Wnioskowanie Boolowskie i teoria zbiorów przybliżonych

4. **Zbiory przybliżone**
   - Wprowadzenie do teorii zbiorów przybliżonych
   - Złożoność problemu szukania reduktów
5. **Wnioskowanie Boolowskie w obliczaniu reduktów i reguł decyzyjnych**
   - Metody wnioskowań Boolowskich w szukaniu reduktów
   - Systemy decyzyjne oparte o zbiory przybliżone
6. **Metoda drzew decyzyjnych**
   - Wprowadzenie
   - Konstrukcja drzew decyzyjnych
7. **Problem dyskretyzacji**
   - Przypomnienia podstawowych pojęć
   - Problem dyskretyzacji
   - Dyskretyzacja metodą wnioskowania Boolowskiego

Teoria zbiorów przybliżonych

- Teoria zbiorów przybliżonych została wprowadzona w latach 80-tych przez prof. Zdzisława Pawlaka.
- Głównym celem jest dostarczanie narzędzi dla problemu aproksymacji pojęć (zbiorów).
- Zastosowania w systemach decyzyjnych:
  - Redukcja danych, selekcja ważnych atrybutów;
  - Generowanie reguł decyzyjnych;
  - Odkrywanie wzorców z danych: szablony, reguły asocjacyjne;
  - Odkrywanie zależności w danych.

Systemy informacyjne

Definicja

Jest to para $\mathbb{S} = (U, A)$, gdzie

- U – skończony niepusty zbiór obiektów (ang. cases, states, patients, observations ...);
- A – skończony, niepusty zbiór atrybutów. Każdy $a \in A$ odpowiada pewnej funkcji $a : U \to V_a$ zwanej *wartościowaniem*, gdzie V_a jest nazwana dziedziną atrybutu a.

Dla $B \subseteq A$, definiujemy

- *B-sygnatura* obiektu $x \in U$ (ang. **B-information vector**) jako

$$inf_B(x) = \{(a, a(x)) : a \in B\}$$

- Zbiór sygnatur względem B o obiektach z U (ang. **B-information set**):

$$INF(\mathbb{S}) = \{inf_B(x) : x \in U\}$$

Tablica decyzyjna

Tablica decyzyjna powstaje ze zwykłych tablic danych poprzez sprecyzowanie:

- **Atrybutów (nazwanych warunkowymi)**: cechy, których wartości na obiektach są dostępne, np. pomiary, parametry, dane osobowe, ...
- **Decyzji (atrybut decyzyjny):**, t.j. cecha "ukryta" związana z pewną znaną częściowo wiedzą o pewnym pojęciu:
  - Decyzja jest znana tylko dla obiektów z (treningowej) tablicy decyzyjnej;
  - Jest podana przez eksperta (np. lekarza) lub na podstawie późniejszych obserwacji (np. ocena giełdy);
  - Chcemy podać metodę jej wyznaczania dla dowolnych obiektów na podstawie wartości atrybutów warunkowych na tych obiektach.

Przykład

Przedstawiona tablica decyzyjna zawiera:

- 8 obiektów będących opisami pacjentów
- 3 atrybuty: Headache Muscle pain, Temp.
- Decyzję stwierdzącą czy pacjent jest przeziębiony czy też nie. lub nie

Example

U	Ból głowy	Ból mięśni	Temp.	Grypa
p1	Tak	Tak	N	Nie
p2	Tak	Tak	H	Tak
p3	Tak	Tak	VH	Tak
p4	Nie	Tak	N	Nie
p5	Nie	Nie	H	Nie
p6	Nie	Tak	VH	Tak
p7	Nie	Tak	H	Tak
p8	Nie	Nie	VH	Nie

Relacja rozróżnialności

Dane są obiekty $x, y \in U$ i zbiór atrybutów $B \subset A$, mówimy, że

- x, y są **rozróżnialne przez** B wtw, gdy istnieje $a \in B$ taki, że $a(x) \neq a(y)$;
- x, y są **nierozróżnialne przez** B, jeśli one są identyczne na B, tzn. $a(x) = a(y)$ dla każdego $a \in B$;
- $[x]_B$ = zbiór obiektów nierozróżnialnych z x przez B.

Relacja rozróżnialności

- Dla każdych obiektów x, y:
  - albo $[x]_B = [y]_B$;
  - albo $[x]_B \cap [y]_B = \emptyset$.
- Relacja

$$x \; IND_B \; y := x, y \text{ są nierozróżnialne przez } B$$

  jest relacją równoważności.
- Każdy zbiór atrybutów $B \subset A$ wyznacza podział zbioru obiektów na klasy nierozróżnialności.

Przykład

Dla $B = \{Blgowy, Blmini\}$

- obiekty $p1, p2, p3$ są nierozróżnialne;
- są 3 klasy nierozróżnialności relacji IND_B:
  - $[p1]_B = \{p1, p2, p3\}$
  - $[p4]_B = \{p4, p6, p7\}$
  - $[p5]_B = \{p5, p8\}$

Example

U	Ból głowy	Ból mięśni	Temp.	Grypa
p1	Tak	Tak	N	Nie
p2	Tak	Tak	H	Tak
p3	Tak	Tak	VH	Tak
p4	Nie	Tak	N	Nie
p5	Nie	Nie	H	Nie
p6	Nie	Tak	VH	Tak
p7	Nie	Tak	H	Tak
p8	Nie	Nie	VH	Nie

Relacja rozróżnialności i aproksymacja pojęć

- Każdy zbiór obiektów X (np. klasa decyzyjna, pojęcie) może być opisany za pomocą atrybutów ze zbioru B *dokładnie* lub *w przybliżeniu*
  - *dokładny opis:* jeśli X jest sumą pewnych klas nierozróżnialności definiowanych przez B (ZBIORY DOKŁADNE)
  - *przybliżony opis:* w przeciwnym przypadku (ZBIORY PRZYBLIŻONE)
- W obu przypadkach X może być opisany przez 2 "dokładne zbiory" zwane dolną i górną aproksymacją zbioru X

$$\underline{B}(X) = \{x : [x]_B \subset X\} \quad \overline{B}(X) = \{x : [x]_B \cap X \neq \emptyset\}$$

Aproksymacja pojęć

- Obszar brzegowy (ang. B-boundary region) pojęcia X zawiera obiekty, dla których nie możemy jednoznacznie zdecydować czy należą one do X czy nie na podstawie atrybutów z B
- Obszar wewnętrzny (ang. B-inside region of X) zawiera obiekty, które możemy pewnie klasyfikować jako elementy pojęcia X mając do dyspozycji atrybuty z B.
- Zbiór jest przybliżony (ang. rough set) jeśli obszar brzegowy jest niepusty, w przeciwnym przypadku zbiór jest nazwany dokładny (ang. crisp set).

Przykład

Niech $B = \{a_1, a_2\}$

$$IND(B) = \{\{1,2\}, \quad (sunny, hot)$$
$$\{3,13\}, \quad (overcast, hot)$$
$$\{4,10,14\}, \quad (rainy, mild)$$
$$\{5,6\}, \quad (rainy, cool)$$
$$\{8,11\}, \quad (sunny, mild)$$
$$\{7\},\{9\},\{12\}\}$$

Chcemy aproksymować pojęcie definiowane przez klasę decyzyjną (play=no)

$$X = CLASS_{no} = \{1,2,6,8,14\}$$

Aproksymacje pojęcia X:

$$\mathbf{L}_B(X) = \{1,2\}$$
$$\mathbf{U}_B(X) = \{1,2,5,6,8,11,4,10,14\}$$

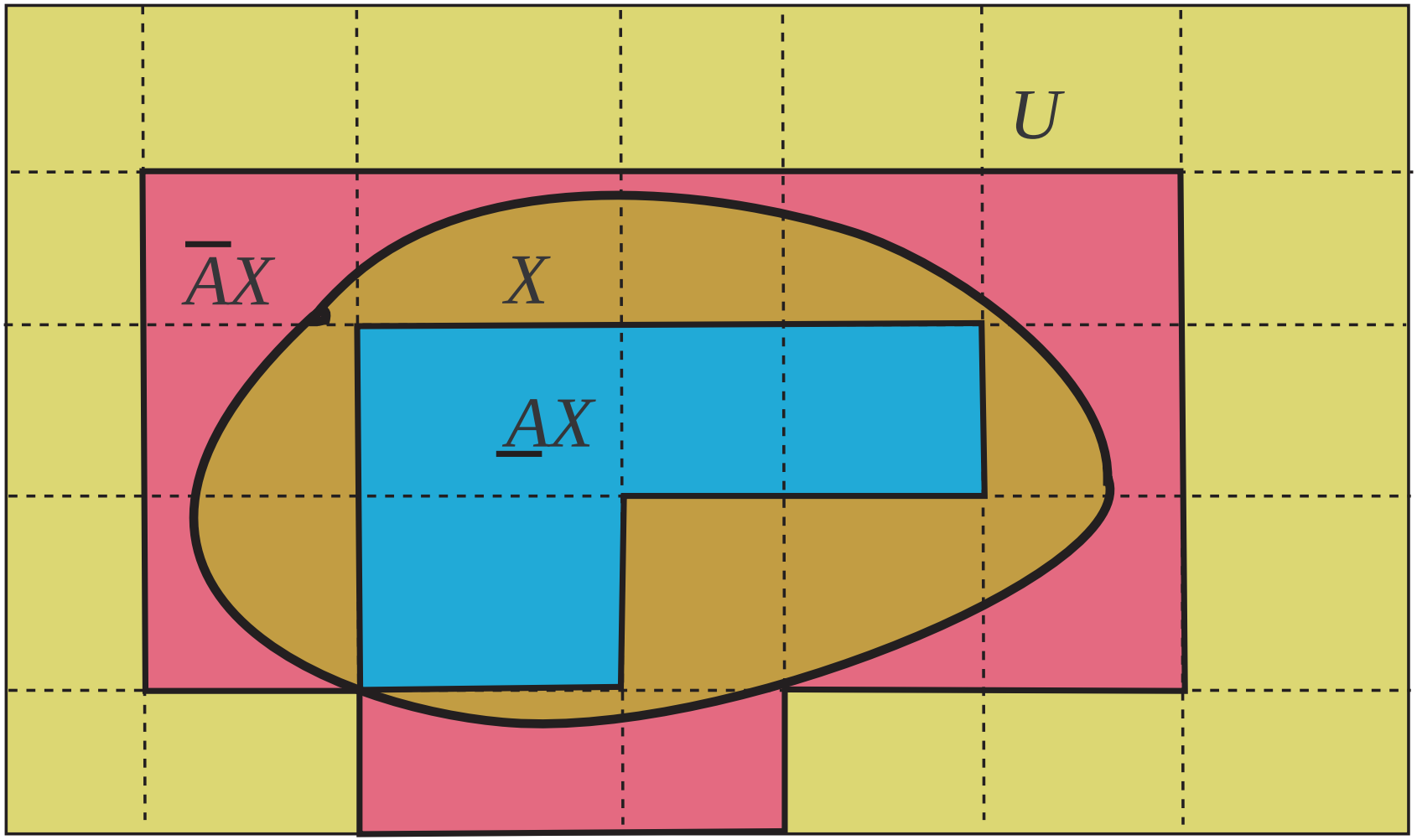
Reguła pewna:
If $B(x) = (sunny, hot)$ **then** $d(x) = no$

$\mathbb{A}$	a_1	a_2	a_3	a_4	d
ID	outlook	temp.	hum.	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

Reguły przybliżone:

If $B(x) = (rainy, cool)$ **then** $d(x) = no$
If $B(x) = (rainy, mild)$ **then** $d(x) = no$
If $B(x) = (sunny, mild)$ **then** $d(x) = no$

Pączek



Jakość aproksymacji

Jakość aproksymacji (ang. accuracy of approximation)

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

- $0 \leq \alpha_B(X) \leq 1$
- Jeśli $\alpha_B(X) = 1$, to zbiór X jest dokładnie definiowany przez B;
- Jeśli $\alpha_B(X) < 1$, to zbiór X jest aproksymacyjnie definiowany przez B;

Motywacje

- W systemach decyzyjnych, nie wszystkie atrybuty są potrzebne do procesie podejmowania decyzji;
- Chcemy wybrać pewne podzbiory atrybutów niezbędnych do tego celu;
- Redukty to minimalne podzbiory atrybutów zachowujących charakterystykę całego zbioru atrybutów.
- W teorii zbiorów przybliżonych, istnieją co najmniej 2 pojęcia reduktów: informacyjne i decyzyjne.

Definicja

Definicja reduktu informacyjnego

Zbiór atrybutów $B \subset A$ nazywamy reduktem tablicy decyzyjnej A wtw, gdy

- B zachowuje rozróżnialność zbioru A:
  t.j. dla dowolnych obiektów $x, y \in U$,
  jeśli x, y są rozróżnialne przez A, to są również rozróżnialne przez B
- B jest niezredukowalny:
  tzn. żaden właściwy podzbiór B nie zachowuje rozróżnialności zbioru A (t.j., B jest minimalny pod względem rozróżnialności)

Definicja

Definicja reduktu decyzyjnego

Zbiór atrybutów $B \subset A$ nazywamy reduktem tablicy A wtw, gdy

- B zachowuje rozróżnialność zbioru A względem decyzji dec:
  t.j. dla dowolnych obiektów $x, y \in U$,
  jeśli $dec(x) \neq dec(y)$ i x, y są rozróżnialne przez A, to są również rozróżnialne przez B
- B jest niezredukowalny:
  tzn. żaden właściwy podzbiór B nie zachowuje rozróżnialności zbioru A (B jest minimalny pod względem rozróżnialności)

Zbiór reduktów

- $RED(\mathbb{S})$ = zbiór wszystkich reduktów tablicy decyzyjnej $\mathbb{S}$;
- Jeśli $\mathbb{S} = (U, A \cup \{dec\})$ i $|A| = n$ to

$$|RED(\mathbb{S})| \leq \binom{n}{n/2}$$

- **rdzeń:** zbiór atrybutów będących we wszystkich reduktach

$$K = \bigcap_{B \in RED(\mathbb{S})} B.$$

Problemy obliczeniowe związane z reduktami

- Znaleźć rdzeń danej tablicy decyzyjnej;
- Znaleźć jakiś redukt;
- Znaleźć krótkie redukty;
- Znaleźć długie redukty.

Sformułowanie problemu

Problem najkrótszego reduktu

Dane: Tablica decyzyjna $\mathbb{S} = (U, A \cup \{dec\})$;

Szukane: "najkrótszy redukt tablicy decyzyjnej $\mathbb{S}$", tzn. taki redukt decyzyjny $B \in \mathbf{RED}(\mathbb{S}, dec)$, że

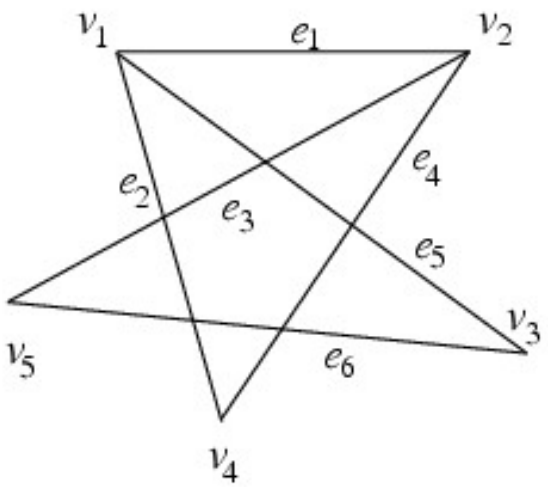
$$\forall_{X \in \mathbf{RED}(\mathbb{S}, dec)} |B| \leq |X|$$

Twierdzenie

Problem szukania najkrótszego reduktu jest NP-zupełny.

Idea dowodu

- **Ogólnie musimy pokazać, że** jakiś znany NP-zupełny problem jest "łatwiejszy" od problemu najkrótszego reduktu;
- Wybierzmy problem minimalnego pokrycia wierzchołkami: "Dany jest graf $G = (V, E)$. Znaleźć minimalny zbiór wierzchołków $X \subset V$ o takiej własności, że każda krawędź z E posiada co najmniej jeden z końców w X."
- Wielomianowa transformacja:



$\Longrightarrow$

$\mathbb{S}(\mathbf{G})$	a_{v_1}	a_{v_2}	a_{v_3}	a_{v_4}	a_{v_5}	a^*
x^*	0	0	0	0	0	0
u_{e_1}	1	1	0	0	0	1
u_{e_2}	1	0	0	1	0	1
u_{e_3}	0	1	0	0	1	1
u_{e_4}	0	1	0	1	0	1
u_{e_5}	1	0	1	0	0	1
u_{e_6}	0	0	1	0	1	1