ACBC-kryteria oceny reguł i reprezentowanie reguł asocjacyjnych spełniających je z użyciem szablonów

Marzena Kryszkiewicz
Politechnika Warszawska

Miary ACBC oceny jakości reguł

Miara ACBC oceny jakości reguł

Miara oceny jakości reguły $X \rightarrow Y$ jest definiowana jako miara *ACBC*, jeśli może być wyrażona przy użyciu co najwyżej:

- prawdopodobieństwa $P(X)$ poprzednika (ang. *Antecedent*) X tej reguły,
- prawdopodobieństwa $P(Y)$ jej następnika (ang. *Consequent*) Y,
- prawdopodobieństwa $P(Z)$ jej bazy (ang. *Base*) $Z = X \cup Y$,
- Stałych (ang. *Constants*).

3

Przykłady miar ACBC oceny jakości reguł

measure	definition
$sup(X \rightarrow Y)$	$P(XY) \times \mid \mathcal{D} \mid$
$conf(X \rightarrow Y)$	$P(XY) / P(X)$
$novelty(X \rightarrow Y)$	$P(XY) - P(X) \times P(Y)$
$lift(X \rightarrow Y)$	$P(XY) / (P(X) \times P(Y))$
$certainty\text{-}Factor(X \rightarrow Y)$	$\begin{cases} \frac{conf(X \rightarrow Y) - P(Y)}{1 - P(Y)} & \text{if } conf(X \rightarrow Y) > P(Y) \\ 0 & \text{if } conf(X \rightarrow Y) = P(Y) = \\ -\frac{P(Y) - conf(X \rightarrow Y)}{P(Y) - 0} & \text{if } conf(X \rightarrow Y) < P(Y) \end{cases}$ $\begin{cases} \frac{P(XY) - P(X) \times P(Y)}{P(X) - P(X) \times P(Y)} & \text{if } P(XY) > P(X) \times P(Y) \\ 0 & \text{if } P(XY) = P(X) \times P(Y) \\ -\frac{P(X) \times P(Y) - P(XY)}{P(X) \times P(Y) - 0} & \text{if } P(XY) < P(X) \times P(Y) \end{cases}$
$cosine(X \rightarrow Y)$	$P(XY) / \sqrt{P(X) \times P(Y)}$
$Jaccard(X \rightarrow Y)$	$P(XY) / (P(X) + P(Y) - P(XY))$
$accuracy(X \rightarrow Y)$	$P(XY) + P(\bar{X}\bar{Y}) = 1 + 2P(XY) - P(X) - P(Y)$

- Wszystkie te miary oceny reguł $X \rightarrow Y$ (jak i wiele innych) można wyrazić przy użyciu co najwyżej $P(X)$, $P(Y)$, $P(XY)$ oraz stałych.

Uwaga: $P(\bar{X}\bar{Y}) = P(\bar{X}) - P(\bar{X}Y) = (1 - P(X)) - (P(Y) - P(XY))$.

4

Relatywne wsparcie reguły asocjacyjnej

- *Relatywne wsparcie* reguły asocjacyjnej $X \rightarrow Y$ jest definiowane jako prawdopodobieństwo współwystąpienia X i Y:

$$rSup(X \rightarrow Y) = P(XY) = sup(XY) / |D|.$$

- **Własność.** *Relatywne wsparcie* zależy od prawdopodobieństwa wystąpienia bazy reguły, ale nie zależy od prawdopodobieństwa wystąpienia jej poprzednika, ani następnika.

5

Zaufanie reguły asocjacyjnej

- *Zaufanie* reguły asocjacyjnej $X \rightarrow Y$ jest definiowane jako warunkowe prawdopodobieństwo wystąpienia Y pod warunkiem wystąpienia X:

$$conf(X \rightarrow Y) = P(XY) / P(X).$$

- **Własność.** *Zaufanie* zależy od prawdopodobieństw wystąpienia poprzednika i bazy reguły, ale nie zależy od prawdopodobieństwa wystąpienia jej następnika.

6

(Nie)zależność zdarzeń

- X i Y są *niezależne*, jeśli:
  $P(XY) = P(X) \times P(Y)$.
- Wpp. są zwane *zależnymi*.
- **Naturalna interpretacja:**
  - Jeżeli $P(XY) > P(X) * P(Y)$, to X i Y są zależne pozytywnie.
  - Jeżeli $P(XY) = P(X) * P(Y)$, to X i Y są niezależne.
  - Jeżeli $P(XY) < P(X) * P(Y)$, to X i Y są zależne negatywnie.
- **Własność.** Fakt, że $P(XY) \neq P(X) * P(Y)$ nie określa stopnia, w jakim X i Y są zależne.

7

Zaufanie a (nie)zależność zdarzeń…

TId	X	Y
1		x
2		x
3		x
4		x
5		x
6		x
7		x
8		x
9	x	x
10	x	

- $P(X) = \frac{2}{10}$, $P(Y) = \frac{9}{10} \Rightarrow P(X) \times P(Y) = 0.18$.
- $P(XY) = \frac{1}{10} = 0.1 < 0.18 = P(X) \times P(Y)$, czyli **$X$ i Y są zależne negatywnie.**
- ***conf($X \rightarrow Y$) = P(XY) / P(X) = $\frac{1}{2}$.***

- $P(X) = \frac{2}{10}$, $P(\overline{Y}) = \frac{1}{10} \Rightarrow P(X) \times P(\overline{Y}) = 0.02$.
- $P(X\overline{Y}) = \frac{1}{10} = 0.1 > 0.02 = P(X) \times P(\overline{Y})$, czyli **$X$ i $\overline{Y}$ są zależne pozytywnie.**
- ***conf($X \rightarrow \overline{Y}$) = P($X\overline{Y}$) / P(X) = $\frac{1}{2}$.***

8

Zaufanie a (nie)zależność zdarzeń

TId	X	Y
1		x
2		x
3		x
4		x
5		x
6		x
7		x
8	x	x
9	x	x
10	x	

- $P(X) = \frac{3}{10}$, $P(Y) = \frac{9}{10} \Rightarrow P(X) \times P(Y) = 0.27$.
- $P(XY) = \frac{2}{10} = 0.2 < 0.27 = P(X) \times P(Y)$, czyli **$X$ i Y są zależne negatywnie.**
- ***conf($X \rightarrow Y$) = P(XY) / P(X) = $\frac{2}{3}$.***

- $P(X) = \frac{3}{10}$, $P(\overline{Y}) = \frac{1}{10} \Rightarrow P(X) \times P(\overline{Y}) = 0.03$.
- $P(X\overline{Y}) = \frac{1}{10} = 0.1 > 0.03 = P(X) \times P(\overline{Y})$, czyli **$X$ i $\overline{Y}$ są zależne pozytywnie.**
- ***conf($X \rightarrow \overline{Y}$) = P($X\overline{Y}$) / P(X) = $\frac{1}{3}$.***

9

Współczynnik podniesienia (lift)

- *Współczynnik podniesienia* (ang. *lift*) dla reguły asocjacyjnej $X \rightarrow Y$ jest definiowany jako stosunek prawdopodobieństwa warunkowego wystąpienia Y pod warunkiem wystąpienia X do prawdopodobieństwa wystąpienia Y:

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{P(Y)} = \frac{P(XY)}{P(X) \times P(Y)}.$$

- **Własność.**

(Nie)zależność	Warunek na (nie)zależność	Równoważny warunek	Równoważny warunek
Y i X są zależne pozytywnie	$P(XY) > P(X) \times P(Y)$	$lift(X \rightarrow Y) > 1$	$lift(Y \rightarrow X) > 1$
Y i X są niezależne	$P(XY) = P(X) \times P(Y)$	$lift(X \rightarrow Y) = 1$	$lift(Y \rightarrow X) = 1$
Y i X są zależne negatywnie	$P(XY) < P(X) \times P(Y)$	$lift(X \rightarrow Y) < 1$	$lift(Y \rightarrow X) < 1$

- **Własność.** $lift(X \rightarrow Y) = lift(Y \rightarrow X)$.

10

Lift a (nie)zależność zdarzeń

TId	X	Y	Z	V
1	x	x		
2	x	x		
3	x	x		
4	x	x		
5	x	x		
6	x	x		
7	x	x		
8	x	x		
9	x	x		
10			x	x

- $P(X) = \frac{9}{10}$, $P(Y) = \frac{9}{10} \Rightarrow P(X) \times P(Y) = 0.81$.
- $P(XY) = \frac{9}{10} = 0.9 > 0.81 = P(X) \times P(Y)$, czyli **$X$ i Y są zależne pozytywnie.**
- ***lift($X \rightarrow Y$)*** $= \frac{P(XY)}{P(X) \times P(X)} = \frac{0.9}{0.9 \times 0.9} = \frac{10}{9} = 1\frac{1}{9}$.

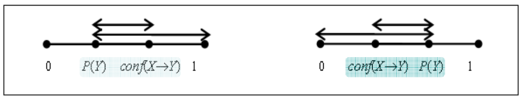
- $P(Z) = \frac{1}{10}$, $P(V) = \frac{1}{10} \Rightarrow P(Z) \times P(V) = 0.01$.
- $P(ZV) = \frac{1}{10} = 0.1 > 0.01 = P(Z) \times P(V)$, czyli **$Z$ i V są zależne pozytywnie.**
- ***lift($Z \rightarrow V$)*** $= \frac{P(ZV)}{P(Z) \times P(V)} = \frac{0.1}{0.1 \times 0.1} = \frac{10}{1} = 10$.

11

Współczynnik pewności (cf)…

- Współczynnik pewności (ang. *certainty factor*, *cf*) dla reguły asocjacyjnej $X \rightarrow Y$ jest definiowany jako stopień, w jakim prawdopodobieństwo wystąpienia Y może ulec zmianie pod warunkiem wystąpienia X:

$$cf(X \rightarrow Y) = \begin{cases} \frac{conf(X \rightarrow Y) - P(Y)}{1 - P(Y)} & \text{if } conf(X \rightarrow Y) > P(Y), \\ 0 & \text{if } conf(X \rightarrow Y) = P(Y), \\ -\frac{P(Y) - conf(X \rightarrow Y)}{P(Y) - 0} & \text{if } conf(X \rightarrow Y) < P(Y). \end{cases}$$



12

Współczynnik pewności (cf)

$$cf(X \to Y) = \begin{cases} \frac{conf(X \to Y) - P(Y)}{1 - P(Y)} & \text{if } conf(X \to Y) > P(Y), \\ 0 & \text{if } conf(X \to Y) = P(Y), \\ -\frac{P(Y) - conf(X \to Y)}{P(Y) - 0} & \text{if } conf(X \to Y) < P(Y). \end{cases}$$

- **Własność:**

asymetryczne dla X i Y

$$cf(X \to Y) = \begin{cases} \frac{P(XY) - P(X) \times P(Y)}{P(X) - P(X) \times P(Y)} & \text{if } P(XY) > P(X) \times P(Y), \\ 0 & \text{if } P(XY) = P(X) \times P(Y), \\ -\frac{P(X) \times P(Y) - P(XY)}{P(X) \times P(Y) - 0} & \text{if } P(XY) < P(X) \times P(Y). \end{cases}$$

symetryczne dla X i Y

13

Wartości ekstremalne miar regułowych

miara	maks.	min.
$P(XY)$	1	0
$conf(X \to Y)$	1	0
$lift(X \to Y)$	∞	0
$cf(X \to Y)$	1 dla zależności pozytywnej pomiędzy X i Y	-1 dla zależności negatywnej pomiędzy X i Y

14

Wartości ekstremalne P(XY) dla zadanych P(X) & P(Y)…

- Własność.

a) $max\_P(XY|_{P(X),\,P(Y)}) = \min\{P(X), P(Y)\}$

b) $min\_P(XY|_{P(X),\,P(Y)}) = \begin{cases} 0 & \text{if } P(X) + P(Y) \le 1 \\ P(X) + P(Y) - 1 & \text{if } P(X) + P(Y) > 1 \end{cases}$

$= \max\{0,\ P(X) + P(Y) - 1\}$

X	*Y*
x	x
x	x
x	

X	*Y*
x	
x	
x	
	x
	x

X	*Y*
x	
x	
x	x
x	x
x	x
	x

15

Wartości ekstremalne P(XY) dla zadanych P(X) & P(Y)

- Własność.

zamiast 1

zamiast 0

miara	maks. dla ustalonych wartości $P(X)$ i $P(Y)$	min. dla ustalonych wartości $P(X)$ i $P(Y)$
$P(XY)$	$\min\{P(X), P(Y)\}$	$\max\{0,\ P(X) + P(Y) - 1\}$
$conf(X \to Y)$	$\frac{\min\{P(X), P(Y)\}}{P(X)}$	$\frac{\max\{0, P(X) + P(Y) - 1\}}{P(X)}$
$lift(X \to Y)$	$\frac{\min\{P(X), P(Y)\}}{P(X) \times P(Y)}$	$\frac{\max\{0, P(X) + P(Y) - 1\}}{P(X) \times P(Y)}$
$cf(X \to Y)$	$\frac{\min\{P(X), P(Y)\} - P(X) \times P(Y)}{P(X) - P(X) \times P(Y)}$ dla zależności pozytywnej pomiędzy X i Y	$-\frac{P(X) \times P(Y) - \max\{0, P(X) + P(Y) - 1\}}{P(X) \times P(Y) - 0}$ dla zależności negatywnej pomiędzy X i Y

16

Współczynnik zależności (df)

- *Współczynnik zależności* (ang. *dependence factor, df*) dla reguły asocjacyjnej $X \to Y$ jest definiowany jako stopień, w jakim prawdopodobieństwo wystąpienia Y może ulec zmianie pod warunkiem wystąpienia X i Y z prawdopodobieństwem odpowiednio $P(X)$ i $P(Y)$:

zamiast 1

$$df(X \to Y) = \begin{cases} \frac{conf(X \to Y) - P(Y)}{max\_conf(X \to Y|_{P(X),P(Y)}) - P(Y)} & \text{if } conf(X \to Y) > P(Y), \\ 0 & \text{if } conf(X \to Y) = P(Y), \\ -\frac{P(Y) - conf(X \to Y)}{P(Y) - min\_conf(X \to Y|_{P(X),P(Y)})} & \text{if } conf(X \to Y) < P(Y). \end{cases}$$

zamiast 0

17

Ekstremalne wartości współczynnika zależności df

- **Własność.**

miara	maks. (dla pozytywnej zależności pomiędzy X i Y)	min. (dla negatywnej zależności pomiędzy X i Y)
$df(X \to Y)$ przy ustalonych wartościach $P(X)$ i $P(Y)$	1	-1

zamiast 1

$$df(X \to Y) = \begin{cases} \frac{conf(X \to Y) - P(Y)}{max\_conf(X \to Y|_{P(X),P(Y)}) - P(Y)} & \text{if } conf(X \to Y) > P(Y), \\ 0 & \text{if } conf(X \to Y) = P(Y), \\ -\frac{P(Y) - conf(X \to Y)}{P(Y) - min\_conf(X \to Y|_{P(X),P(Y)})} & \text{if } conf(X \to Y) < P(Y). \end{cases}$$

zamiast 0

18

Współczynnik zależności df w terminach prawdopodobieństw…

zamiast 1

$$df(X \rightarrow Y) = \begin{cases} \dfrac{P(XY) - P(X) \times P(Y)}{max\_P(XY|_{P(X),P(Y)}) - P(X) \times P(Y)} & \text{if } P(XY) > P(X) \times P(Y), \\ 0 & \text{if } P(XY) = P(X) \times P(Y), \\ -\dfrac{P(X) \times P(Y) - P(XY)}{P(X) \times P(Y) - min\_P(XY|_{P(X),P(Y)})} & \text{if } P(XY) < P(X) \times P(Y). \end{cases}$$

zamiast 0

19

Współczynnik zależności df w terminach prawdopodobieństw

zamiast 1

$$df(X \rightarrow Y) = \begin{cases} \dfrac{P(XY) - P(X) \times P(Y)}{max\_P(XY|_{P(X),P(Y)}) - P(X) \times P(Y)} & \text{if } P(XY) > P(X) \times P(Y), \\ 0 & \text{if } P(XY) = P(X) \times P(Y), \\ -\dfrac{P(X) \times P(Y) - P(XY)}{P(X) \times P(Y) - min\_P(XY|_{P(X),P(Y)})} & \text{if } P(XY) < P(X) \times P(Y). \end{cases}$$

zamiast 0

- **Własność.** $df(X \rightarrow Y) = df(Y \rightarrow X)$.

20

Współczynnik zależności df a współczynnik pewności cf

- **Własność.**
  - $df(X \rightarrow Y) = cf(X \rightarrow Y) = 0$, jeśli $P(XY) = P(X) \times P(Y)$.
  - $|df(X \rightarrow Y)| \geq |cf(X \rightarrow Y)|$, jeśli $P(XY) \neq P(X) \times P(Y)$.
  - If $P(XY) > P(X) \times P(Y)$, then
    $df(X \rightarrow Y) = \max\{cf(X \rightarrow Y), cf(Y \rightarrow X)\}$.
  - If $P(XY) < P(X) \times P(Y)$, then:
    $df(X \rightarrow Y) = cf(X \rightarrow Y)$, jeśli $P(X) + P(Y) \leq 1$.
    $|df(X \rightarrow Y)| > |cf(X \rightarrow Y)|$, jeśli $P(X) + P(Y) > 1$.

21

Współczynnik zależności: $P(X) + P(Y) \leq 1$

- **Tabela.** Porówanie wartości $lift(X \rightarrow Y)$, $cf(X \rightarrow Y)$ oraz $df(X \rightarrow Y)$, gdy $P(X) + P(Y) \leq 1$

$P(X)$	$P(Y)$	$P(XY)$	$P(X) \times P(Y)$	$lift(X \rightarrow Y)$	$cf(X \rightarrow Y)$	$cf(Y \rightarrow X)$	$df(X \rightarrow Y) = df(Y \rightarrow X)$
0.60	**0.30**	**0.30**	**0.18**	**1.67**	**0.29**	**1.00**	**1.00**
0.60	0.30	0.25	0.18	1.39	0.17	0.58	0.58
0.60	0.30	0.20	0.18	1.11	0.05	0.17	0.17
0.60	0.30	0.18	0.18	1.00	0.00	0.00	0.00
0.60	0.30	0.15	0.18	0.83	-0.17	-0.17	-0.17
0.60	0.30	0.10	0.18	0.56	-0.44	-0.44	-0.44
0.60	0.30	0.00	0.18	0.00	-1.00	-1.00	-1.00

22

Współczynnik zależności: $P(X) + P(Y) > 1$

- **Tabela.** Porówanie wartości $lift(X \rightarrow Y)$, $cf(X \rightarrow Y)$ oraz $df(X \rightarrow Y)$, gdy $P(X) + P(Y) > 1$

$P(X)$	$P(Y)$	$P(XY)$	$P(X) \times P(Y)$	$lift(X \rightarrow Y)$	$cf(X \rightarrow Y)$	$cf(Y \rightarrow X)$	$df(X \rightarrow Y) = df(Y \rightarrow X)$
0.80	0.60	0.60	0.48	1.25	0.38	1.00	1.00
0.80	0.60	0.55	0.48	1.15	0.22	0.58	0.58
0.80	0.60	0.50	0.48	1.04	0.06	0.17	0.17
0.80	0.60	0.48	0.48	1.00	0.00	0.00	0.00
0.80	0.60	0.45	0.48	0.94	-0.06	-0.06	-0.37
0.80	**0.60**	**0.40**	**0.48**	**0.83**	**-0.17**	**-0.17**	**-1.00**

23

Współczynnik zależności a miara lift

zamiast ∝

$$df(X \rightarrow Y) = \begin{cases} \dfrac{lift(X \rightarrow Y) - 1}{max\_lift(X \rightarrow Y|_{P(X),P(Y)}) - 1} & \text{if } lift(X \rightarrow Y) > 1, \\ 0 & \text{if } lift(X \rightarrow Y) = 1, \\ -\dfrac{1 - lift(X \rightarrow Y)}{1 - min\_lift(X \rightarrow Y|_{P(X),P(Y)})} & \text{if } lift(X \rightarrow Y) < 1. \end{cases}$$

zamiast 0

24

Zależność pomiędzy *X* i *Y*, a zależności pomiędzy $\bar{X}$ i *Y*, *X* i $\bar{Y}$, $\bar{X}$ i $\bar{Y}$

- **Twierdzenie.**

$$df(X \to Y) = df(\bar{X} \to \bar{Y}) = -df(X \to \bar{Y}) = -df(\bar{X} \to Y).$$

25

Reprezentacja szablonowa reguł spełniających różnorodne kryteria oparte na miarach ACBC

Reguły spełniające różnorodne kryteria oparte na miarach ACBC

Niech:

- M będzie zbiorem $\{\mu_1, ..., \mu_n\}$ miar *ACBC*

oraz

- $E = \{\varepsilon_1,...,\varepsilon_n\}$ będzie zbiorem odpowiadających im wartości progowych.

Reguła $X \to Y$ jest definiowana jako *silna ze względu na miary M i wartości progowe E* (krótko: jako (M, E)-*reguła*), jeżeli:

$$\forall \mu_i \in M\ (\mu_i(X \to Y) > \varepsilon_i).$$

27

Własność miary ACBC

Własność. Niech:

- $X \to Y$ i $Z \to V$ będą regułami, dla których zachodzi:

$$P(X) = P(Z),\ P(Y) = P(V),\ P(XY) = P(ZV)$$

oraz

- μ będzie miarą *ACBC*.

Wtedy:

$$\mu(X \to Y) = \mu(Z \to V).$$

28

Reprezentowanie reguł spełniających wiele kryteriów opartych ma miarach ACBC

- Liczba reguł spełniających wiele kryteriów opartych ma miarach ACBC może być olbrzymia.
- Szablony regułowe mogą być użyte do reprezentowania reguł spełniających różnorodne kryteria oparte ma miarach ACBC.
- Szablony regułowe są tworzone z użyciem dwóch typów zbiorów:
  - *zbiorów zamkniętych* oraz
  - *generatorów kluczowych.*

29

Zbiory zamknięte - przypomnienie

- *Domknięciem zbioru X* o niezerowym wsparciu (czyli zawartym w co najmniej jednej transakcji) jest przecięcie wszystkich transakcji zawierających *X*. Wpp. *domknięciem zbioru X* jest uniwersum (zbiór wszystkich pozycji).
- **Własność.** *Domknięcie zbioru X* jest największym nadzbiorem zbioru *X*, występującym w tych samych transakcjach co zbiór X (czyli występującym w tej samej liczbie transakcji co zbiór *X*).
- *Zbiór X* jest definiowany jako *zbiór zamknięty*, jeśli *X* jest swoim domknięciem; czyli, jeśli:

$$\forall Y \supset X,\ t(Y) \subset t(X)$$

(lub równoważnie, $\forall Y \supset X$, $sup(Y) < sup(X)$).

30

Istotna własność zbiorów zamkniętych

Niech X będzie zbiorem, a Y zbiorem zamkniętym, stanowiącym domknięcie zbioru X. Wtedy:

- X i Y występują w tych samych transakcjach.
- X i Y maja to samo wsparcie.
- Y jest domknięciem każdego zbioru U, spełniającego warunek: $X \subseteq U \subseteq Y$.
- Każdy zbiór U, spełniający warunek: $X \subseteq U \subseteq Y$, ma to samo wsparcie co zbiór Y.

31

Generatory kluczowe - przypomnienie

- Zbiór Y jest definiowany jako *generator* zbioru X, jeśli Y jest minimalnym podzbiorem zbioru X, który występuje w tych samych transakcjach co zbior X.
- **Własność.** Każdy zbiór ma co najmniej jeden generator.
- Zbiór X jest definiowany jako *generator* kluczowy, jeśli X jest swoim (jedynym) generatorem, czyli, jeżeli:

$$\forall Y \subset X,\ t(Y) \supset t(X)$$

(lub równoważnie, $\forall Y \subset X,\ sup(Y) > sup(X)$).

32

Przykład: Generatory kluczowe i zbiory zamknięte

Tabela 1. Przykładowy zbiór D z transakcjami

Transaction Id	Transaction
#1	{*abcde*}
#2	{*abcdef*}
#3	{*abcdehi*}
#4	{*abe*}
#5	{*bcdehi*}

Tabela 2. Częste zbiory zamknięte i częste generatory kluczowe odkryte w zbiorze D transakcji

Częste zbiory zamknięte	Częste generatory kluczowe	Wsparcie relatywne
{*be*}	$\emptyset$	5/5
{*abe*}	{*a*}	4/5
{*bcde*}	{*c*}, {*d*}	4/5
{*abcde*}	{*ac*}, {*ad*}	3/5
{*bcdehi*}	{*h*}, {*i*}	2/5
{*abcdef*}	{*f*}	1/5
{*abcdehi*}	{*ah*}, {*ai*}	1/5

33

Przykład: Wnioskowanie z użyciem generatorów kluczowych i zbiorów zamkniętych

Transaction Id	Transaction
#1	{*abcde*}
#2	{*abcdef*}
#3	{*abcdehi*}
#4	{*abe*}
#5	{*bcdehi*}

- {*h*} jest generatorem kluczowym.
- {*bcdehi*} jest zbiorem zamkniętym, który jest domknięciem zbioru {*h*}.
- Każdy zbiór U, spełniający warunek: $\{h\} \subseteq U \subseteq \{bcdehi\}$, ma takie samo wsparcie co zbiór {*h*} i zbiór {*bcdehi*} (tutaj: 2). Liczba takich zbiorów U wynosi 32.

34

Szablony regułowe

Para dwóch reguł ($X \to Y$, $Z \to V$) jest definiowana jako szablon regułowy (ang. *rule template*), jeśli:

- X i Y są generatorami kluczowymi oraz $X \cap Y = \emptyset$,
- Z i V są zbiorami zamkniętymi,
- Z jest domknięciem zbioru X, a V jest domknięciem zbioru Y.

$X \to Y$ jest zwana *regułą dolną* szablonu ($X \to Y$, $Z \to V$).
$Z \to V$ jest zwana *regułą górną* szablonu ($X \to Y$, $Z \to V$).

35

Szablony regułowe i ich własności

Reguła $U \to W$ jest definiowana jako *pokryta przez szablon regułowy* ($X \to Y$, $Z \to V$), jeśli $X \subseteq U \subseteq Z$ i $Y \subseteq W \subseteq V$.

Twierdzenie. Niech ($X \to Y$, $Z \to V$) będzie szablonem regułowym pokrywającym regułę $U \to W$. Wtedy:

- $P(X) = P(U) = P(Z)$.
- $P(Y) = P(W) = P(V)$.
- $P(XY) = P(UW) = P(ZV)$.
- Dla każdej ewaluacyjnej miary ACBCμ:

$$\mu(X \to Y) = \mu(U \to W) = \mu(Z \to V).$$

36

Moc pokrycia szablonu regułowego

Twierdzenie. Szablon regułowy ($X \rightarrow Y$, $Z \rightarrow V$) pokrywa

$2^m \times 2^n \times 3^k$ unikatowych reguł asocjacyjnych, gdzie:

- $\Delta_{Common} = (Z \cap V) \setminus (X \cup Y)$,
- $\Delta_{antecedent} = Z \setminus (X \cup V)$;
- $\Delta_{consequent} = V \setminus (Y \cup Z)$;
- $k = |\Delta_{Common}|$; $m = |\Delta_{antecedent}|$; $n = |\Delta_{consequent}|$.

Przykład. Niech *R* oznacza szablon regułowy:

$(\{a\} \rightarrow \{ij\}, \{abcdef\} \rightarrow \{efghij\})$.

Wtedy: $\Delta_{Common} = \{ef\}$, $\Delta_{antecedent} = \{bcd\}$, $\Delta_{consequent} = \{gh\}$.

Zatem, szablon regułowy R pokrywa $2^{|\{bcd\}|} \times 2^{|\{gh\}|} \times 3^{|\{ef\}|} = 2^3 \times 2^2 \times 3^2 = 288$ reguł asocjacyjnych.

37

(M, E)-szablony regułowe

- Niech M będzie zbiorem miar regułowych ACBC. Szablon regułowy ($X \rightarrow Y$, $Z \rightarrow V$) jest nazywany (M, E)-szablonem regułowym, jeśli:

$$\forall \mu_i \in M,\ \varepsilon_i \in E\ (\mu_i(X \rightarrow Y) > \varepsilon_i)$$

(lub alternatywnie, $\forall \mu_i \in M,\ \varepsilon_i \in E\ (\mu_i(Z \rightarrow V) > \varepsilon_i)$).

38

(M, E)-szablony regułowe jako reprezentacja (M, E)-silnych reguł asocjacyjnych

Twierdzenie. Niech M będzie zbiorem miar regułowych ACBC, a $U \rightarrow W$ będzie regułą asocjacyjną.

- Jeśli istnieje (M, E)-szablon regułowy ($X \rightarrow Y$, $Z \rightarrow V$) pokrywający $U \rightarrow W$, to $U \rightarrow W$ jest (M, E)-regułą asocjacyjną oraz

$$\forall \mu_i \in M\ (\mu_i(U \rightarrow W) = \mu_i(X \rightarrow Y) = \mu_i(Z \rightarrow V)).$$

- Wpp., $U \rightarrow W$ nie jest (M, E)-regułą asocjacyjną .

39

Przykład: (M, E)-szablony regułowe jako reprezentacja (M, E)-reguł asocjacyjnych

Zbiór danych D

TId	Transaction
#1	{abcde}
#2	{abcdef}
#3	{abcdehi}
#4	{abe}
#5	{bcdehi}

($\mathcal{M}$, E)-szablon regułowy	rSup	novelty	lift	certaintyFactor
$(\{h\}\rightarrow\{i\}, \{bcdehi\}\rightarrow\{bcdehi\})$	2/5	0.24	2.5	1
$(\{h\}\rightarrow\{ai\}, \{bcdehi\}\rightarrow\{abcdehi\})$	1/5	0.12	2.5	0.375
$(\{i\}\rightarrow\{h\}, \{bcdehi\}\rightarrow\{bcdehi\})$	2/5	0.24	2.5	1
$(\{i\}\rightarrow\{ah\}, \{bcdehi\}\rightarrow\{abcdehi\})$	1/5	0.12	2.5	0.375
$(\{ah\}\rightarrow\{i\}, \{abcdehi\}\rightarrow\{bcdehi\})$	1/5	0.12	2.5	1
$(\{ai\}\rightarrow\{h\}, \{abcdehi\}\rightarrow\{bcdehi\})$	1/5	0.12	2.5	1

- Niech zbiór miar regułowych ACBC
  M = {*rSup*, *novelty*, *lift*, *certaintyFactor*} oraz
- zbiór odpowiadających im wartości progowych
  E = {0, 0.1, 2.0, 0.3}.

Wtedy jest 6 (M, E)-szablonów regułowych, które pokrywają 486 (M, E)-reguł ascocjacyjnych.

40

Literatura

- Marzena Kryszkiewicz: Concise Representations of Frequent Patterns and Association Rules, Warsaw: Publishing House of Warsaw University of Technology (2002)
- Marzena Kryszkiewicz: A Lossless Representation for Association Rules Satisfying Multiple Evaluation Criteria. Proceedings of ACIIDS 2016, LNCS, vol. 9622, 2016, Springer
- Marzena Kryszkiewicz: ACBC-Adequate Association and Decision Rules Versus Key Generators and Rough Sets Approximations. Fundam. Inform. 148(1-2): 65-85 (2016)
- Marzena Kryszkiewicz: Representative Rule Templates for Association Rules Satisfying Multiple Canonical Evaluation Criteria. ACIIDS (1) 2018: 550-561

41

Ćwiczenia…

1. Pokaż, że współczynnik zależności df (ang. dependence factor) jest miarą ACBC.

2. Niech zbiór ACBC-miar regułowych:
   - M = {*rSup*, *conf*, *dependenceFactor*}

     oraz zbiór odpowiadających im wartości progowych:
   - E = {2/5, 0.45, 1}.

     Wyznacz (M, E)-szablony regułowe na podstawie zbioru danych ze slajdu 33.

42

Ćwiczenia

3. Wyznacz liczbę reguł asocjacyjnych pokrytych przez szablon regułowy ({*ah*}→{*i*}, {*abcdehi*}→{*bcdehi*}).

4. Wyznacz zbiór reguł asocjacyjnych pokrytych przez szablon regułowy ({*ah*}→{*i*}, {*abcdehi*}→{*bcdehi*}).

5. Na podstawie zbioru szablonów reguł zamieszczonych na slajdzie 40 wyznacz te szablony, które pokrywają regułę {*bch*}→{*aei*} oraz na tej podstawie wyznacz wartości miar *rSup*, *novelty*, *lift* i *certaintyFactor* dla tej reguły.

43