Minería de patrones frecuentes y reglas de asociación

Máster Online en Ciencia de Datos

UCO ONLINE

Dr. José Raúl Romero

Profesor Titular de la Universidad de Córdoba y Doctor en Ingeniería Informática por la Universidad de Málaga. Sus líneas actuales de trabajo se centran en la democratización de la ciencia de datos (*Automated ML* y *Explainable Artificial Intelligence*), aprendizaje automático evolutivo y analítica de software (aplicación de aprendizaje y optimización a la mejora del proceso de desarrollo de software).

Miembro del Consejo de Administración de la *European Association for Data Science*, e investigador senior del Instituto de Investigación Andaluz de *Data Science and Computational Intelligence*.

Director del **Máster Online en Ciencia de Datos** de la Universidad de Córdoba.

Generación de reglas de asociación

UCO ONLINE

Recordemos…

- Habitualmente, el proceso de **minería de reglas de asociación** se divide en dos subproblemas (Han and Kamber, 2001):
  - **Encontrar los itemsets frecuentes** cuyas ocurrencias exceden un umbral de soporte mínimo predefinido
  - **Derivar reglas de asociación** a partir de aquellos itemsets frecuentes (considerando un umbral para la confianza, si procede)
- Estos subproblemas se resuelven iterativamente hasta que no aparezca más nuevas reglas (**condición de cierre**)

Reglas de asociación

Una implicación de la forma **X** $\rightarrow$ **Y**, donde **X** e **Y** son *itemsets*.

Ejemplo:
{Leche, Pañales} $\rightarrow$ {Cerveza}

TID	*Items*
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Coca-cola
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Coca-cola

Medidas de evaluación de una regla:

Soporte (**s**)

Fracción de transacciones que contienen los *ítemsets* **X** e **Y**

Confianza (**c**)

Mide con qué frecuencia los ítems de **Y** aparecen en las transacciones que contienen **X**

Ejemplo: $\{\text{Leche}, Pañales\} \Rightarrow Cerveza$

$$s = \frac{\sigma(\text{Leche}, Pañales, \text{Cerveza})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Leche}, \text{Pañales}, \text{Cerveza})}{\sigma(\text{Leche}, \text{Pañales})} = \frac{2}{3} = 0.67$$

Generación de reglas de asociación

Dado un *itemset* frecuente **L**, encontrar todos los subconjuntos no vacios **f** $\subset$ **L** tal que **f** $\rightarrow$ **L – f** satisfaciendo el **umbral de confianza**

- Si {A,B,C,D} es un *itemset* frecuente, las reglas candidatas son:

  ABC $\rightarrow$D, ABD $\rightarrow$C, ACD $\rightarrow$B, BCD $\rightarrow$A,
  A $\rightarrow$BCD, B $\rightarrow$ACD, C $\rightarrow$ABD, D $\rightarrow$ABC,
  AB $\rightarrow$CD, AC $\rightarrow$ BD, AD $\rightarrow$ BC, BC $\rightarrow$AD,
  BD $\rightarrow$AC, CD $\rightarrow$AB,

Si **|L| = k**, hay **2k – 2** reglas de asociación candidatas (ignorando **L** $\rightarrow \varnothing$ y $\varnothing \rightarrow$ **L**)

¿Cómo generar eficientemente reglas a partir de los *itemsets* frecuentes?

- En general, la confianza no tiene la propiedad anti-monótona

  c(ABC $\rightarrow$D) puede ser mayor o menor que c(AB $\rightarrow$D)

- Pero la confianza de las reglas generadas a partir del mismo *itemset* **sí tienen la propiedad anti-monótona**

  e.g., L = {A,B,C,D}:

  $$c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$$

- La confianza es anti-monótona con respecto al número de ítems del consecuente de la regla

Generación de reglas de asociación

Algoritmo de generación de reglas

UCO ONLINE

Generación de reglas

Función de generación de reglas de asociación

AssocRules()

forall large datasets l_k, $k > 2$

$GenRules(l_k, l_k)$;

Generar todas las reglas válidas **$a \rightarrow (l_k - a)$**, para todas $a \subset a_m$

GenRules(l_k, a_m)

$A = \{(m-1) - itemsets \;\; a_{m-1} \mid a_{m-1} \subset a_m\}$

forall $a_{m-1} \in A$ **begin**

$$conf = \frac{\text{support}(l_k)}{\text{support}(a_{m-1})}$$

Con confianza *conf*

Con support = $support(l_k)$

if $(conf \geq \text{minconf})$ **then**

$\text{output} - \text{rule}\, a_{m-1} \rightarrow (l_k - a_{m-1})$

Generar reglas con subconjuntos de a_{m-1} como antecedente

if $(m - 1 > 1)$ **then**

GenRules(l_k, a_{m-1})

Ejemplo (recuperamos: lección Apriori)

TID	Lista de Items
T100	I1, I2, I5
T101	I2, I4
T102	I2, I3
T103	I1, I2, I4
T104	I1, I3
T105	I2, I3
T106	I1, I3
T107	I1, I2 ,I3, I5
T108	I1, I2, I3

- La base de datos **D** contiene 9 transacciones
- Se supone que el soporte mínimo requerido es 2 (min_sup = 2/9 = 22%)
- Se establece un umbral mínimo de confianza del 70%
- Primero, debemos averiguar los itemsets frecuentes utilizando el algoritmo correspondiente
- Después, se generan las reglas de asociación según los umbrales mínimos de soporte y confianza

Ejemplo

- **Procedimiento**:
  - Para cada itemset frecuente **l**, se generan todos los subconjuntos de **l** no vacíos
  - Para cada subconjunto no vacío **a** de **l**, se considera la regla **a -> (l – a)** si su confianza supera el umbral mínimo:

$$support\_count(l) / support\_count(s) >= min\_conf$$

- Siguiendo con el ejemplo…

L = {{I1}, {I2}, {I3}, {I4}, {I5}, {I1,I2}, {I1,I3}, {I1,I5}, {I2,I3}, {I2,I4}, {I2,I5}, {I1,I2,I3}, {I1,I2,I5}}.

  - Tomamos ***l*** = {I1,I2,I5} y **min_conf** = 70%
  - Todos sus subconjuntos no vacíos son {I1,I2}, {I1,I5}, {I2,I5}, {I1}, {I2}, {I5}

Ejemplo

A cada regla resultante se le calcula su confianza:

- R1: I1 ^ I2 → I5
  - Confianza = supp{I1,I2,I5} / supp{I1,I2} = 2 / 4 = 50%
  - R1 es rechazada
- R2: I1 ^ I5 → I2
  - Confianza = supp{I1,I2,I5} / supp{I1,I5} = 2 / 2 = 100%
  - R2 es seleccionada
- R3: I2 ^ I5 → I1
  - Confianza = supp{I1,I2,I5} / supp{I2,I5} = 2/2 = 100%
  - R3 es seleccionada

Ejemplo

- R4: I1 → I2 ^ I5
  - Confianza = supp{I1,I2,I5} / supp{I1} = 2 / 6 = 33%
  - R4 es rechazada
- R5: I2 → I1 ^ I5
  - Confianza = supp{I1,I2,I5} / {I2} = 2 / 7 = 29%
  - R5 es rechazada
- R6: I5 → I1 ^ I2
  - Confidence = sc{I1,I2,I5}/ {I5} = 2/2 = 100%
  - R6 es seleccionada

R2, R3 y R6 son reglas de asociación exactas

Ejemplo 2

L = {{1}, {2}, {3}, {5}, {1,3}, {2,3}, {2,5}, {3,5}, {2,3,5}}

Fila	1	2	3	4	5
1	x		x	x	
2		x	x		x
3	x	x	x		x
4		x			x

min_supp = 2
min_conf = 0.75

Reglas de asociación:

{1} → {3} : 1
{2} → {3} : 0.67
{2} →{5} : 1
{3} → {5} : 0.67
{2,3} → {5} : 1
{3,5} → {2} : 1

{3} → {1} : 0.67
{3} → {2} : 0.67
{5} → {2} : 1
{5} → {3} : 0.67
{2,5} → {3} : 0.67

Generación de reglas de asociación

Uso de la herramienta Weka

UCO ONLINE

¡Gracias!

UCO ONLINE