperature và Headache
  - ✓ R2 = T∧V //Temperature và Vomitting
- Lõi (core) là R1 ∩ R2 = {T}: là thuộc tính bắt buộc phải có trong phân lớp dữ liệu

Luật phân lớp

Các bước

1. Phân lớp U theo D: $U/D = \{X_1, X_2\}$
2. Phân lớp U theo từng tập rút gọn: $R_1,...R_k$
   - ✓ R_1: $U/R_1 = \{Z_1, Z_2,..., Z_n\}$
   - ✓ R_2: $U/R_2 = \{Z_{n+1}, Z_{n+2},...Z_m\}$
   - ✓ ...
3. Nếu thấy $Z_i \subseteq X_j$ thì có luật phân lớp $Z_i \rightarrow X_j$ với $Z_i \subseteq X_j$

Ví dụ

1. Phân lớp U theo D: U/D = $\{X_1, X_2\}$

   ✓ $X_1 = \{x \in U \mid d(x) = Yes\} = \{\#1, \#2, \#3, \#6\}$

   ✓ $X_2 = \{x \in U \mid d(x) = No\} = \{\#4, \#5\}$

	Headache	Vomitting	Temperature	*Viral illness*
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

2. Phân lớp U theo R_1={Temperature, Headache}, R_2={Temperature, Vomitting}

- ✓ U/R_1: Z_1 = {#1}, Z_2 = {#2,#5}, Z_3 = {#3}, Z_4 = {#4}, Z_5 = {#6}
- ✓ U/R_2: Z_6 = {#1}, Z_7 = {#2,#5}, Z_8 = {#3,#6}, Z_9 = {#4}

	Headache	Vomitting	Temperature	***Viral illness***
#1	No	Yes	High	*Yes*
#2	Yes	No	High	*Yes*
#3	Yes	Yes	Very high	*Yes*
#4	No	Yes	Normal	*No*
#5	Yes	No	High	*No*
#6	No	Yes	Very high	*Yes*

1. Phân lớp U theo D: U/D = $\{X_1, X_2\}$
   - ✓ $X_1 = \{x \in U \mid d(x) = Yes\} = \{\#1, \#2, \#3, \#6\}$
   - ✓ $X_2 = \{x \in U \mid d(x) = No\} = \{\#4, \#5\}$
2. Phân lớp U theo R_1={Temperature, Headache}, R_2={Temperature, Vomitting}
   - ✓ U/R_1: $Z_1 = \{\#1\}$, $Z_2 = \{\#2,\#5\}$, $Z_3 = \{\#3\}$, $Z_4 = \{\#4\}$, $Z_5 = \{\#6\}$
   - ✓ U/R_2: $Z_6 = \{\#1\}$, $Z_7 = \{\#2,\#5\}$, $Z_8 = \{\#3,\#6\}$, $Z_9 = \{\#4\}$
3. Rút luật
   - ✓ Vì $Z_1 \subseteq X_1$ nên có Z_1 → X_1: nếu Temperature="High" và Headache="No" thì Viral illeness="Yes"
   - ✓ Vì $Z_8 \subseteq X_1$ nên có Z_8 → X_1: nếu Temperature="Very high" và Vomitting="Yes" thì Viral illeness="Yes"

Bài tập nhóm

- Thời gian: 10 phút
- Cho hệ quyết định (slide sau)
- Hãy tìm
  - ✓ Ma trận phân biệt
  - ✓ Hàm phân biệt
  - ✓ Rút gọn hàm
  - ✓ Đưa ra các luật

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

Q & A