Systemy ekspertowe

Część trzecia

Eksploracja danych z wykorzystaniem tablic decyzyjnych i zbiorów przybliżonych

Autor
Roman Simiński

Kontakt
siminski@us.edu.pl
www.us.edu.pl/~siminski

System informacyjny

Definicja

System informacyjny *SI* zdefiniowany jest jako dwójka:

$$SI = (U, A)$$

gdzie

- U jest niepustym, skończonym zbiorem obiektów,
- A jest niepustym, skończonym zbiorem atrybutów.

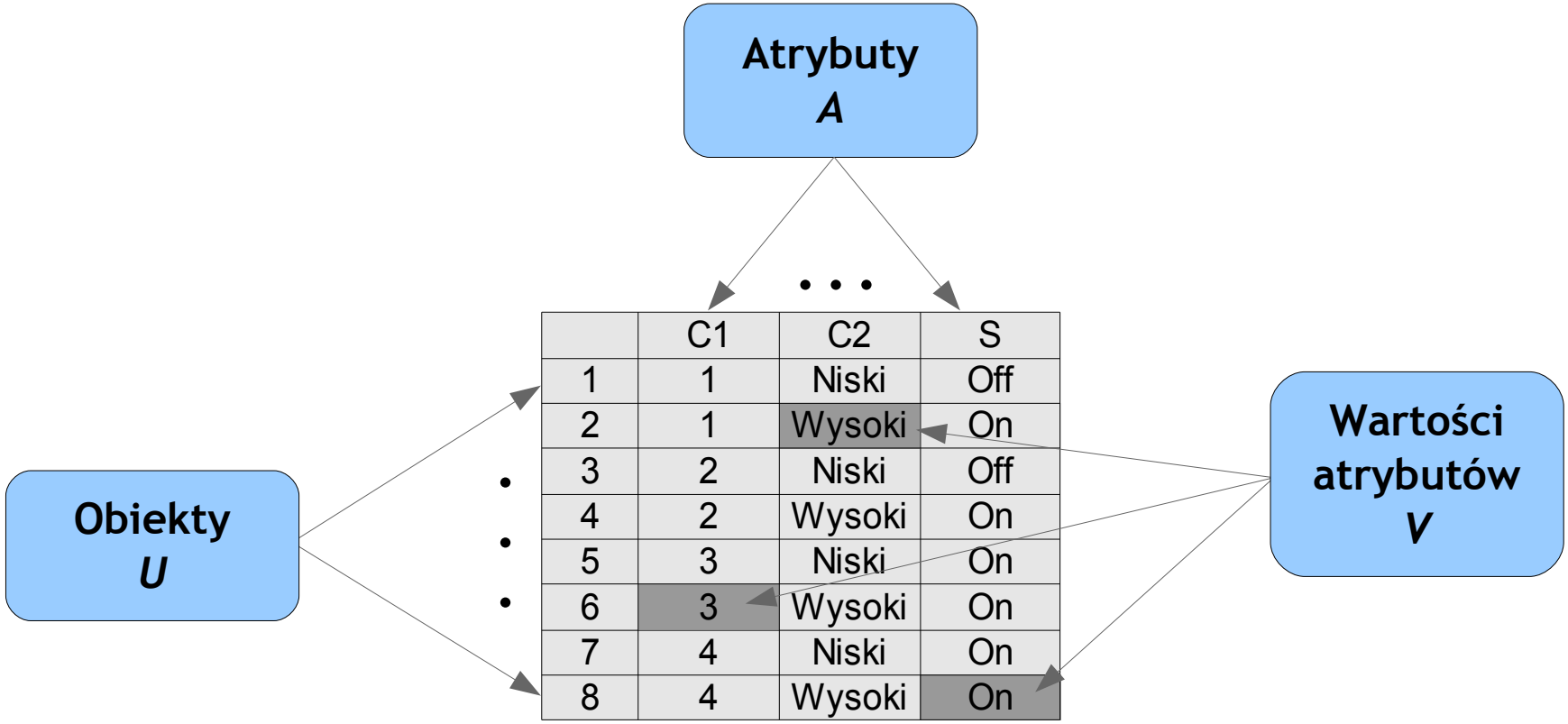
Zbiór V_a jest dziedziną atrybutu $a \in A$. $V = \bigcup_{a \in A} V_a$.

Definiuje się również funkcję informacyjną

f: $U \times A \rightarrow V$, taką, że $\forall a \in A, x \in U \; f(a, x) \in \mathrm{V_a}$.

System informacyjny

Jak należy rozumieć definicję SI?



	C1	C2	S
1	1	Niski	Off
2	1	Wysoki	On
3	2	Niski	Off
4	2	Wysoki	On
5	3	Niski	On
6	3	Wysoki	On
7	4	Niski	On
8	4	Wysoki	On

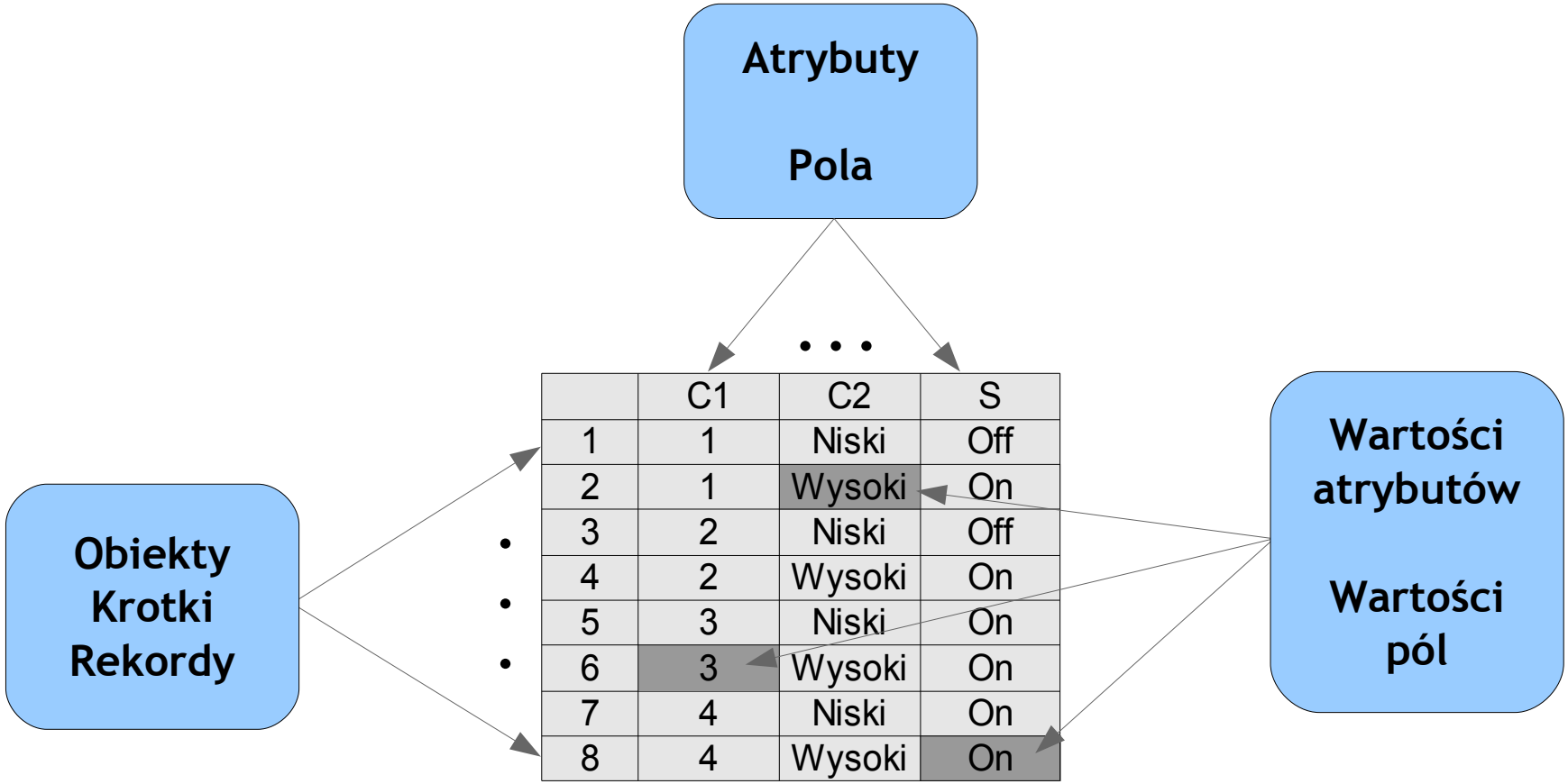
f(C2, 1) = Niski f(C1, 4) = 2 f(S, 3) = Off

f(C2, 2) = Wysoki f(C1, 2) = 1 f(S, 7) = On

$f: U \times A \rightarrow V : \forall a \in A, x \in U \; f(a, x) \in V_a$, gdzie V_a jest dziedziną atrybutu $a \in A$.

System informacyjny

System informacyjny a tabela bazy danych?



	C1	C2	S
1	1	Niski	Off
2	1	Wysoki	On
3	2	Niski	Off
4	2	Wysoki	On
5	3	Niski	On
6	3	Wysoki	On
7	4	Niski	On
8	4	Wysoki	On

Pojęcie systemu informacyjnego odpowiada pojęciowo pojęciu tabeli (relacji) w bazach danych.

Tablica decyzyjna

Definicja

Tablicą decyzyjną DT nazywać będziemy system informacyjny w postaci:

$$DT = (U, A \cup \{d\})$$

gdzie $d \notin A$ jest atrybutem decyzyjnym niezaliczanym do zbioru atrybutów A systemu. Atrybuty $a \in A$ nazywamy atrybutami warunkowymi.

	C1	C2	S
1	1	Niski	Off
2	1	Wysoki	On
3	2	Niski	Off
4	2	Wysoki	On
5	3	Niski	On
6	3	Wysoki	On
7	4	Niski	On
8	4	Wysoki	On

Atrybuty warunkowe

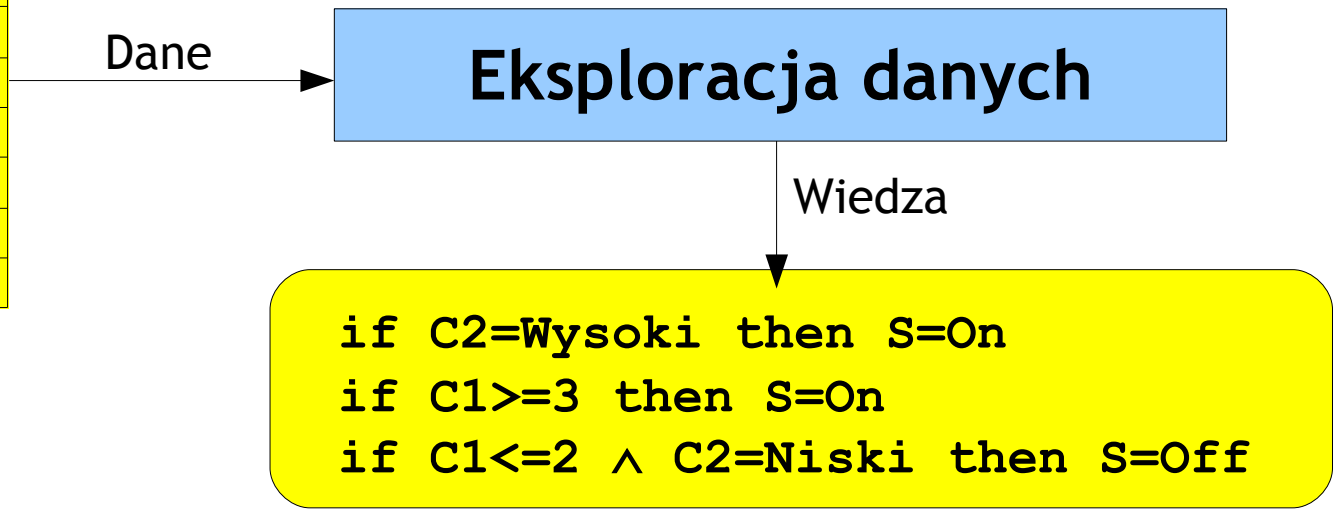
Atrybut decyzyjny

Tablica decyzyjna

Tablica decyzyjna a eksploracja danych

- Celem eksploracji danych będzie odkrycie reguł opisujących wiedzę zapisaną niejawnie w tablicy decyzyjnej.
- Reguły będą opisywały zależności pomiędzy atrybutami warunkowymi i ich wartościami a atrybutem decyzyjnym.
- Chcemy otrzymać *minimalną liczbę jak najprostszych reguł*. Minimalizujemy zatem ich liczbę oraz liczbę atrybutów warunkowych występujących w przesłankach reguł.

	C1	C2	S
1	1	Niski	Off
2	1	Wysoki	On
3	2	Niski	Off
4	2	Wysoki	On
5	3	Niski	On
6	3	Wysoki	On
7	4	Niski	On
8	4	Wysoki	On



Zanim poznamy relację nierozróżnialności...

Powtórka - iloczyn kartezjański

Iloczyn kartezjański zbiorów A i B to zbiór wszystkich par uporządkowanych (a, b), takich, że a należy do zbioru A, zaś b należy do zbioru B. Oznacza się go symbolem $A \times B$. Formalnie:

$$A \times B = \{(a, b) : a \in A, b \in B\}$$

Iloczyn kartezjański może być zbudowany na tym samym zbiorze, np. $A \times A$, co bywa oznaczane A^2. Formalnie:

$$A \times A = \{(a, b) : a \in A, b \in A\}$$

Iloczyn kartezjański dla zbioru obiektów U tablicy decyzyjnej DT

	A			
U	a	b	c	d
1	**1**	**0**	**2**	**2**
2	**0**	**1**	**1**	**2**
3	**0**	**0**	**2**	**2**
4	**2**	**2**	**1**	**0**

Iloczyn kartezjański $U \times U$ to zbiór par obiektów.

$$U \times U = \{(x, y) : x \in U, y \in U\}$$

$$U \times U = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 2), (2, 3), (2, 4), (3, 3), (3, 4), (4, 4)\}$$

Zanim poznamy relację nierozróżnialności...

Powtórka - relacja

- *Relacja* pomiędzy elementami zbioru A a elementami zbioru B to wybrany podzbiór iloczynu kartezjańskiego $A \times B$. Relację tworzą pary elementów wybrane z iloczynu kartezjańskiego według pewnego kryterium.
- W praktyce najpopularniejsze i najszerzej stosowane są *relacje dwuargumentowe* (dwuczłonowe, binarne), zwykle nazywane po prostu relacjami.
- Jeśli założymy, że relacja nazywa się np. R, to zapis $x\ R\ y$ oznacza, że *x jest w relacji R z y*.

Relacja nierozróżnialności

Definicja

Niech $SI = (U, A)$ będzie systemem informacyjnym i niech $B \subseteq A$.

W zbiorze U definiujemy dwuargumentową relację $IND(B)$, generowaną przez zbiór B, zwaną relacją *nierozróżnialności* (ang. *indescernibility relation*):

$$IND(B) = \{(x, y) \in U \times U : \forall a \in B, a(x) = a(y)\}$$

gdzie znak „=” między $a(x)$ i $a(y)$ należy rozumieć w ten sposób, że dla obiektów x i y, należących do U, atrybut a przyjmuje taką samą wartość.

Zapis w postaci:

$$x \; IND(B) \; y$$

Oznacza, że x jest w relacji $IND(B)$ z y.

Mówiąc konkretnie: obiekt x systemu informacyjnego SI jest nierozróżnialny od obiektu y tegż samego systemu, ze względu na wybrany podzbiór atrybutów B.

Relacja nierozróżnialności

Przykład

U	A			
	a	b	c	d
1	**1**	**0**	**2**	**2**
2	**0**	**1**	**1**	**2**
3	**0**	**0**	**2**	**2**
4	**2**	**2**	**1**	**0**

Dla *relacji nierozróżnialności* kryterium wyboru podzbioru $U \times U$ jest *nieodróżnialność obiektów* według atrybutów pochodzących z $B \subseteq A$:

$$\forall a \in B, a(x) = a(y).$$

Iloczyn kartezjański:

$$U \times U = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 2), (2, 3), (2, 4), (3, 3), (3, 4), (4, 4)\}$$

Różne wersje relacji nierozróżnialności, uzależnione od zawartości zbioru B:

Dla $B = \{a\}$: $IND(B) = \{(1, 1), (2, 2), (2, 3), (3, 3), (4, 4)\}$

Dla $B = \{c\}$: $IND(B) = \{(1, 1), (1, 3), (2, 2), (2, 4), (3, 3), (4, 4)\}$

Dla $B = \{b, c\}$: $IND(B) = \{(1, 1), (1, 3), (2, 2), (3, 3), (4, 4)\}$

Relacja nierozróżnialności

Własność

Niech $SI = (U, A)$ będzie systemem informacyjnym i niech $B \subseteq A$.

Relacja nierozróżnialności $IND(B)$ jest relacją równoważnościową, spełniającą następujące warunki:

1. Jest relacją zwrotną

$$\forall x \in U \; (x \; IND \; (B) \; x)$$

2. Jest relacją symetryczną

$$\forall x, y \in U \; (x \; IND \; (B) \; y) \Leftrightarrow (x \; IND \; (B) \; y)$$

3. Jest relacją przechodnią

$$\forall x, y, z \in U \; ((x \; IND \; (B) \; y) \wedge (y \; IND \; (B) \; z)) \Rightarrow (x \; IND \; (B) \; z)$$

Zapis $x \; IND \; (B) \; y$ oznacza, że x jest w relacji $IND(B)$ z y, czyli $\forall a \in B, a(x) = a(y)$.

Relacja nierozróżnialności

Dowód, IND(B) jest relacją zwrotną

Niech $SI = (U, A)$ będzie systemem informacyjnym i niech $B \subseteq A$.

Relacja $IND(B)$ jest *zwrotną*, bo:

Weźmy dowolny obiekt $x \in U$, mamy więc:

$$\forall a \in B,\ a(x) = a(x)$$

a więc z definicji:

$$(x, x) \in IND(B).$$

Relacja nierozróżnialności

Dowód, IND(B) jest relacją symetryczną

Niech $SI = (U, A)$ będzie systemem informacyjnym i niech $B \subseteq A$.

Relacja $IND(B)$ jest *symetryczną*, bo:

Weźmy dowolne obiekt $x, y \in U$, załóżmy, że $(x, y) \in IND(B)$, mamy wtedy:

$$\forall a \in B, a(x) = a(y)$$

stąd:

$$\forall a \in B, a(y) = a(x)$$

a więc:

$$(y, x) \in IND(B).$$

Relacja nierozróżnialności

Dowód, IND(B) jest relacją przechodnią

Niech $SI = (U, A)$ będzie systemem informacyjnym i niech $B \subseteq A$.

Relacja $IND(B)$ jest *przechodnią*, bo:

Weźmy dowolne obiekt $x, y, z \in U$, załóżmy, że $(x, y) \in IND(B)$ oraz $(y, z) \in IND(B)$, mamy wtedy:

$$\forall a \in B, (a(x) = a(y) \wedge a(y) = a(z))$$

stąd:

$$\forall a \in B, a(x) = a(z)$$

a więc:

$$(x, z) \in IND(B).$$

Relacja nierozróżnialności

Klasy abstrakcji

Relacja nierozróżnialności $IND(B)$ będąc relacją równoważnościową, dzieli zbiór obiektów U na rozłączne, niepuste klasy abstrakcji.

- Klasy abstrakcji relacji nierozróżnialności $IND(B)$ oznacza się $U/IND(B)$.
- Każda klasa abstrakcji relacji nierozróżnialności $IND(B)$ to zbiór obiektów nierozróżnialnych ze względu na atrybuty ze zbioru B.
- Klasy abstrakcji $U/IND(B)$ relacji nierozróżnialności $IND(B)$ to zatem zbiór zbiorów takich obiektów, które są nierozróżnialne ze względu na atrybuty ze zbioru B.
- Klasa abstrakcji dla obiektu $x \in U$ relacji $IND(B)$ zdefiniowana jest następująco:

$$[x]_{IND(B)} = \{y \in U, \quad \forall a \in B, \ (a(x) = a(y)\ )\ \}$$

Relacja nierozróżnialności

Klasy abstrakcji — przykładowa tablica decyzyjna

U	A			
	a	b	c	d
1	**1**	**0**	**2**	**2**
2	**0**	**1**	**1**	**2**
3	**0**	**0**	**2**	**2**
4	**2**	**2**	**1**	**0**

Dla $B = \{a\}$:

$U/IND(B) = \{\{1\}, \{2, 3\}, \{4\}\}$

Dla $B = \{b, c\}$:

$U/IND(B) = \{\{1, 3\}, \{2\}, \{4\}\}$

Dla $B = \{a, b, c\}$:

$U/IND(B) = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

Dla $B = \{c\}$ i $x=1$:

$[x]_{IND(B)} = \{1, 3\}$

Dla $B = \{c\}$ i $x=2$:

$[x]_{IND(B)} = \{2, 4\}$

Dla $B = \{a, b\}$ i $x=1$:

$[x]_{IND(B)} = \{1\}$

Relacja nierozróżnialności

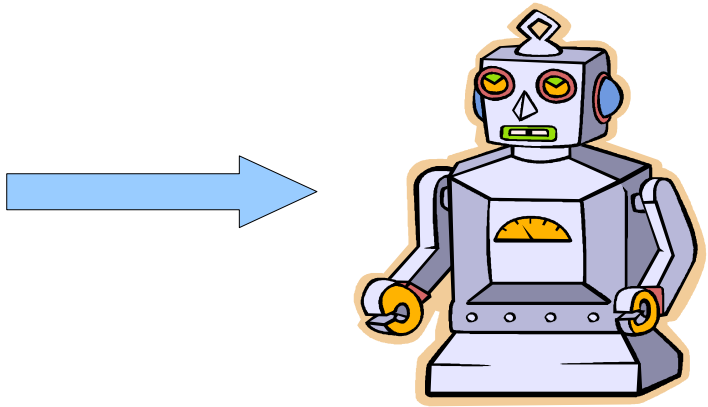
Klasy abstrakcji — tablica decyzyjna z życia wzięta

- Robot ma zebrać dojrzałe jabłka. Rozpoznaje tylko ich wielkość i kolor.
- Tylko na podstawie tych dwóch cech ma ocenić czy jabłko jest dojrzałe czy nie.
- Aby nauczyć robota rozpoznawania jabłek dojrzałych sporządzono tablicę opisującą różne przypadki jabłek dojrzałych i niedojrzałych.

Relacja nierozróżnialności

Klasy abstrakcji — tablica decyzyjna z życia wzięta

	kolor	*wielkość*	*dojrzałe*
x_1	*czerwone*	*duże*	*tak*
x_2	*żółte*	*średnie*	*tak*
x_3	*zielone*	*małe*	*nie*
x_4	*zielone*	*duże*	*tak*
x_5	*żółte*	*średnie*	*nie*
x_6	*czerwone*	*średnie*	*tak*
x_7	*żółte*	*duże*	*tak*
x_8	*czerwone*	*średnie*	*tak*
x_9	*żółte*	*małe*	*nie*
x_{10}	*żółte*	*małe*	*tak*
x_{11}	*czerwone*	*małe*	*tak*
x_{12}	*zielone*	*średnie*	*nie*

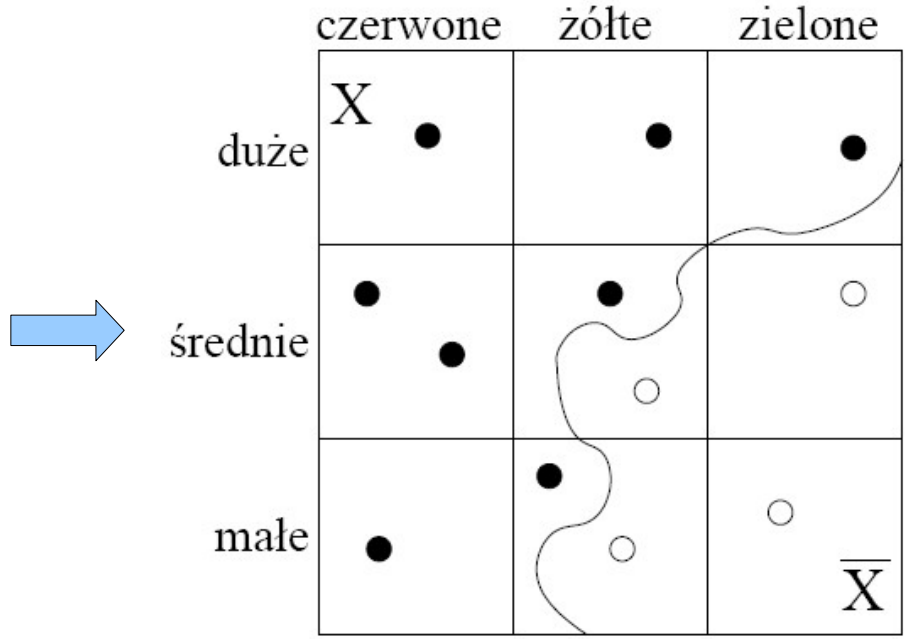


Na podstawie tej tablicy należy nauczyć robota rozpoznawania jabłek dojrzałych i niedojrzałych, ale *tylko na podstawie koloru i wielkości*.

Relacja nierozróżnialności

Klasy abstrakcji — tablica decyzyjna z życia wzięta

	kolor	*wielkość*	*dojrzałe*
x_1	*czerwone*	*duże*	*tak*
x_2	*żółte*	*średnie*	*tak*
x_3	*zielone*	*małe*	*nie*
x_4	*zielone*	*duże*	*tak*
x_5	*żółte*	*średnie*	*nie*
x_6	*czerwone*	*średnie*	*tak*
x_7	*żółte*	*duże*	*tak*
x_8	*czerwone*	*średnie*	*tak*
x_9	*żółte*	*małe*	*nie*
x_{10}	*żółte*	*małe*	*tak*
x_{11}	*czerwone*	*małe*	*tak*
x_{12}	*zielone*	*średnie*	*nie*



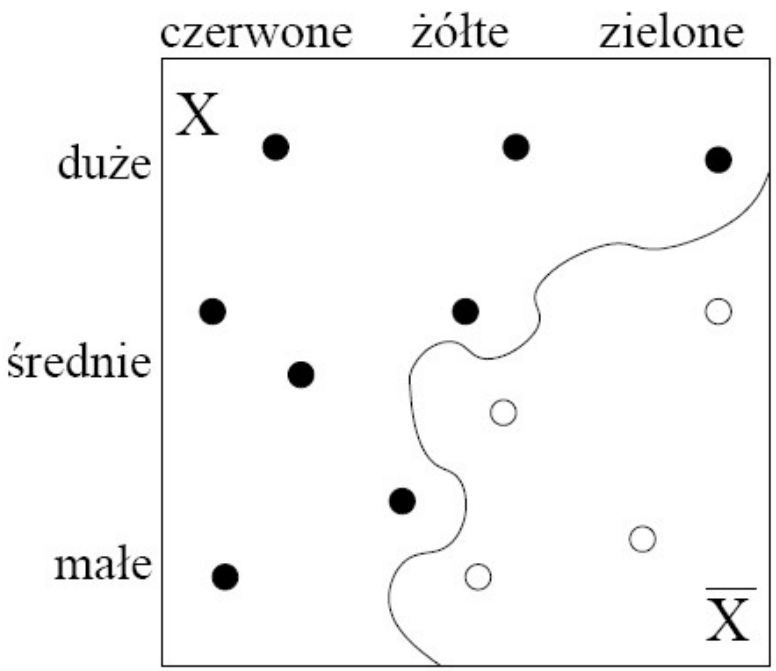
Klasy abstrakcji relacji nierozróżnialności $IND(B)$, dla B={*kolor*, *wielkość*}.

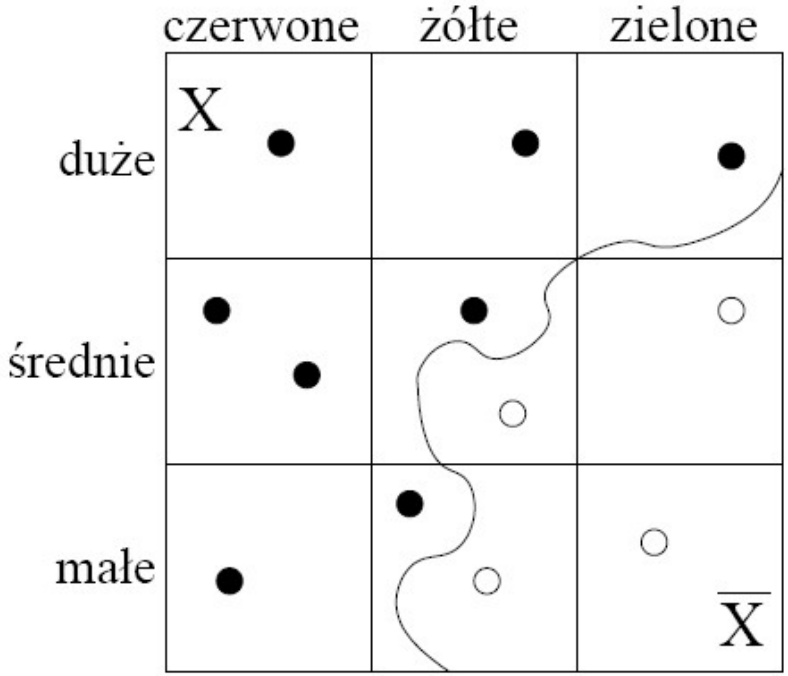
W obrębie każdego kwadratu obiekty są nierozróżnialne ze względu na pojęcia zdefiniowane przez zbiór B.

Relacja nierozróżnialności

Klasy abstrakcji — wniosek do zapamiętania

- Klasę abstrakcji nazywa się często pojęciem elementarnym lub pojęciem atomowym, gdyż jest najmniejszym podzbiorem uniwersum U, jaki możemy *sklasyfikować*.
- *Sklasyfikować* znaczy *odróżnić* od pozostałych elementów *za pomocą cech* — czyli atrybutów klasyfikujących obiekty do poszczególnych pojęć podstawowych.





X – jabłka dojrzałe, $\overline{X}$ – jabłka niedojrzałe.

Aproksymacja zbiorów - przybliżenie dolne, górne, brzeg zbioru

Problem z jednoznaczną klasyfikacją obiektów do pewnego podzbioru

- Jednym z celów wnioskowania w systemach decyzyjnych jest próba stwierdzenia czy obiekt (lub ich grupa) należy do pewnej klasy, lub nie. Inaczej mówiąc – czy należą do pewnego pojęcia czy nie.
- Proces taki opiera się na opisie obiektu wyrażonym przy pomocy atrybutów. Wybrany podzbiór atrybutów systemu informacyjnego determinuje podział obiektów na rozłączne klasy abstrakcji.
- Ważnym problemem jest zdolność radzenia sobie z niedoskonałymi danymi. Jednym ze źródeł trudności w zadaniach opisu czy klasyfikacji jest istnienie niespójności w dostępnych danych.
- Obiekty posiadające identyczne (lub podobne) opisy, lecz zaliczone do różnych pojęć, uniemożliwiają stworzenie jednoznacznej definicji tychże pojęć.
- Niespójności nie powinny być traktowane wyłącznie jako wynik błędu czy szumu informacyjnego. Mogą one także wynikać z niedostępności części informacji, naturalnej granularności i niejednoznaczności języka reprezentacji.

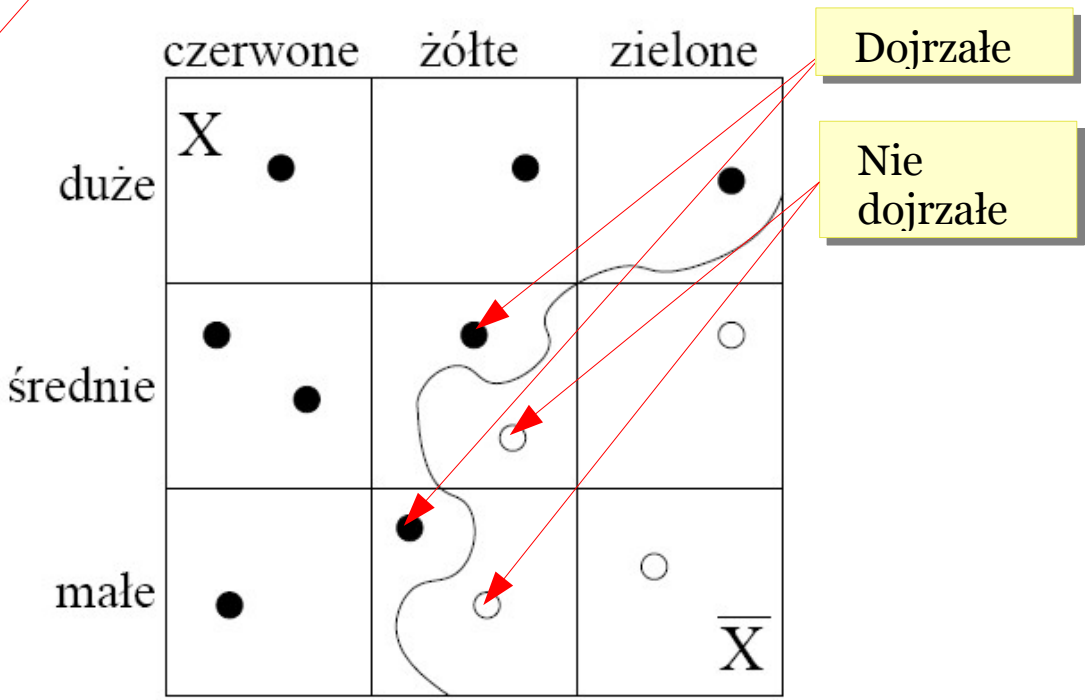
Aproksymacja zbiorów - przybliżenie dolne, górne, brzeg zbioru

Problem z jednoznaczną klasyfikacją obiektów do pewnego podzbioru

	kolor	*wielkość*	*dojrzałe*
x_1	*czerwone*	*duże*	*tak*
x_2	*żółte*	*średnie*	*tak*
x_3	*zielone*	*małe*	*nie*
x_4	*zielone*	*duże*	*tak*
x_5	*żółte*	*średnie*	*nie*
x_6	*czerwone*	*średnie*	*tak*
x_7	*żółte*	*duże*	*tak*
x_8	*czerwone*	*średnie*	*tak*
x_9	*żółte*	*małe*	*nie*
x_{10}	*żółte*	*małe*	*tak*
x_{11}	*czerwone*	*małe*	*tak*
x_{12}	*zielone*	*średnie*	*nie*

Gdy *żółte* i *średnie* to dojrzałe czy nie?

Gdy *żółte* i *małe* to dojrzałe czy nie?



Aproksymacja zbiorów - przybliżenie dolne, górne, brzeg zbioru

Zbiory przybliżone a problem z jednoznaczną klasyfikacją obiektów

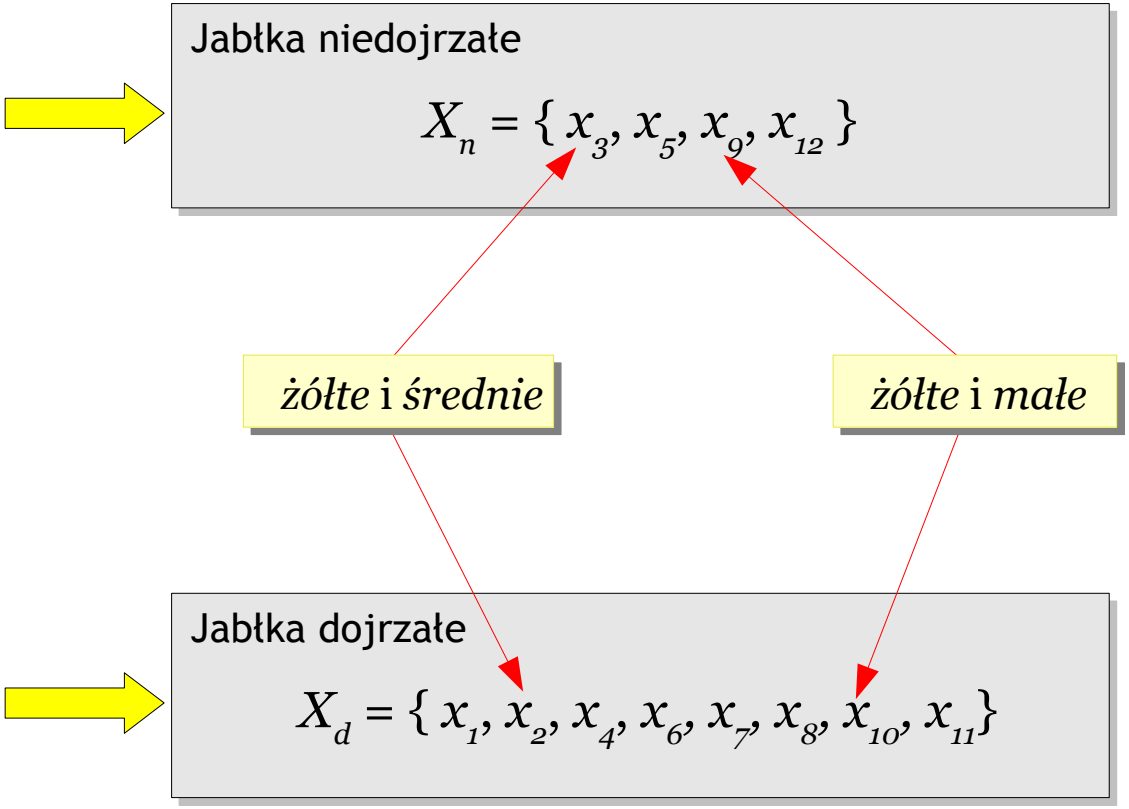
- Teoria *zbiorów przybliżonych* (ang. *rough sets*) zaproponowana przez Zdzisława Pawlaka jest dogodnym narzędziem analizy tego typu niespójności informacji.
- Teoria oparta jest na założeniu, że posiadając informację reprezentowaną za pomocą atrybutów i ich wartości na obiektach, możliwe jest określenie relacji zachodzącej pomiędzy tymi obiektami. Obiekty posiadające ten sam opis, wyrażony za pomocą atrybutów, są nierozróżnialne ze względu na dostępną informację.
- W przypadku niemożliwości precyzyjnego zdefiniowania zbioru obiektów (pojęcia, klasy decyzyjnej) tworzy ona *dolne* i *górne przybliżenie* tego zbioru na podstawie klas relacji nierozróżnialności pomiędzy obiektami.

Aproksymacja zbiorów - przybliżenie dolne, górne, brzeg zbioru

Aproksymacje zbiorów - zastosowanie praktyczne

Dzielimy jabłka na dwa zbiory: X_d — *zbiór jabłek dojrzałych* i X_n — *zbiór jabłek niedojrzałych*. Cały czas pamiętamy, że robot rozróżnia tylko dwie cechy: *kolor* i *wielkość*.

	kolor	*wielkość*	*dojrzałe*
x_1	*czerwone*	*duże*	*tak*
x_2	*żółte*	*średnie*	*tak*
x_3	*zielone*	*małe*	*nie*
x_4	*zielone*	*duże*	*tak*
x_5	*żółte*	*średnie*	*nie*
x_6	*czerwone*	*średnie*	*tak*
x_7	*żółte*	*duże*	*tak*
x_8	*czerwone*	*średnie*	*tak*
x_9	*żółte*	*małe*	*nie*
x_{10}	*żółte*	*małe*	*tak*
x_{11}	*czerwone*	*małe*	*tak*
x_{12}	*zielone*	*średnie*	*nie*



Problem — jak nauczyć rozpoznawania, które jabłko jest dojrzałe (na podstawie koloru i wielkości), skoro w tabeli są jabłka dojrzałe i niedojrzałe przy tych samych cechach.

Aproksymacja zbiorów - przybliżenie dolne, górne, brzeg zbioru

Aproksymacje zbiorów - definicje

Niech $SI = (U, A)$ będzie systemem informacyjnym, $B \subseteq A$ będzie pewnym zbiorem atrybutów. Niech będzie dany pewien podzbiór obiektów $X \subseteq U$, reprezentujący *pewne pojęcie*.

Chcemy zbadać, które obiekty w zbiorze U należą (lub nie) do pojęcia X.

Dla każdego obiektu $x \in U$ przez $[x]_{IND(B)}$ oznaczamy klasę abstrakcji relacji nierozróżnialności $IND(B)$, do której należy obiekt x.

Dolnym przybliżeniem (aproksymacją) pojęcia X w systemie SI w oparciu o opis B nazywamy zbiór:

$$B_{IND(B)}X = \{ x \in U : [x]_{IND(B)} \subseteq X \}$$

Górnym przybliżeniem (aproksymacją) pojęcia X w systemie SI w oparciu o opis B nazywamy zbiór:

$$B^{IND(B)}X = \{ x \in U : [x]_{IND(B)} \cap X \neq \emptyset \}$$

Brzegiem pojęcia X w systemie SI w oparciu o opis B nazywamy zbiór:

$$BN_B(X) = B^{IND(B)}X - B_{IND(B)}X$$

Aproksymacja zbiorów - przybliżenie dolne, górne, brzeg zbioru

Aproksymacje zbiorów - interpretacja na przykładzie

Niech:

$B = \{ kolor, wielkość \}$

$X_n = \{ x_3, x_5, x_9, x_{12} \}$

	kolor	*wielkość*	*dojrzałe*
x_1	*czerwone*	*duże*	*tak*
x_2	*żółte*	*średnie*	*tak*
x_3	*zielone*	*małe*	*nie*
x_4	*zielone*	*duże*	*tak*
x_5	*żółte*	*średnie*	*nie*
x_6	*czerwone*	*średnie*	*tak*
x_7	*żółte*	*duże*	*tak*
x_8	*czerwone*	*średnie*	*tak*
x_9	*żółte*	*małe*	*nie*
x_{10}	*żółte*	*małe*	*tak*
x_{11}	*czerwone*	*małe*	*tak*
x_{12}	*zielone*	*średnie*	*nie*

$[x_1]_{IND(B)} = \{ x_1 \}$

$[x_2]_{IND(B)} = \{ x_2, x_5 \}$

$[x_3]_{IND(B)} = \{ x_3 \}$

$[x_4]_{IND(B)} = \{ x_4 \}$

$[x_5]_{IND(B)} = \{ x_2, x_5 \}$

$[x_6]_{IND(B)} = \{ x_6, x_8 \}$

$[x_7]_{IND(B)} = \{ x_7 \}$

$[x_8]_{IND(B)} = \{ x_6, x_8 \}$

$[x_9]_{IND(B)} = \{ x_9, x_{10} \}$

$[x_{10}]_{IND(B)} = \{ x_9, x_{10} \}$

$[x_{11}]_{IND(B)} = \{ x_{11} \}$

$[x_{12}]_{IND(B)} = \{ x_{12} \}$

Przybliżenie dolne:

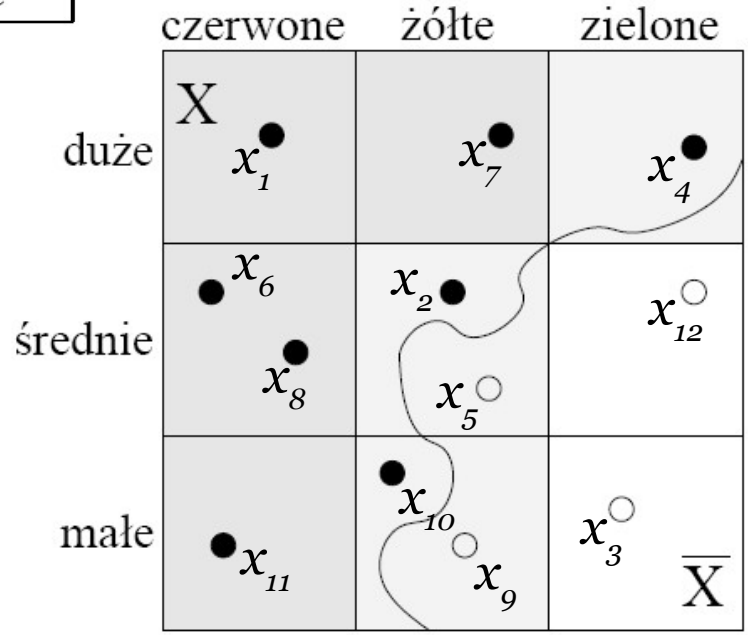
$$B_{IND(B)}X_n = \{ x_3, x_{12} \}$$

Przybliżenie górne:

$$B^{IND(B)}X_n = \{ x_2, x_3, x_5, x_9, x_{10}, x_{12} \}$$

Brzeg:

$$BN_N(X_n) = \{ x_2, x_5, x_9, x_{10} \}$$



Aproksymacja zbiorów - przybliżenie dolne, górne, brzeg zbioru

Aproksymacje zbiorów - interpretacja

- Za pomocą dolnej i górnej aproksymacji jesteśmy w stanie określić nieostre pojęcie w ścisły sposób.
- *Dolna aproksymacja* pojęcia, to wszystkie te obiekty, które *należą bez wątpienia* do pojęcia X. Należą one bowiem do takich klas abstrakcji, które w całości zawierają się w pojęciu X.
- *Górna aproksymacja* pojęcia, to zbiór takich obiektów, co do których *nie możemy wykluczyć, że należą do pojęcia* X. Jest to spowodowane tym, że należą do klas abstrakcji mających niepuste przecięcie z pojęciem X. Są zatem nierozróżnialne z pewnymi obiektami należącymi do tego pojęcia.
- Brzeg zbioru X zawiera obiekty, których nie można jednoznacznie przydzielić do X z uwagi na sprzeczny opis.

Aproksymacja zbiorów - przybliżenie dolne, górne, brzeg zbioru

Aproksymacje zbiorów - interpretacja

Zbiór przybliżony X może być scharakteryzowany ilościowo za pomocą:

- Współczynnika *dokładności przybliżenia*:

$$\alpha BX = \frac{|B_{IND(B)}X|}{|B^{IND(B)}X|}$$ gdzie $|X|$ to liczność niepustego zbioru X

- Współczynnika *dokładności przybliżenia dolnego*:

$$\alpha_B X = \frac{|B_{IND(B)}X|}{|U|}$$

- Współczynnika *dokładności przybliżenia górnego*:

$$\alpha_B X = \frac{|B^{IND(B)}X|}{|U|}$$

Redukcja atrybutów - pojęcie jądra i reduktów

Nadmiar informacji jest szkodliwy

- W celu precyzyjnego i konkretnego opisana relacji pomiędzy obiektami występującymi w bazie wiedzy, stosuje się redukcję liczby atrybutów opisujących owe relacje.
- Poszukuje się takich podzbiorów atrybutów, które zachowują podział obiektów na klasy decyzyjne taki sam, jak wszystkie atrybuty.
- Te zbiory atrybutów nie mogą być wyznaczone w dowolny sposób. W teorii zbiorów przybliżonych wykorzystuje się koncepcję *reduktu* będącego niezależnym podzbiorem atrybutów zachowującym taki sam podział na klasy decyzyjne jak wszystkie atrybuty.
- Węższym pojęciem jest pojęcie *jądra*, określającego zbiór atrybutów niezbędnych dla zachowania rozróżnialności obiektów w systemie.

Definicja jądra

Niech dany będzie system informacyjny $SI = (U, A)$ oraz zbiór atrybutów $B \subseteq A$. Zbiór wszystkich niezbędnych atrybutów w B nazywamy *jądrem* (rdzeniem) i oznaczamy przez $CORE(B)$.

Redukcja atrybutów - pojęcie jądra i reduktów

Kiedy atrybut jest niezbędny?

Niech $B \subseteq A$ i $a \in B$. Mówimy, że atrybut a jest zbędny w B, gdy $IND(B) = IND(B - \{a\})$.
W przeciwnym przypadku atrybut a jest niezbędny w B.

Zbiór atrybutów B jest *niezależny*, gdy dla każdego $a \in B$ atrybut a jest niezbędny.
W przeciwnym przypadku zbiór jest *zależny*.

Wyznaczanie jądra z definicji

U	A			
	a	b	c	d
1	1	0	2	2
2	0	1	1	2
3	0	0	2	2
4	2	2	1	0

$U/IND(B) = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

$U/IND(B-\{a\}) = \{\{1, 3\}, \{2\}, \{4\}\} \rightarrow a$ jest niezbędny

$U/IND(B-\{b\}) = \{\{1\}, \{2\}, \{3\}, \{4\}\} \rightarrow b$ jest zbędny

$U/IND(B-\{c\}) = \{\{1\}, \{2\}, \{3\}, \{4\}\} \rightarrow c$ jest zbędny

Zakładamy, że $B = \{a, b, c\}$

Jądro $CORE(B) = \{a\}$

Redukcja atrybutów - pojęcie jądra i reduktów

Ogólny algorytm wyznaczania jądra z definicji

1. Wyznacz klasy abstrakcji relacji nierozróżnialności $U/IND(B)$, gdzie B jest to zbiór wszystkich rozważanych atrybutów.
2. Wyznacz klasy abstrakcji z pominięciem i-tego atrybutu $U/\ IND(B-\{a_i\})$.
3. Jeżeli $U/\ IND(B) = U/IND(B-\{a_i\})$ to
4. Atrybut a_i jest zbędny,
5. W przeciwnym wypadku
6. Atrybut a_i jest niezbędny i wchodzi do jądra $CORE(B)$.
7. Powtarzaj pkt. 2, aż wykorzystane zostaną wszystkie atrybuty z B.

Redukcja atrybutów - pojęcie jądra i reduktów

Szczegółowy algorytm wyznaczania jądra z definicji

Dane wejściowe:

System informacyjny: $SI = (U, A)$

Zbiór atrybutów $B \subseteq A$: $B = \{a_1, a_2, \ldots a_i, \ldots a_n\}$

Dane wyjściowe:

Jądro, zbiór atrybutów: $CORE(B)$

Algorytm:

CORE(B):= {}

Wyznacz $U/\ INB(B)$

Dla każdego $a_i \in B$ wykonaj

Jeżeli $U/\ INB(B) <> U/\ IND(B-\{a_i\})$ To

$CORE\ (B) := CORE(B) \cup \{\ a_i\}$

gdzie:

- $CORE\ (B)$ – jądro (zbiór atrybutów).
- B - rozważany zbiór atrybutów.
- a_i - i-ty atrybut ze zbioru B.
- $U/INB(B)$ - klasa abstrakcji relacji nierozróżnialności dla pełnego zbioru atrybutów.
- $U/IND(B-\{a\})$ - klasy abstrakcji relacji nierozróżnialności dla zbioru atrybutów z pominięciem atrybutu a.

Redukcja atrybutów - pojęcie jądra i reduktów

Jeszcze jeden przykład

	a	b	c	d	e
1	0	1	0	1	T
2	1	0	0	1	T
3	1	1	1	0	T
4	1	1	1	1	T
5	0	1	0	0	N

1. gdzie $B = \{a,b,c,d\}$
2. $U/\ IND\ (B) = \{\{1\},\{2\},\{3\},\{4\},\{5\}\}$
3. $U/\ IND(B-\{a\}) = \{\{1\},\{2\},\{3\},\{4\},\{5\}\}$
4. $U/\ IND(B) = U/IND(B-\{a\})$ -> atrybut *a* jest zbędny
5. $U/\ IND(B-\{b\}) = \{\{1\},\{2\},\{3\},\{4\},\{5\}\}$
6. $U/\ IND(B) = U/\ \mathrm{IND}(B-\{b\})$ -> atrybut *b* jest zbędny
7. $U/\ IND(B-\{c\}) = \{\{1\},\{2\},\{3\},\{4\},\{5\}\}$
8. $U/\ IND(B) = U/\ \mathrm{IND}(B-\{c\})$ -> atrybut *c* jest zbędny
9. $U/\ IND(B-\{d\}) = \{\{1,5\},\{2\},\{3,4\}\}$
10. $U/\ IND(B) <> U/\ \mathrm{IND}(B-\{d\})$ -> atrybut *d* jest niezbędny,
11. $CORE(B) = \{d\}$

Redukcja atrybutów - pojęcie jądra i reduktów

Definicja reduktu

Podzbiór atrybutów $B \subseteq A$ nazywamy *reduktem* zbioru atrybutów A, gdy zbiór atrybutów B jest *niezależny* oraz $IND(B) = IND(A)$.

Zbiór wszystkich reduktów oznaczamy przez $RED(A)$.

Redukt to najmniejszy zbiór atrybutów, przy którym zostaje zachowana dotychczasowa klasyfikacja (rozróżnialność) obiektów

Ważne!

Redukt musi spełniać *dwa* kryteria:

1. musi być *niezależnym zbiorem atrybutów* (tylko atrybuty niezbędne),
2. musi *zachowywać taką samą rozróżnialność obiektów* jak zbiór redukowany.

Uwaga

Redukty można wyznaczać dla dowolnego podzbioru A. Do tej pory rozważaliśmy zawsze jakiś podzbiór atrybutów $B \subseteq A$. Dla takiego podzbiory B też możemy liczyć redukty. Wtedy reduktem będzie jakiś podzbiór atrybutów $C \subseteq B$, a zbiór wszystkich reduktów B oznaczać będziemy $RED(B)$.

Redukcja atrybutów - pojęcie jądra i reduktów

Związek pomiędzy jądrem a reduktem

Jądro systemu informacyjnego rozpatrywanego dla podzbioru atrybutów $B \subseteq A$ jest częścią wspólną wszystkich reduktów tego systemu.

$$CORE(B) = \cap RED(A).$$

Uwaga! To właściwość wiążąca jądro i redukty a nie definicja jądra!

Redukcja atrybutów - pojęcie jądra i reduktów

Wyznaczanie reduktów z definicji

U	A			
	a	b	c	d
1	1	0	2	2
2	0	1	1	2
3	0	0	2	2
4	2	2	1	0

Zakładamy, że $B = \{a, b, c\}$

Potencjalne redukty to:

$\{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}$

Ale wiemy już, że $CORE(B) = \{a\}$

Skoro jądro utrzymuje rozróżnialność obiektów w systemie, to nie możemy atrybutów z jądra zredukować.

Pozwala to zawęzić zbiór potencjalnych reduktów do:

$\{a\}, \{a, b\}, \{a, c\}, \{a, b, c\}$

Samo jądro jest interesującym kandydatem na redukt. Z definicji spełnia już *pierwszy* warunek dla reduktu – jest zbioremnie zależnym.

Czy jednak $U/IND(\{a\}) = U/IND(\{a, b, c\})$? Sprawdźmy *drugi* warunek:

$U/IND(\{a, b, c\}) = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

$U/IND(\{a\}) = \{\{1\}, \{2, 3\}, \{4\}\}$ $\rightarrow$ Jądro CORE(B) = $\{a\}$ nie jest reduktem

Redukcja atrybutów - pojęcie jądra i reduktów

Wyznaczanie reduktów z definicji

	A			
U	a	b	c	d
1	1	0	2	2
2	0	1	1	2
3	0	0	2	2
4	2	2	1	0

Czy $B_1 = \{a, b\}$ jest reduktem?

1. Czy B_1 jest niezależny?

$U/IND(B_1) = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

$U/IND(B_1\text{-}\{a\}) = \{\{1, 3\}, \{2\}, \{4\}\}$ $\rightarrow$ $\{a\}$ jest niezbędny

$U/IND(B_1\text{-}\{b\}) = \{\{1\}, \{2, 3\}, \{4\}\}$ $\rightarrow$ $\{b\}$ jest niezbędny

Zbiór B_1 jest niezależny, spełnia pierwszy warunek reduktu

Redukcja atrybutów - pojęcie jądra i reduktów

Wyznaczanie reduktów z definicji

	A			
U	a	b	c	d
1	1	0	2	2
2	0	1	1	2
3	0	0	2	2
4	2	2	1	0

Czy $B_1 = \{a, b\}$ jest reduktem?

2. Czy B_1 zachowuje taką samą rozróżnialność obiektów jak B?

$U/IND(\{a, b, c\}) = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

$U/IND(B_1) = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

Widzimy, że:

$U/IND(\{a, b, c\}) = U/IND(B_1)$

Zatem

Zbiór B_1 jest reduktem

Redukcja atrybutów - pojęcie jądra i reduktów

Wyznaczanie reduktów z definicji

	A			
U	a	b	c	d
1	1	0	2	2
2	0	1	1	2
3	0	0	2	2
4	2	2	1	0

Czy $B_2 = \{a, c\}$ jest reduktem?

1. Czy B_2 jest niezależny?

$U/IND(B_2) = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

$U/IND(B_2\text{-}\{a\}) = \{\{1, 3\}, \{2, 4\}\}$ $\rightarrow$ $\{a\}$ jest niezbędny

$U/IND(B_2\text{-}\{c\}) = \{\{1\}, \{2, 3\}, \{4\}\}$ $\rightarrow$ $\{c\}$ jest niezbędny

Zbiór B_2 jest niezależny, spełnia pierwszy warunek reduktu

Redukcja atrybutów - pojęcie jądra i reduktów

Wyznaczanie reduktów z definicji

	A			
U	a	b	c	d
1	1	0	2	2
2	0	1	1	2
3	0	0	2	2
4	2	2	1	0

Czy $B_2 = \{a, c\}$ jest reduktem?

2. Czy B_2 zachowuje taką samą rozróżnialność obiektów jak B?

$U/IND(\{a, b, c\}) = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

$U/IND(B_2) = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

Widzimy, że:

$U/IND(\{a, b, c\}) = U/IND(B_2)$

Zatem

Zbiór B_2 jest reduktem

Redukcja atrybutów - pojęcie jądra i reduktów

Wyznaczanie reduktów z definicji

	A			
U	a	b	c	d
1	1	0	2	2
2	0	1	1	2
3	0	0	2	2
4	2	2	1	0

Ostatecznie, dla rozważanego systemu informacyjnego i zbioru $B = \{a, b, c\}$ zbiór reduktów RED(B):

$$\text{RED}(B) = \{\{a, b\}, \{a, c\}\}$$

Redukcja atrybutów - pojęcie jądra i reduktów

Ogólny algorytm wyznaczania reduktów z definicji

1. Wyznacz klasy abstrakcji $U/ IND(B)$, gdzie B jest to zbiór wszystkich rozważanych atrybutów.
2. Sprawdź, czy jądro $CORE(B)$ nie jest reduktem:
3. Ponieważ jądro to zbiór atrybutów niezbędnych, to sprawdź, czy $U/IND(B) = U/IND(CORE(B))$, jeżeli tak to jądro to jedyny redukt i przejdź do Punktu 6.
4. Sprawdź kolejne podzbiory atrybutów $B_i \subseteq B$.
5. Jeżeli podzbiór B_i jest niezależny to:

   Jeżeli U/ IND(B) = U/ IND(B_i) to:

   Dopisz podzbiór B_i do zbioru reduktów.
6. Koniec.

Redukcja atrybutów - pojęcie jądra i reduktów

Szczegółowy algorytm wyznaczania reduktów z definicji

Dane wejściowe:

System informacyjny: $SI = (U, A)$

Zbiór atrybutów $B \subseteq A$: $B = \{a_1, a_2, \ldots a_i, \ldots a_n\}$

Dane wyjściowe:

Zbiór reduktów: $RED(B)$

Jądro, zbiór atrybutów: $CORE(B)$

Redukcja atrybutów - pojęcie jądra i reduktów

Szczegółowy algorytm wyznaczania reduktów z definicji

Algorytm:

$RED(B) := \{\}$

$CORE(B) := \{\}$

Wyznacz $U/\ IND(B)$

Wyznacz $CORE(B)$

Jeżeli $U/\ IND(B) = U/\ IND(CORE(B))$ To:

$RED(B) := CORE(B)$

W przeciwnym wypadku:

Dla każdego podzbioru atrybutów $B_i \subseteq B$ wykonaj

Jeżeli B_i jest niezależny To

Jeżeli $U\ /\ IND(B) = U\ /\ IND(B_i)$ To

$RED(B) := RED(B) \cup B_i$

Redukcja atrybutów - pojęcie jądra i reduktów

Wyznaczenie jądra i reduktów z definicji jest niewygodne

Wyznaczanie reduktów ułatwi *macierz nierozróżnialności*. Dla danego systemu informacyjnego $SI=(U, A)$, gdzie $U = \{ x_1, x_2, ..., x_n\}$ oraz podzbioru atrybutów $B \subseteq A$, *macierz nierozróżnialności* $M(SI)=[c_{ij}]_{nxn}$ definiujemy następująco:

$$c_{ij} = \{ a \in \text{A}: a(x_i) \neq a(x_j), i, j = 1, 2, ..., n \}$$

Każdy element macierzy c_{ij} jest zbiorem atrybutów różniących i-ty i j-ty obiekt z U.

Macierz nierozróżnialności - przykład

Zakładamy, że $B = \{a, b, c\}$

$SI=(U, A)$

	A			
U	a	b	c	d
1	1	0	2	2
2	0	1	1	2
3	0	0	2	2
4	2	2	1	0

$M(SI)=[c_{ij}]_{nxn}$

	1	2	3	4
1	∅	a,b,c	a	a,b,c,d
2	a,b,c	∅	b,c	a,b,d
3	a	b,c	∅	a,b,c,d
4	a,b,c,d	a,b,d	a,b,c,d	∅

Redukcja atrybutów - pojęcie jądra i reduktów

Wyznaczenie jądra i reduktów z macierzy nierozróżnialności

- Wyznaczanie jądra $CORE(B)$:

Do rdzenia wchodzą atrybuty występujące w macierzy nierozróżnialności pojedynczo.

$CORE(B) = \{ a \in A: c_{ij} = \{ a \}$, dla pewnego $1 <= i,j <= n \}$

- Wyznaczanie reduktów $RED(B)$:

Pewien podzbiór atrybutów $C \subseteq B$ jest reduktem jeśli jest minimalny (w sensie zawierania zbiorów) oraz posiada niepuste przecięcie z każdym niepustym elementem macierzy $M(SI)$.

Macierz nierozróżnialności - przykład

$$M(SI)=[c_{ij}]_{nxn}$$

	1	2	3	4
1	∅	a,b,c	a	a,b,c,d
2	a,b,c	∅	b,c	a,b,d
3	a	b,c	∅	a,b,c,d
4	a,b,c,d	a,b,d	a,b,c,d	∅

$CORE(B) = \{ a \}$, bo $c_{13} = \{ a \}$

$RED(B) = \{ \{ a, b \}, \{ a, c \} \}$

ponieważ:
$\{ a, b \} \cap c_{ij} \neq \emptyset$ i $\{ a, c \} \cap c_{ij} \neq \emptyset$
i te zbiory są minimalne.