Procesy i systemy Business Intelligence

Odkrywanie asocjacji

Cel

- **Celem** procesu odkrywania asocjacji jest znalezienie interesujących zależności lub korelacji (nazywanych ogólnie asocjacjami) pomiędzy danymi w dużych zbiorach danych.

- **Wynikiem** procesu odkrywania asocjacji jest zbiór reguł asocjacyjnych opisujących znalezione zależności lub korelacje między danymi.

Analiza koszyka zakupów

- **Geneza** problemu odkrywania reguł asocjacyjnych: problem analizy koszyka zakupów (MBA – Market Basket Analysis)
  - **dane**: zbiór danych zawierający informacje o zakupach realizowanych przez klientów supermarketu
  - **cel**: znalezienie grup produktów, które klienci supermarketu najczęściej kupują razem
- Celem analizy MBA jest znalezienie naturalnych wzorców zachowań konsumenckich klientów

Przykładowy wzorzec

Ktoś kto kupuje pieluszki, najczęściej kupuje też piwo

- Znalezione wzorce zachowań mogą być wykorzystane:
  - organizacji pólek w supermarkecie
  - opracowania akcji promocyjnych
  - opracowania katalogu oferowanych produktów

Analiza koszyka zakupów

- MBA znajduje zastosowanie wszędzie tam, gdzie zbiór klientów („koszyki”) nabywa łącznie pewien zbiór „produktów” (dóbr lub usług):
  - telekomunikacja
  - analiza pogody
  - bankowość
  - diagnostyka medyczna
  - karty kredytowe
- Model koszyka zakupów jest abstrakcją umożliwiającą modelowanie relacji między „produktami” i „koszykami”

Tablica obserwacji

- Koszyk zakupów można opisać za pomocą tzw. *tablicy obserwacji*
  - **dane**: zbiór atrybutów $A = \{A_1, A_2, ..., A_n\}$, zbiór koszyków (obserwacji) $T = \{T_1, T_2, ..., T_m\}$
  - atrybuty tablicy reprezentują wystąpienia „produktów”
  - wiersze tablicy reprezentują wystąpienia „koszyków”
  - atrybut *tr_id* - identyfikatory poszczególnych obserwacji
  - pozycja $T_i[A_j] = 1$ tablicy wskazuje, że i-ta obserwacja zawiera wystąpienie j-tego atrybutu

tr_id	coca_cola	piwo	orzeszki	pieluszki
t1	1	0	1	0
t2	0	1	1	1
t3	1	0	0	0
t4	1	1	1	0
t5	0	1	1	1

Tablica obserwacji (2)

„Koszyki” = studenci, „produkty” = wykłady oferowane przez uczelnię

poszukiwanie wykładów, które studenci wybierają najczęściej łącznie

„Koszyki” = strony WWW, „produkty” = słowa kluczowe

poszukiwanie stron WWW opisanych tymi samymi, lub podobnymi, zbiorami słów kluczowych (prawdopodobnie, znalezione strony dotyczą podobnej problematyki)

„Koszyki” = zdania, „produkty” = słowa występujące w tych zdaniach

poszukiwanie zbitek słów lub haseł występujących często razem

„Koszyki” = zdania, „produkty” = dokumenty zawierające te zdania

dokumenty występujące zbyt często razem mogą przedstawiać plagiaty

Reguły asocjacyjne

- Wynikiem analizy tablicy obserwacji (zbioru koszyków zakupów) jest zbiór reguł asocjacyjnych następującej postaci:

$$\{(A_{i1} = 1) \wedge \ldots \wedge (A_{ik} = 1) \rightarrow \{(A_{ik+1} = 1) \wedge \ldots \wedge (A_{ik+l} = 1)$$

interpretacja reguły: „jeżeli klient kupił produkty A_{i1}, A_{i2}, ..., A_{ik}, to prawdopodobnie kupił również produkty A_{ik+1}, A_{ik+2}, ..., A_{ik+l}”

Miary oceny reguł asocjacyjnych

- **Wsparcie** (ang. *support*) reguły asocjacyjnej to miara ważności i powszechności reguły

$$\text{support}(X \rightarrow Y) = P(X \cup Y)$$

- **Ufność** (ang. confidence) reguły asocjacyjnej to miara siły z jaką reguła wiąże elementy

$$\text{confidence}(X \rightarrow Y) = P(Y \mid X) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}$$

Interpretacja miar oceny reguł asocjacyjnych

kiełbasa = 1 $\wedge$ keczup = 1 $\rightarrow$ piwo = 1

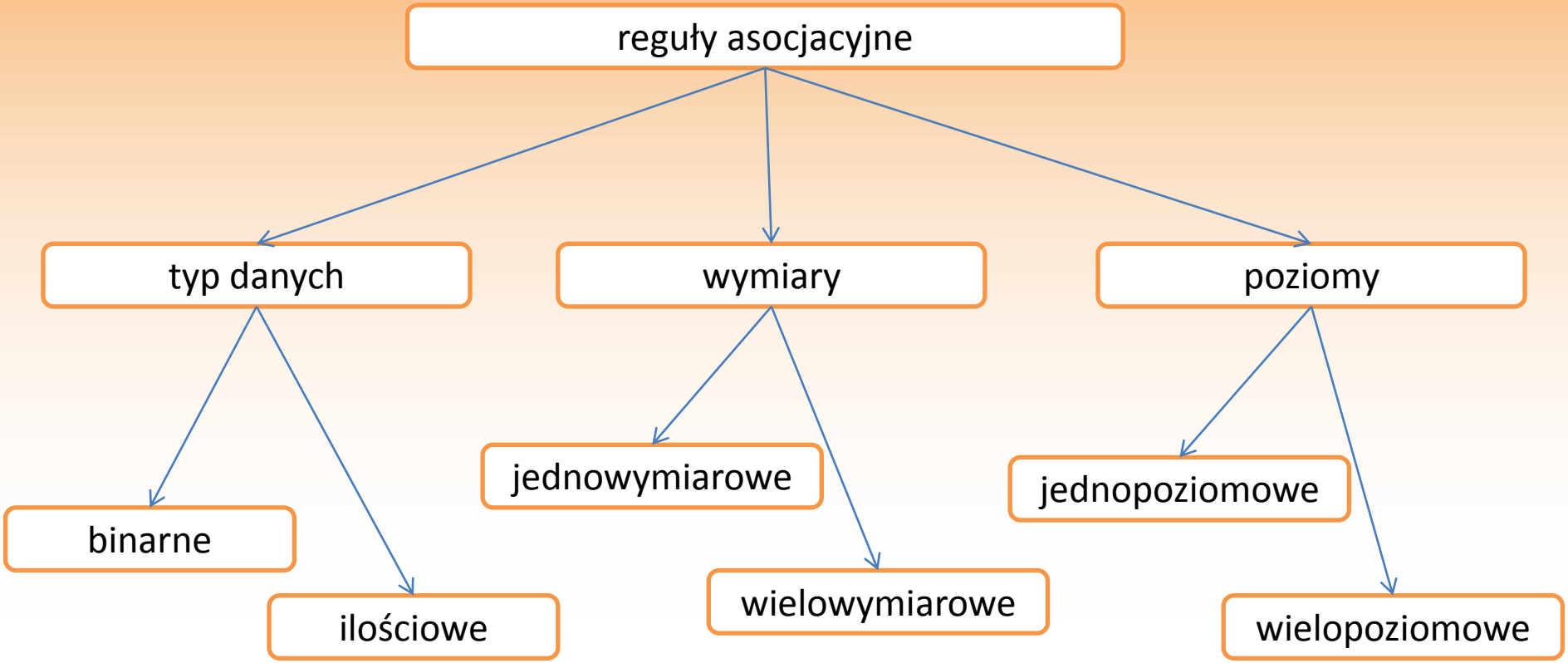
- $\text{support}(\text{kiełbasa} = 1 \wedge \text{keczup} = 1 \rightarrow \text{piwo} = 1) = 3\ \%$
  - 3% klientów kupiło w trakcie pojedynczej wizyty w sklepie kiełbasę, keczup i piwo
- $\text{confidence}(\text{kiełbasa} = 1 \wedge \text{keczup} = 1 \rightarrow \text{piwo} = 1) = 80\ \%$
  - 80% klientów którzy kupili w trakcie pojedynczej wizyty w sklepie kiełbasę i keczup, kupiło także piwo

Silne reguły asocjacyjne

- **Cel**: znalezienie wszystkich reguł, których wsparcie jest większe niż *minsup* i których ufność jest większa niż *minconf*
  - *minsup*, *minconf* – parametry definiowane przez użytkownika

- Reguły, które spełniają warunek minimalnego wsparcia i minimalnej ufnosci, nazywamy *silnymi regułami asocjacyjnymi* (ang. *strong association rules*)

Klasyfikacja reguł asocjacyjnych



Przykłady reguł asocjacyjnych

Binarna reguła asocjacyjna

- $pieluszki = 1 \rightarrow piwo = 1$

Ilościowa reguła asocjacyjna

- $wiek = \text{'30...40'} \wedge wykształcenie = \text{'wyższe'} \rightarrow średnie\_zarobki = 3\,100\ zł$

Jednowymiarowa reguła asocjacyjna

- $kiełbasa = 5 \wedge piwo = 12 \rightarrow czipsy = 2$

Wielowymiarowa reguła asocjacyjna

- $płeć = M \wedge wiek > 40 \wedge pali = tak \rightarrow choroba\ serca = tak$

Jednopoziomowa reguła asocjacyjna

- $kiełbasa = 1 \rightarrow keczup = 1$

Wielopoziomowa reguła asocjacyjna

- $kiełbasa = 1 \wedge keczup = 1 \rightarrow napój\ bezalkoholowy = 1$

Odkrywanie binarnych reguł asocjacyjnych

- **dane**:
  - $I=\{i_1, i_2, ..., i_m\}$: zbiór elementów (ang. *items*)
  - transakcja: zbiór elementów $T \subseteq I$ i $T \neq \emptyset$
  - rozmiar transakcji *size*(T) – liczba elementów w transakcji T
  - baza danych D: zbiór transakcji

  - transakcja T *wspiera* element $x \in I$, jeżeli $x \in T$
  - transakcja T *wspiera* zbiór $X \subseteq I$, jeżeli T wspiera każdy element ze zbioru X, $X \subseteq T$

- **wsparcie zbioru** X w bazie danych D, *support*(X) to procent transakcji z bazy danych D wspierających zbiór X

Przykład

tr_id	produkt	data	liczba	cena
1	orzeszki	2/22/98	6	0,99
1	coca-cola	2/22/98	3	0,20
2	piwo	2/22/98	4	0,49
2	orzeszki	2/22/98	1	0,99
2	pieluszki	2/22/98	1	1,49
3	coca-cola	2/23/98	10	0,20
4	coca-cola	2/24/98	6	0,20
4	piwo	2/24/98	2	0,49
4	orzeszki	2/24/98	4	0,99
5	piwo	2/24/98	2	0,49
5	orzeszki	2/24/98	4	0,99
5	pieluszki	2/24/98	10	1,49

Dla minsup=0,4 i minconf=0,5:

piwo → orzeszki	sup= 0,60 conf= 1,00
orzeszki → piwo	sup= 0,60 conf= 0,75
piwo ^ pieluszki → orzeszki	sup= 0,40 conf= 1,00
pieluszki ^ orzeszki → piwo	sup= 0,40 conf= 1,00
pieluszki → piwo ^ orzeszki	sup= 0,40 conf= 1;00
pieluszki → piwo	*sup*= 0,40 *conf*= 1,00
pieluszki → orzeszki	*sup*= 0,40 *conf*= 1,00
piwo ^ orzeszki → pieluszki	*sup*= 0,40 *conf*= 0,67

Algorytm naiwny

- Dany jest zbiór elementów I, baza danych transakcji D, oraz minimalne progi wsparcia i ufności minsup i minconf
  - wygeneruj wszystkie możliwe podzbiory zbioru I
  - dla każdego podzbioru oblicz wsparcie tego zbioru w bazie danych D
  - dla każdego zbioru, którego wsparcie jest niemniejsze niż *minsup*, wygeneruj regułę asocjacyjną – dla każdej otrzymanej reguły oblicz ufność reguły

- Liczba wszystkich możliwych podzbiorów zbioru I wynosi $2^{|I|}$ (rozmiar I ≈ 200 000 elementów)
- Liczba wszystkich możliwych binarnych reguł asocjacyjnych dla zbioru elementów I wynosi $3^{|I|} - 2^{|I|+1} + 1$

Ważna obserwacja

- Jeżeli wsparcie zbioru ($X \cup Y$) jest mniejsze niż *minsup*, to wówczas możemy pominąć obliczanie ufności reguł asocjacyjnych $X \rightarrow Y$ oraz $Y \rightarrow X$, gdyż reguły te zostaną odrzucone

- Jeżeli wsparcie zbioru (X, Y, Z) jest mniejsze niż *minsup*, to możemy pominąć obliczanie ufności następujących 6 reguł asocjacyjnych:

  $(X \rightarrow Y, Z)$ $(Y \rightarrow X, Z)$ $(Z \rightarrow X, Y)$ $(X, Y \rightarrow Z)$ $(X, Z) \rightarrow Y$ $(Y, Z) \rightarrow X$

- W ogólnym przypadku odrzucenie zbioru *k*-elementowego $(X_1, X_2, \ldots, X_k)$, którego wsparcie jest mniejsze niż *minsup*, pozwala pominąć obliczanie ufności $2^k - 2$ reguł asocjacyjnych

Ogólny algorytm odkrywania reguł asocjacyjnych

Algorytm 1.1: Ogólny algorytm odkrywania reguł asocjacyjnych

dane: baza danych D, *minsup*, *minconf*

wynik: zbiór silnych binarnych reguł asocjacyjnych

1. znajdź wszystkie zbiory elementów $I_i=\{i_{i1}, i_{i2}, ..., i_{im}\}$, $I_i \subseteq I$, których $support(I_i) \geq minsup$ (zbiory I_i nazywać będziemy *zbiorami częstymi*)
2. na podstawie zbiorów częstych znalezionych w kroku 1 wygeneruj wszystkie silne binarne reguły asocjacyjne – zastosuj algorytm 1.2

Ogólny algorytm odkrywania reguł asocjacyjnych

Algorytm 1.2: Generowanie reguł asocjacyjnych

dane: kolekcja zbiorów częstych U_j I_j

wynik: zbiór silnych binarnych reguł asocjacyjnych

1. dla każdego zbioru częstego I_j znajdź jego wszystkie niepuste podzbiory $subI_j$
2. dla każdego podzbioru $subI_j$ dla którego zachodzi

   $$support(I_j) / support(subI_j) \geq minconf$$

   wygeneruj regułę asocjacyjną $subIj \rightarrow (I_j - subI_j)$

Algorytm Apriori (1)

- Założenia i oznaczenia:
  - zakładamy, że wszystkie transakcje są wewnętrznie uporządkowane (np. leksykograficznie)
  - L_k oznacza kolekcję zbiorów częstych o rozmiarze k, nazywanych *częstymi* zbiorami k-elementowymi
  - C_k oznacza kolekcję zbiorów kandydujących o rozmiarze k, nazywanych *kandydującymi* zbiorami k-elementowymi

Algorytm Apriori (2)

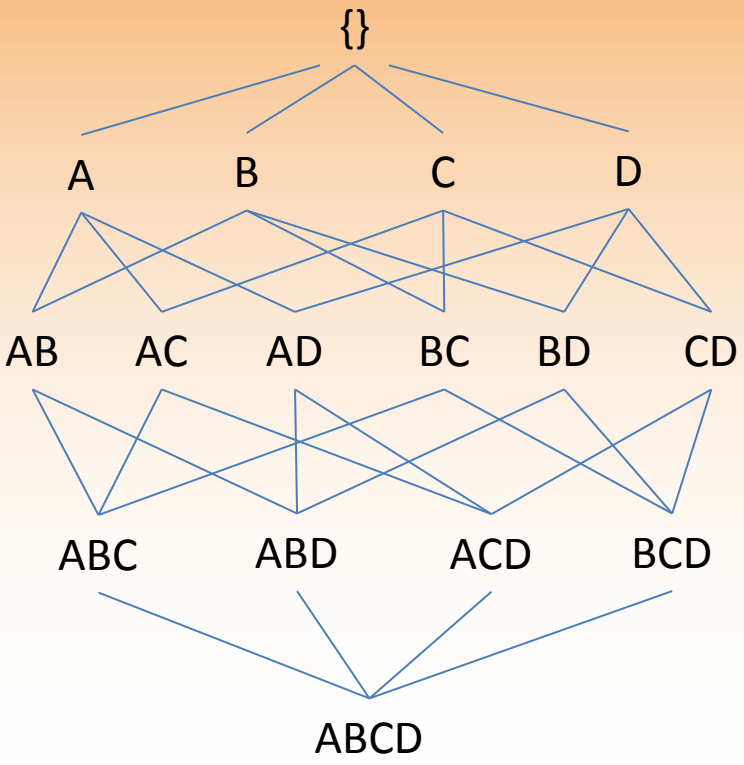
Algorytm 1.3: Algorytm odkrywania zbiorów częstych Apriori

We: baza danych D, próg minimalnego wsparcia minsup

Wy: wszystkie zbiory częste z bazy danych D

1. $L1 \leftarrow$ 1-elementowe zbiory częste ;
2. **for** ($k = 2$; $L_{k-1} \neq \emptyset$; k++) **do**
3. $Ck \leftarrow$ apriori_gen(L_{k-1});
4. **for all** transakcji $t \in D$ **do**
5. $C_t \leftarrow$ subset(C_k, t);
6. **for all** zbiorów kandydujących $c \in C_t$ **do**
7. c.count++;
8. **end for**
9. **end for**
10. $Lk \leftarrow \{c \in C_k \mid c.count \geq minsup\}$;
11. **end for**
12. **return** $\bigcup_k L_k$

Odkrywanie zbiorów częstych

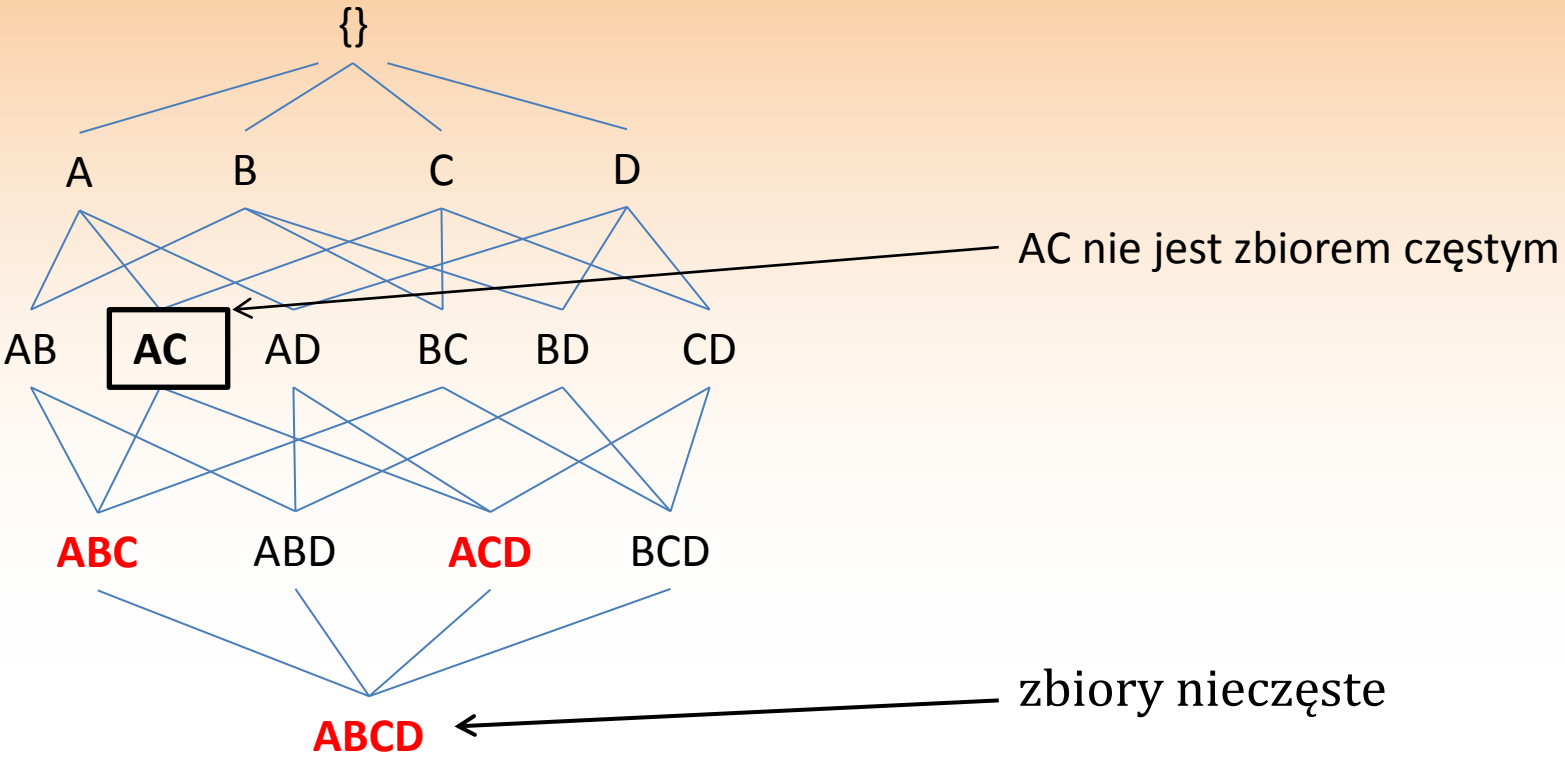


Celem każdego algorytmu odkrywania zbiorów częstych jest ograniczenie liczby analizowanych zbiorów elementów występujących w kracie

antymonotoniczność - wszystkie podzbiory zbioru częstego muszą być częste, jeżeli B jest zbiorem częstym i $A \subseteq B$, to A jest również zbiorem częstym

jeżeli zbiór A nie jest zbiorem częstym, to żaden nadzbiór B zbioru A, $A \subseteq B$, nie będzie zbiorem częstym, nie musimy rozważać wsparcia zbioru *X*, którego jakikolwiek podzbiór nie jest zbiorem częstym

Antymonotoniczność



Przykład (1)

tr_id	produkt
1	coca-cola, orzeszki
2	orzeszki, pieluszki, piwo
3	coca-cola
4	coca-cola, orzeszki, piwo
5	orzeszki, pieluszki, piwo

- *minsup* = 30% (0,3)
- *minconf* = 70% (0,7)

Przykład (2)

$C_1 = L_1$

zbiór	wsparcie
coca-cola	0,6
orzeszki	0,8
pieluszki	0,4
piwo	0,6

L_2

zbiór	wsparcie
coca-cola, orzeszki	0,4
orzeszki, pieluszki	0,4
orzeszki, piwo	0,6
pieluszki, piwo	0,4

C_2

zbiór	wsparcie
coca-cola, orzeszki	0,4
coca-cola, pieluszki	0,0
coca-cola, piwo	0,2
orzeszki, pieluszki	0,4
orzeszki, piwo	0,6
pieluszki, piwo	0,4

Przykład (3)

C_3

zbiór	wsparcie
orzeszki, pieluszki, piwo	0,4

L_3

zbiór	wsparcie
orzeszki, pieluszki, piwo	0,4

$C_4 = \varnothing$ → $L_4 = \varnothing$

To jest koniec pierwszego etapu - generowania zbiorów częstych

W kolejnym kroku, na podstawie otrzymanych zbiorów częstych są generowane binarne reguły asocjacyjne zgodnie z algorytmem 1.2.
W kroku tym pomijane są 1-elementowe zbiory częste

Przykład (4)

Na podstawie zbioru L_2 można wygenerować następujący zbiór reguł:

zbiór częsty	wsparcie	reguła	ufność
pieluszki, piwo	0,4	pieluszki → piwo	0,67
pieluszki, piwo	0,4	piwo → pieluszki	1,00
orzeszki, piwo	0,6	orzeszki → piwo	0,75
orzeszki, piwo	0,6	piwo → orzeszki	1,00
pieluszki, orzeszki	0,4	pieluszki → orzeszki	1,00
pieluszki, orzeszki	0,4	orzeszki → pieluszki	0,50
coca-cola, orzeszki	0,4	coca-cola → orzeszki	0,67
coca-cola, orzeszki	0,4	orzeszki → coca-cola	0,50

Przykład (5)

Na podstawie zbioru L_3 można wygenerować następujący zbiór reguł:

zbiór częsty	wsparcie	reguła	ufność
orzeszki, pieluszki, piwo	0,4	orzeszki $\wedge$ pieluszki $\rightarrow$ piwo	1,00
orzeszki, pieluszki, piwo	0,4	orzeszki $\wedge$ piwo$\rightarrow$ pieluszki	0,67
orzeszki, pieluszki, piwo	0,4	pieluszki $\wedge$ piwo $\rightarrow$ orzeszki	0,67
orzeszki, pieluszki, piwo	0,4	orzeszki $\rightarrow$ pieluszki $\wedge$ piwo	0,50
orzeszki, pieluszki, piwo	0,4	pieluszki $\rightarrow$ orzeszki $\wedge$ piwo	1,00
orzeszki, pieluszki, piwo	0,4	piwo $\rightarrow$ orzeszki $\wedge$ pieluszki	0,67

Przykład (6)

Tylko kilka ze znalezionych reguł spełnia warunek minimalnej ufności.
Stąd, ostateczny wynik działania algorytmu Apriori jest następujący:

reguła asocjacyjna	**wsparcie**	**ufność**
piwo $\rightarrow$ pieluszki	0,4	1,00
orzeszki $\rightarrow$ piwo	0,6	0,75
piwo $\rightarrow$ orzeszki	0,6	1,00
pieluszki $\rightarrow$ orzeszki	0,4	1,00
orzeszki $\wedge$ pieluszki $\rightarrow$ piwo	0,4	1,00
pieluszki $\rightarrow$ orzeszki $\wedge$ piwo	0,4	1,00

Problem z regułami asocjacyjnymi

- Rozważmy tabelę przedstawiającą wyniki ankiety dotyczącej preferencji klientów w zakresie herbaty i kawy:

	kawa	nie-kawa	suma
herbata	20	5	25
nie-herbata	70	5	75
suma	90	10	100

- Znaleziono następującą regułę asocjacyjną:

herbata $\rightarrow$ kawa (support $= 20\%$, confidence $= 80\%$)

wsparcie $= 20/100 = 20\%$
ufność $= 20/25 = 80\%$

Czy miłośnicy herbaty piją kawę?

- „Kto lubi herbatę, najczęściej, lubi również kawę”
  - jest to reguła o dużym wsparciu i wysokiej ufności
  - 90% ankietowanych lubi kawę!!!

- Skąd różnica w wartości wsparcia?
  - ludzie, którzy lubią herbatę najczęściej nie lubią kawy
  - istnieje negatywna korelacja pomiędzy preferencją „lubię herbatę” i „lubię kawę”

$$\text{nie-herbata} \rightarrow \text{kawa (support} = 70\%, \text{confidence} = 70/75{=}93\%)$$

Korelacja

- Do oceny korelacji miedzy poprzednikiem i następnikiem reguły asocjacyjnej wykorzystuje się miarę *interest* lub miarę *lift*
- Dwa zdarzenia A i B są niezależne, jeżeli P(A B) = P(A)P(B), w przeciwnym razie zdarzenia A i B są skorelowane

$$\text{interest}(A \to B) = \frac{\text{support}(A \to B)}{\text{support}(A) \cdot \text{support}(B)} \qquad \text{lift}(A \to B) = \frac{\text{confidence}(A \to B)}{\text{support}(B)}$$

- Miara *interest* określa korelacje pomiędzy zdarzeniami A i B
  - *interest* = 1 – zdarzenia niezależne
  - *interest* < 1 – zdarzenia skorelowane negatywnie
  - *interest* > 1 – zdarzenia skorelowane pozytywnie

---

interest (herbata → kawa) = 0.89, zdarzenia skorelowane negatywnie