Minería de Datos

Reglas de asociación representativas

Ángel Ríos San Nicolás

20 de marzo de 2021

Ejercicio. Se da el siguiente conjunto de datos $\mathcal{D} = \{abc, abc, abc, ab, ab, ab, ab, bc, a, b, c\}$. Se pide calcular F_τ, FC_τ, FG_τ, $RI_{\tau,\gamma}$ y $RR_{\tau,\gamma}$ para $\tau = 0,25$ y $\gamma = 0,75$ (los dos en valores relativos).

Solución. El conjunto de items es $\mathcal{I} = \{a, b, c\}$ que tiene $\#\left(2^{\mathcal{I}}\right) = 2^3 = 8$ subconjuntos posibles.

$$2^{\mathcal{I}} = \{\emptyset, a, b, c, ab, ac, bc, abc\}$$

Calculamos los soportes absolutos y relativos teniendo en cuenta que el número total de datos es $\#(\mathcal{D}) = 13$.

Subconjunto X	Soporte absoluto $s(X)$	Soporte relativo $s_r(X)$
$\emptyset$	13	1
a	10	0.7692307692307693
b	11	0.8461538461538461
c	6	0.4615384615384615
ab	9	0.6923076923076923
ac	4	0.3076923076923077
bc	5	0.3846153846153846
abc	4	0.3076923076923077

Como todos los subconjuntos tienen soporte relativo mayor que $\tau = 0.25$, todos los conjuntos son frecuentes, es decir, $F_\tau = \{X \subsetneq \mathcal{I} : s_r(X) \geq \tau\} = 2^{\mathcal{I}} = \{\emptyset, a, b, c, ab, ac, bc, abc\}$.

Calculamos los conjuntos cerrados frecuentes $FC_\tau = \{X \in F_\tau : \forall Z \supsetneq X, s(Z) < s(X)\}$.

- abc

  Ningún conjunto lo contiene estrictamente, con lo que se cumple la propiedad y $abc \in FC_\tau$.

- bc

  Únicamente $abc \subsetneq bc$ y se cumple $s(abc) = 4 < 5 = s(bc)$, con lo que $bc \in FC_\tau$.

- ac

  Únicamente $abc \subsetneq ac$ y se cumple $s(abc) = 4 = s(ac)$, con lo que $ac \notin FC_\tau$.

- ab

  Únicamente $abc \subsetneq ab$ y se cumple $s(abc) = 4 < 9 = s(ab)$, con lo que $ab \in FC_\tau$.

- c

  Tenemos que $\left.\begin{array}{rr} abc \supsetneq c & s(abc) = 4 \\ ac \supsetneq c \text{ y se cumplen} & s(ac) = 4 \\ bc \supsetneq c & s(bc) = 5 \end{array}\right\} < 6 = s(c)$, con lo que $c \in FC_\tau$.

- b

  Tenemos que $\left.\begin{array}{rcr} abc \supsetneq b & & s(abc) = 4 \\ ab \supsetneq b & \text{y se cumplen} & s(bc) = 5 \\ bc \supsetneq b & & s(bc) = 5 \end{array}\right\} < 11 = s(b)$, con lo que $b \in FC_\tau$.

- a

  Tenemos que $\left.\begin{array}{rcr} abc \supsetneq a & & s(abc) = 4 \\ ab \supsetneq a & \text{y se cumplen} & s(ab) = 9 \\ ac \supsetneq a & & s(ac) = 4 \end{array}\right\} < 10 = s(a)$, con lo que $a \in FC_\tau$.

- $\emptyset$

  No hay conjuntos cerrados no vacíos con soporte 13, con lo que $\emptyset \in FC_\tau$.

Por lo tanto, el conjunto de cerrados frecuentes es $FC_\tau = \{abc, bc, ab, c, b, a, \emptyset\}$.

Calculamos los generadores minimales frecuentes $FG_\tau = \{X \in F_\tau : \forall Y \subsetneq X, s(Y) > s(X)\}$.

- $\emptyset$

  Ningún conjunto está contenido estrictamente, con lo que $\emptyset \in FG_\tau$.

- a

  Únicamente $\emptyset \subsetneq a$ y se cumple $s(\emptyset) = 13 > 10 = s(a)$, con lo que $a \in FG_\tau$.

- b

  Únicamente $\emptyset \subsetneq b$ y se cumple $s(\emptyset) = 13 > 11 = s(b)$, con lo que $b \in FG_\tau$.

- c

  Únicamente $\emptyset \subsetneq c$ y se cumple $s(\emptyset) = 13 > 6 = s(c)$, con lo que $c \in FG_\tau$.

- ab

  Tenemos que $\left.\begin{array}{rcr} \emptyset \subsetneq ab & & s(\emptyset) = 13 \\ a \subsetneq ab & \text{y se cumple} & s(a) = 10 \\ b \subsetneq ab & & s(b) = 11 \end{array}\right\} > 9 = s(ab)$, con lo que $ab \in FG_\tau$.

- ac

  Tenemos que $\left.\begin{array}{rcr} \emptyset \subsetneq ac & & s(\emptyset) = 13 \\ a \subsetneq ac & \text{y se cumple} & s(a) = 10 \\ c \subsetneq ac & & s(c) = 6 \end{array}\right\} > 4 = s(ac)$, con lo que $ac \in FG_\tau$.

- bc

  Tenemos que $\left.\begin{array}{rcr} \emptyset \subsetneq bc & & s(\emptyset) = 13 \\ b \subsetneq bc & \text{y se cumple} & s(b) = 11 \\ c \subsetneq bc & & s(c) = 6 \end{array}\right\} > 5 = s(bc)$, con lo que $bc \in FG_\tau$.

- abc

  Consideramos $ac \subsetneq abc$ que cumple $s(ac) = 4 = s(abc)$, con lo que $abc \notin FG_\tau$.

Por lo tanto, el conjunto de generadores minimales frecuentes es $FG_\tau = \{\emptyset, a, b, c, ab, ac, bc\}$.

Calculamos el conjunto $RI_{\tau,\gamma} = \{Z \in FC_\tau : \gamma \cdot \text{mxgs}_{\tau,\gamma}(Z) > \text{mxs}_\tau(Z)\}$. Para ello, necesitamos calcular primero mxs_τ para cada cerrado frecuente.

$$\text{mxs}_\tau(X) = \max(\{s(Z) : Z \in FC_\tau, Z \supsetneq X\} \cup \{0\})$$

El conjunto abc no está contenido estrictamente en ningún otro, con lo que $\text{mxs}_\tau(abc) = 0$. Los conjuntos ab y bc están contenidos estrictamente únicamente en abc, así que

$$\text{mxs}_\tau(ab) = \text{mxs}_\tau(bc) = s(abc) = 4.$$

Respecto a los unipuntuales

$$\begin{aligned}\text{mxs}_\tau(a) &= \max(\{s(ab), s(ac), s(abc)\} \cup \{0\}) = \max(\{9, 4, 4, 0\}) = 9\\ \text{mxs}_\tau(b) &= \max(\{s(ab), s(bc), s(abc)\} \cup \{0\}) = \max(\{9, 5, 4, 0\}) = 9\\ \text{mxs}_\tau(c) &= \max(\{s(ac), s(bc), s(abc)\} \cup \{0\}) = \max(\{4, 5, 4, 0\}) = 5\end{aligned}$$

El vacío está contenido estrictamente en todos los subconjuntos no vacíos y tenemos

$$\text{mxs}_\tau(\emptyset) = \max(\{s(Z) : Z \in FC_\tau\} = \max(\{10, 11, 6, 9, 4, 5, 4, 0\}) = 11.$$

Tenemos que calcular también $\text{mxgs}_{\tau,\gamma}$ para cada cerrado frecuente.

$$\text{mxgs}_{\tau,\gamma}(X) = \max(\{s(Y) : Y \in FG_\tau, Y \subsetneq X, \gamma \cdot s(Y) \leq s(X)\} \cup \{0\})$$

- $\emptyset$

  No hay conjuntos estrictamente contenidos en el vacío, con lo que $\text{mxgs}_{\tau,\gamma}(\emptyset) = 0$.

- c

  Únicamente $\emptyset \subsetneq c$, pero $0,75 \cdot s(\emptyset) = 0,75 \cdot 13 = 9,75 \not\leq 6 = s(c)$, con lo que $\text{mxgs}_{\tau,\gamma}(c) = 0$.

- b

  Únicamente $\emptyset \subsetneq b$, y se cumple $0,75 \cdot s(\emptyset) = 0,75 \cdot 13 = 9,75 \leq 11 = s(b)$.

  $$\text{mxgs}_{\tau,\gamma}(b) = \max(\{s(\emptyset)\} \cup \{0\}) = \max(\{13, 0\}) = 13$$

- a

  Únicamente $\emptyset \subsetneq a$, y se cumple $0,75 \cdot s(\emptyset) = 0,75 \cdot 13 = 9,75 \leq 10 = s(a)$.

  $$\text{mxgs}_{\tau,\gamma}(a) = \max(\{s(\emptyset)\} \cup \{0\}) = \max(\{13, 0\}) = 13$$

- bc

  Se debe cumplir $0,75 \cdot s(Y) \leq s(bc) = 5$, es decir $s(Y) \leq \frac{5}{0,75} = 6,\overline{6}$ con $Y \subsetneq bc$.

  $$\text{mxgs}_{\tau,\gamma}(bc) = \max(\{s(c)\} \cup \{0\}) = \max(\{6, 0\}) = 6$$

- ab

  Se debe cumplir $0,75 \cdot s(Y) \leq s(ab) = 9$, es decir $s(Y) \leq \frac{9}{0,75} = 12$ con $Y \subsetneq ab$.

  $$\text{mxgs}_{\tau,\gamma}(ab) = \max(\{s(a), s(b)\} \cup \{0\}) = \max(\{10, 11, 0\}) = 11$$

- abc

  Se debe cumplir $0,75 \cdot s(Y) \leq s(abc) = 4$, es decir $s(Y) \leq \frac{4}{0,75} = 5,\overline{3}$ con $Y \subsetneq abc$.

  $$\text{mxgs}_{\tau,\gamma}(abc) = \max(\{s(ac), s(bc)\} \cup \{0\}) = \max(\{5, 4, 0\}) = 5$$

FC_τ	mxs_τ	$\gamma \cdot \text{mxgs}_{\tau,\gamma}$	$\text{mxgs}_{\tau,\gamma}$
abc	0	$3,75$	5
ab	4	$8,25$	11
bc	4	$4,5$	6
a	9	$9,75$	13
b	5	$9,75$	13
c	5	0	0
$\emptyset$	11	0	0

Por lo tanto, $RI_{\tau,\gamma} = \{abc, ab, bc, a, b\}$.

Calculamos el conjunto de reglas representativas

$$RR_{\tau,\gamma} = \{X \to Z \setminus X : Z \in RI_{\tau,\gamma}, X \subsetneq Z, \text{mxs}_\tau(Z) < \gamma \cdot s(X) \leq s(Z) < \gamma \cdot \text{mns}_\tau(X)\}.$$

Para ello calculamos primero mns_τ para cada generador minimal frecuente.

$$\text{mns}_\tau(X) = \min(\{s(Y) : Y \in FG_\tau, Y \subsetneq X\} \cup \{\infty\})$$

No hay ningún conjunto contenido estrictamente en el vacío con lo que $\text{mns}_\tau(\emptyset) = \infty$. El vacío es el único conjunto contenido estrictamente en a, b, c, con lo que

$$\text{mns}_\tau(a) = \text{mns}_\tau(b) = \text{mns}_\tau(c) = \min(\{s(\emptyset\} \cup \{\infty\}) = \min(\{13, \infty\}) = 13.$$

Para los generadores minimales frecuentes de dos elementos

$$\begin{aligned}
\text{mns}_\tau(ab) &= \min(\{s(\emptyset), s(a), s(b)\} \cup \{\infty\}) = \min(\{13, 10, 11, \infty\}) = 10 \\
\text{mns}_\tau(ac) &= \min(\{s(\emptyset), s(a), s(c)\} \cup \{\infty\}) = \min(\{13, 10, 6, \infty\}) = 6 \\
\text{mns}_\tau(bc) &= \min(\{s(\emptyset), s(b), s(c)\} \cup \{\infty\} = \min(\{13, 11, 6, \infty\}) = 6
\end{aligned}$$

Por lo tanto, tenemos

$Z \in RI_{\tau,\gamma}$	$X \subsetneq Z$	$\text{mxs}_\tau(Z)$	$\gamma \cdot s(X)$	$s(Z)$	$\gamma \cdot \text{mns}_\tau(X)$	$RR_{\tau,\gamma}$
a	$\emptyset$	5	$0,75 \cdot 13 = 9,75$	11	$0,75 \cdot \infty = \infty$	$\emptyset \to a$
b	$\emptyset$	9	$0,75 \cdot 13 = 9,75$	10	$0,75 \cdot \infty = \infty$	$\emptyset \to b$
ab	$\emptyset$	4	$0,75 \cdot 13 = 9,75$	9	$0,75 \cdot \infty = \infty$	
	a		$0,75 \cdot 10 = 7,5$		$0,75 \cdot 13 = 9,75$	$a \to b$
	b		$0,75 \cdot 11 = 8,25$		$0,75 \cdot 13 = 9,75$	$b \to a$
bc	$\emptyset$	4	$0,75 \cdot 13 = 9,75$	5	$0,75 \cdot \infty = \infty$	
	b		$0,75 \cdot 11 = 8,25$		$0,75 \cdot 13 = 9,75$	
	c		$0,75 \cdot 6 = 4,5$		$0,75 \cdot 13 = 9,75$	$c \to b$
abc	$\emptyset$	0	$0,75 \cdot 13 = 9,75$	4	$0,75 \cdot \infty = \infty$	
	a		$0,75 \cdot 10 = 7,5$		$0,75 \cdot 13 = 9,75$	
	b		$0,75 \cdot 11 = 8,25$		$0,75 \cdot 13 = 9,75$	
	c		$0,75 \cdot 6 = 4,5$		$0,75 \cdot 13 = 9,75$	
	ab		$0,75 \cdot 9 = 6,75$		$0,75 \cdot 10 = 7,5$	
	ac		$0,75 \cdot 4 = 3$		$0,75 \cdot 6 = 4,5$	$ac \to b$
	bc		$0,75 \cdot 5 = 3,75$		$0,75 \cdot 6 = 4,5$	$bc \to a$

Las reglas de asociación representativas son

$$RR_{\tau,\gamma} = \{\emptyset \to a, \emptyset \to b, a \to b, b \to a, c \to b, ac \to b, bc \to a\}.$$