Minería de patrones de asociación

Christopher Expósito Izquierdo
Airam Expósito Márquez
Israel López Plata
Belén Melián Batista
J. Marcos Moreno Vega

{**cexposit, aexposim, ilopezpl, mbmelian, jmmoreno**}**@ull.edu.es**

Departamento de Ingeniería Informática y de Sistemas
Universidad de La Laguna

Índice

INTRODUCCIÓN

Patrones de asociación

Propósito

- Identificar asociaciones de interés en un conjunto de datos.
- La versión más estudiada es la identificación de patrones frecuentes.

Origen

- *Mining Association Rules Between Sets of Items in Large Databases*

  Rakesh Agrawal, Tomasz Imieliński, Arun Swami [2]

Aplicaciones

- Análisis de la cesta de la compra.

  Encontrar grupos de alimentos frecuentes en las compras de los clientes.

- Análisis de *web logs*.

  Identificar subconjuntos de logs frecuentes o poco frecuentes.

Ejemplo

COMPRAS	
ID	**Productos**
1	fondant, azúcar, colorante
2	fondant, revistas, libros
3	azúcar, revistas, libros
4	fondant, azúcar, revistas, libros
5	fondant, azúcar, colorantes, libros

Definición del problema

Transacciones (T)	Universo (U)					
	u_1	u_2	$\cdots$	u_j	$\cdots$	u_m
T_1	τ_{11}	τ_{12}	$\cdots$	τ_{1j}	$\cdots$	τ_{1m}
T_2	τ_{21}	τ_{22}	$\cdots$	τ_{2j}	$\cdots$	τ_{2m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T_i	τ_{i1}	τ_{i2}	$\cdots$	τ_{ij}	$\cdots$	τ_{im}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T_n	τ_{n1}	τ_{n2}	$\cdots$	τ_{nj}	$\cdots$	τ_{nm}

$$\tau_{ij} = \begin{cases} 1 & \text{si el ítem } j \text{ está presente en la transacción } T_i \\ 0 & \text{en otro caso} \end{cases}$$

Patrones frecuentes

Ejemplo

- **Productos de la compra**

u_1 = fondant, u_2 = azúcar, u_3 = colorante, u_4 = revistas, u_5 = libros

- **Transacciones**

$T_1 = \{u_1, u_2, u_3\}$,
$T_2 = \{u_1, u_4, u_5\}$,
$T_3 = \{u_2, u_4, u_5\}$,
$T_4 = \{u_1, u_2, u_4, u_5\}$,
$T_5 = \{u_1, u_2, u_3, u_5\}$

- **Matriz del problema**

	u_1	u_2	u_3	u_4	u_5
T_1	1	1	1	0	0
T_2	1	0	0	1	1
T_3	0	1	0	1	1
T_4	1	1	0	1	1
T_5	1	1	1	0	1

Patrones frecuentes

Algunas definiciones

- Un **itemset** es un conjunto de uno o más ítems.
- Un ***k*-itemset** es un itemset de k elementos.
- El **soporte** o cobertura de un itemset I, $sup(I)$, se define como la fracción de transacciones que contienen a I como subconjunto.
- Un **itemset frecuente** es aquel cuyo soporte es igual o superior a un umbral establecido previamente.
- Un **itemset frecuente** I es **maximal** a un cierto nivel de soporte mínimo *minsup* si I es frecuente y ningún superconjunto de I es frecuente.

Patrones frecuentes

Ejemplo

- **Matriz del problema**

	u_1	u_2	u_3	u_4	u_5
T_1	1	1	1	0	0
T_2	1	0	0	1	1
T_3	0	1	0	1	1
T_4	1	1	0	1	1
T_5	1	1	1	0	1

- **Algunos soportes**

$$I = \{u_1, u_2\} \Rightarrow sup(I) = 3/5$$
$$I = \{u_2, u_4\} \Rightarrow sup(I) = 2/5$$
$$I = \{u_2, u_3, u_4\} \Rightarrow sup(I) = 0$$

Propiedades

- **Propiedad de antimonotonía del soporte**

El soporte de cualquier subconjunto J de I es mayor o igual que el soporte de I. Es decir:

$$\forall I, J : J \subseteq I \Rightarrow supp(J) \geq supp(I)$$

- **Propiedad de clausura descendente del soporte**

Todo subconjunto J de un conjunto frecuente I es también frecuente.

- De lo anterior se sigue que si un itemset no es frecuente, tampoco lo serán los itemset que lo contengan.

REGLAS DE ASOCIACIÓN

Reglas de asociación

Definición

- Una **regla de asociación** es una correspondencia entre itemsets. Si X e Y son dos itemsets, la regla que asocia a X con Y se escribe:

$$X \rightarrow Y$$

X es el antecedente de la regla e Y el consecuente.

- Algunos ejemplos:

$$\begin{aligned} \{\text{revistas}\} &\rightarrow \{\text{libros}\} \\ \{\text{libros}\} &\rightarrow \{\text{revistas}\} \\ \{\text{azúcar, fondant}\} &\rightarrow \{\text{colorante}\} \\ \{\text{fondant}\} &\rightarrow \{\text{azúcar, colorante}\} \end{aligned}$$

Medidas de evaluación

• El **soporte de una regla de asociación** se define como la proporción de transacciones que contienen al antecedente y al consecuente de la regla. Es decir:

$$supp(X \rightarrow Y) = supp(X \cup Y)$$

• La **confianza de una regla de asociación** se define como la proporción de transacciones que se cumplen cuando se puede aplicar la regla. Es decir:

$$conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

Ejemplo

- **Productos de la compra**

u_1 = fondant, u_2 = azúcar, u_3 = colorante, u_4 = revistas, u_5 = libros

- **Matriz del problema**

	u_1	u_2	u_3	u_4	u_5
T_1	1	1	1	0	0
T_2	1	0	0	1	1
T_3	0	1	0	1	1
T_4	1	1	0	1	1
T_5	1	1	1	0	1

- **Soporte y confianza**

$X = \{\text{libros}\}, Y = \{\text{revistas}\}$

$$\begin{aligned} supp(X \to Y) &= supp(X \cup Y) \\ &= 3/5 = 0.6 \end{aligned}$$

$$\begin{aligned} conf(X \to Y) &= \frac{supp(X \to Y)}{supp(X)} \\ &= \frac{3/5}{4/5} = 0.75 \end{aligned}$$

Observaciones

- La confianza de la regla $X \rightarrow Y$ es una estimación de la probabilidad condicional de Y dado X obtenida de las frecuencias relativas de aparición de X e Y.

- Los resultados de un análisis de asociación en un aplicación concreta deben tomarse con cautela. Del hecho de que una regla $X \rightarrow Y$ tenga una alta confianza no se deriva directamente que X sea la causa de Y. La regla muestra la ocurrencia simultánea de X e Y, pero no que X sea la causa e Y la consecuencia.

Propiedades

- **Propiedad de monotonía de la confianza**

Sean I, X_1, X_2 itemsets tales que $X_1 \subset X_2 \subset I$. Se tiene que:

$$conf(X_1 \rightarrow I - X_1) \leq conf(X_2 \rightarrow I - X_2)$$

Descripción del problema

- Dados un universo U, un conjunto de transacciones T y dos valores $minsupp \in [0, 1]$, $minconf \in [0, 1]$ encontrar todas las reglas de asociación $X \to Y$ que cumplan

1. *Soporte mínimo*

$$supp(X \to Y) \geq minsupp$$

2. *Confianza mínima*

$$conf(X \to Y) \geq minconf$$

ALGORITMOS PARA PATRONES FRECUENTES

Búsqueda exhaustiva

- **Generar** todas las reglas de asociación que pueden obtenerse desde el universo U.
- **Calcular** el soporte y la confianza de cada regla teniendo en cuenta el conjunto de transacciones T.
- **Almacenar** las reglas que superen los umbrales mínimos *minsupp* y *minconf* .

Búsqueda exhaustiva. Ejemplo

- **Reglas derivadas del conjunto de ítems {fondant, revistas, libros}.**

Regla de asociación	Soporte	Confianza
{fondant} → {revistas, libros}	$2/5 = 0.40$	$2/4 = 0.50$
{revistas} → {fondant, libros}	$2/5 = 0.40$	$2/3 = 0.66$
{libros} → {fondant, revistas}	$2/5 = 0.40$	$2/4 = 0.50$
{fondant, revistas} → {libros}	$2/5 = 0.40$	$2/2 = 1.00$
{fondant, libros} → {revistas}	$2/5 = 0.40$	$2/3 = 0.66$
{revistas, libros} → {fondant}	$2/5 = 0.40$	$2/3 = 0.66$

Búsqueda exhaustiva. Observaciones

- Con un conjunto de ítems de tamaño tres se obtienen seis reglas.
- Si el conjunto tiene k elementos, el número de reglas que pueden obtenerse es $2^k - 2$.
- La magnitud de la tarea a la que se enfrenta la búsqueda exhaustiva se refleja en los datos de la siguiente tabla:

k	$2^k - 2$
2	2
4	14
8	126
15	32766
20	1048574

Búsqueda exhaustiva. Otras observaciones

- Todas las reglas anteriores tienen el mismo soporte, ya que han sido construidas desde el mismo conjunto de ítems.
- Sin embargo, la confianza varía, ya que depende de la distribución de los ítems en el antecedente y consecuente de las reglas.
- Por tanto, la búsqueda de reglas que cumplan los requerimientos mínimos de soporte y confianza puede desarrollar en dos etapas.

Algoritmos para patrones frecuentes

Búsqueda exhaustiva mejorada.

- **Generar los conjuntos frecuentes.**

Encontrar los conjuntos de ítems cuyos soporte sea mayor o igual que *minsupp*.

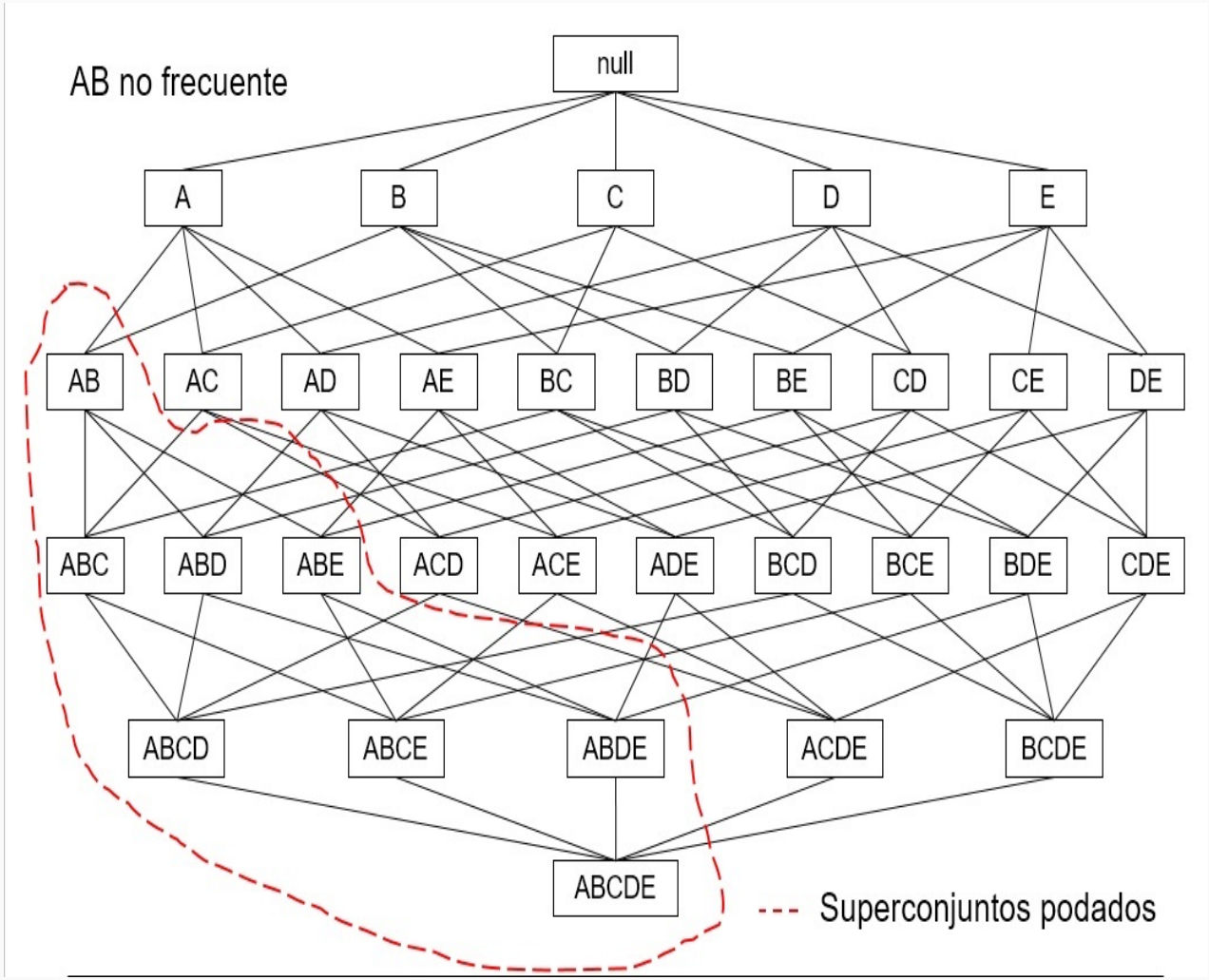
- **Construir las reglas de asociación.**

Para cada uno de los conjuntos obtenidos en la etapa anterior, encontrar las reglas de asociación que tienen una confianza mayor o igual que *minconf* .

Generación eficiente de los conjuntos frecuentes.

- Para generar los conjuntos frecuentes es conveniente usar las propiedades de antimonotonía y clausura descendente del soporte.
- De ellas se deduce que si un conjunto de ítems no es frecuente, tampoco lo serán los conjuntos que lo contengan.
- De esta manera se poda el árbol de búsqueda y se incrementa la eficiencia de la generación de conjuntos frecuentes.

Generación eficiente de los conjuntos frecuentes.



Algoritmos Apriori

datos: T, transacciones

entrada: *minsupp*, umbral de soporte mínimo

variable: C_k, k-ítems candidatos a ser frecuentes

variable: F_k, k-ítems frecuentes

salida: $\bigcup_{i=1}^{k} F_i$, conjuntos frecuentes de ítems

propósito: Encontrar eficientemente los conjuntos frecuentes de ítems

propuesta: Aplicar inteligentemente la propiedad de clausura descendente del soporte.

Algoritmos para patrones frecuentes

Algorithm 1: Apriori(transacciones: T, soporte mínimo: *minsupp*)

$k = 1$;

$F_1 = \{$conjuntos de 1-itemsets frecuentes$\}$;

repeat

  Generar C_{k+1} uniendo pares de elementos de F_k;

  Podar los itemsets de C_{k+1} que violen la clausura descendente;

  Determinar F_{k+1};

  $k = k + 1$;

until $(F_k = \emptyset)$;

return $(\bigcup_{i=1}^{k} F_i)$;

Algoritmos Apriori. Ejemplo

- **Transacciones**

	u_1	u_2	u_3	u_4	u_5
T_1	1	1	1	0	0
T_2	1	0	0	1	1
T_3	0	1	0	1	1
T_4	1	1	0	1	1
T_5	1	1	1	0	1

- **Soporte mínimo:** $minsupp = 3/5$

Algoritmos Apriori. Ejemplo

- **Iteración 1.-**
  - **Conjuntos de tamaño 1 candidatos a ser frecuentes.**

C_1				
$\{u_1\}$	$\{u_2\}$	$\{u_3\}$	$\{u_4\}$	$\{u_5\}$
4/5	4/5	2/5	3/5	4/5

  - **Conjuntos frecuentes de tamaño 1.**

F_1			
$\{u_1\}$	$\{u_2\}$	$\{u_4\}$	$\{u_5\}$
4/5	4/5	3/5	4/5

Algoritmos Apriori. Ejemplo

- **Iteración 2.-**
  - **Conjuntos de tamaño 2 candidatos a ser frecuentes.**

C_2					
$\{u_1, u_2\}$	$\{u_1, u_4\}$	$\{u_1, u_5\}$	$\{u_2, u_4\}$	$\{u_2, u_5\}$	$\{u_4, u_5\}$
$3/5$	$2/5$	$3/5$	$2/5$	$3/5$	$3/5$

  - **Conjuntos frecuentes de tamaño 2.**

F_2			
$\{u_1, u_2\}$	$\{u_1, u_5\}$	$\{u_2, u_5\}$	$\{u_4, u_5\}$
$3/5$	$3/5$	$3/5$	$3/5$

Algoritmos Apriori. Ejemplo

- **Iteración 3.-**
  - **Conjuntos de tamaño 3 candidatos a ser frecuentes.**

C_3	
$\{u_1, u_2, u_5\}$	$\{u_1, u_4, u_5\}$
2/5	2/5

  - **Conjuntos frecuentes de tamaño 3.**

No hay.

Algoritmos Apriori. Ejemplo

- **Conjuntos frecuentes de ítems.-**

Conjunto	Soporte
$\{u_1\}$	$4/5$
$\{u_2\}$	$4/5$
$\{u_4\}$	$3/5$
$\{u_5\}$	$4/5$
$\{u_1, u_2\}$	$3/5$
$\{u_1, u_5\}$	$3/5$
$\{u_2, u_5\}$	$3/5$
$\{u_4, u_5\}$	$3/5$

Generación de reglas

- **Definición del problema.** Dado un conjunto frecuente de ítems, encontrar todas las reglas que pueden derivarse de él cuya confianza supere el umbral mínimo *minconf*.

Generación de reglas

- Dado un conjunto frecuente I de ítems, las reglas que pueden derivarse de él tienen la forma $X \to I - X$, con $X \subset I$.
- Si $I = \{A, B, C, D\}$, se obtienen las siguientes reglas:

$$\begin{aligned}
R_1 &: \{A\} \to \{B, C, D\}; & R_2 &: \{B\} \to \{A, C, D\};\\
R_3 &: \{C\} \to \{A, B, D\}; & R_4 &: \{D\} \to \{A, B, C\};\\
R_5 &: \{A, B\} \to \{C, D\}; & R_6 &: \{A, C\} \to \{B, D\};\\
R_7 &: \{A, D\} \to \{B, C\}; & R_8 &: \{B, C\} \to \{A, D\};\\
R_9 &: \{B, D\} \to \{A, C\}; & R_{10} &: \{C, D\} \to \{A, B\};\\
R_{11} &: \{A, B, C\} \to \{D\}; & R_{12} &: \{A, B, D\} \to \{C\};\\
R_{13} &: \{A, C, D\} \to \{B\}; & R_{14} &: \{B, C, D\} \to \{A\};
\end{aligned}$$

Generación eficiente de las reglas

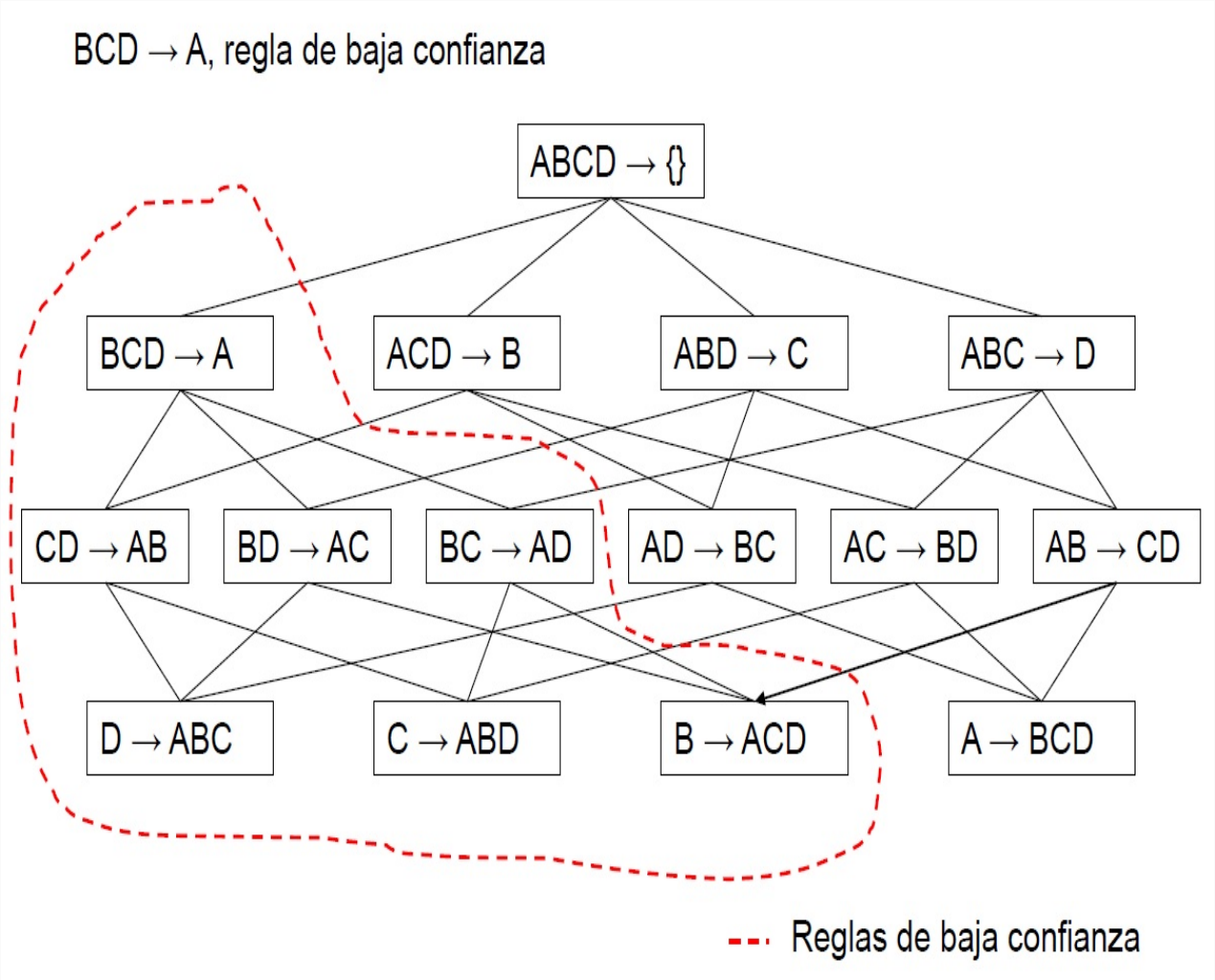
• Para generar las reglas es conveniente usar la propiedad de monotonía de la confianza.

Sean dos reglas $R_1 : X_1 \rightarrow I - X_1$, $R_2 : X_2 \rightarrow I - X_2$, con $X_1 \subset X_2 \subset I$. Se cumple que:

$$\text{si } conf(R_2) \leq minconf \Rightarrow conf(R_1) \leq minconf$$

• De esta manera se poda el árbol de búsqueda y se incrementa la eficiencia de la generación de las reglas.

Generación eficiente de las reglas



OTRAS MEDIDAS DE EVALUACIÓN

Otras medidas de evaluación

Número y relevancia de las reglas de asociación

- El número de reglas de asociación que cumplen los requerimientos mínimos de soporte y confianza puede fácilmente ser del orden de miles o millones en aplicaciones reales.
- Muchas de ellas no tienen interés práctico al no revelar ningún patrón novedoso.
- Por otro lado, las conclusiones derivadas de la confianza de una regla pueden ser débiles.

Otras medidas de evaluación

Tipos de medidas

- **Medidas objetivas:** basadas en valores obtenidos desde los datos. Son independientes de la aplicación y del usuario.
- **Medidas subjetivas:** dependientes de la aplicación y del usuario.
  - **Visualización:** herramientas para la interacción del usuario con el sistema de extracción de reglas que le permitan interpretar y descartar los patrones que no son de su interés.
  - **Procedimiento basado en plantilla:** devuelve únicamente las reglas de asociación que cumplen la plantilla especificada por el usuario.
  - **Medidas subjetivas de interés:** indicadas por el usuario y dependientes de la aplicación.

Otras medidas de evaluación

Medidas objetivas. Tabla de contingencia

	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1+}
$\overline{A}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

- **A, B:** variables.
- $\boldsymbol{f_{11}}$, número de veces que aparecen conjuntamente A y B.
- $\boldsymbol{f_{10}}$, número de veces que aparece A, pero no B.
- $\boldsymbol{f_{01}}$, número de veces que no aparece A, pero sí B.
- $\boldsymbol{f_{00}}$, número de veces que no aparecen ni A ni B.
- $\boldsymbol{N}$, número total de observaciones.

Tabla de contingencia. Ejemplo (i)

	Café	No café	
Té	150	50	200
No té	650	150	800
	800	200	1000

- **Regla:** R_1 : Té $\rightarrow$ Café
- **Soporte:** $supp(R_1) = 0.15$.
- **Confianza:** $conf\,(R_1) = 0.75$.
- **Conclusión:** quiénes beben té tienden a beber café.

Tabla de contingencia. Ejemplo (ii)

	Café	No café	
Té	150	50	200
No té	650	150	800
	800	200	1000

- Es una **conclusión errónea** porque:
  - los bebedores de café, independientemente de si toman o no té, son el 80% de los consumidores,
  - los bebedores de café son el 75% de los bebedores de té.
- Es decir, conocer que una persona bebe té hace que disminuya la probabilidad de que beba café.

Otras medidas de evaluación

Medidas objetivas de interés

- **Lift**

$$Lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{supp(Y)}$$

- Se tiene que:

$$Lift(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) \cdot supp(Y)} = \frac{N \cdot f_{11}}{f_{1+} \cdot f_{+1}}$$

Otras medidas de evaluación

Medidas objetivas de interés

- **Interpretación del Lift**

$Lift(X \rightarrow Y) = 1$, si X e Y son independientes

$Lift(X \rightarrow Y) < 1$, si X e Y están negativamente correlados

$Lift(X \rightarrow Y) > 1$, si X e Y están positivamente correlados

Otras medidas de evaluación

Medidas objetivas de interés

- **Límites del Lift**

	p	$\overline{p}$	
q	880	50	930
$\overline{q}$	50	20	70
	930	70	1000
$Lift(p \rightarrow q) = 1.02$			

	r	$\overline{r}$	
s	20	50	70
$\overline{s}$	50	880	930
	70	930	1000
$Lift(r \rightarrow s) = 4.08$			

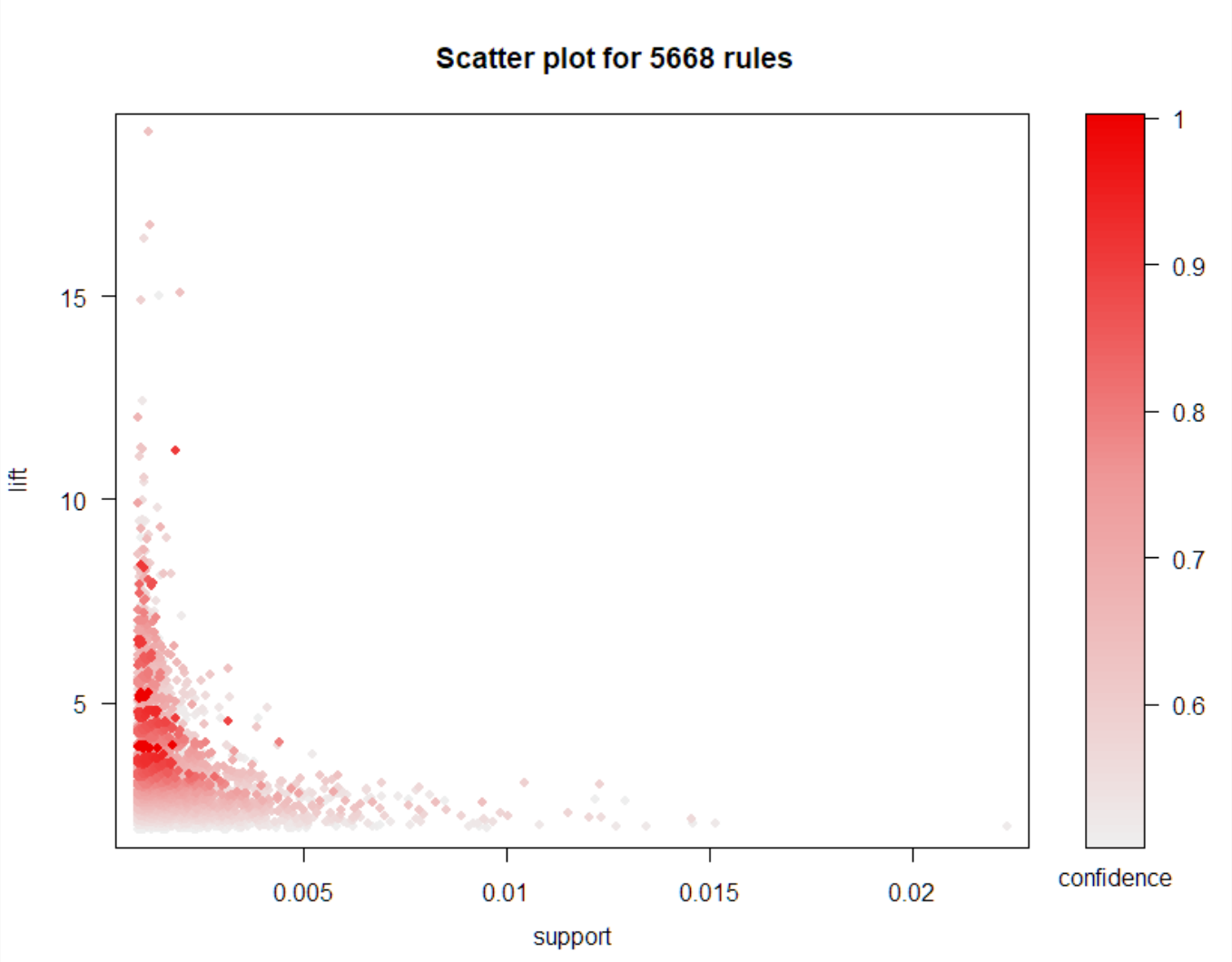
- Aunque p y q aparecen juntos el 88% de las veces, su *lift* es 1.02 (estadísticamente independientes). Sin embargo, aunque r y s aparecen juntos solo el 20% de las veces, su *lift* es 4.08.

- Nótese, no obstante, que la confianza de $p \rightarrow q$ es igual a 94.6% y la confianza de $r \rightarrow s$ es 28.6%.

Lift frente al soporte y la confianza

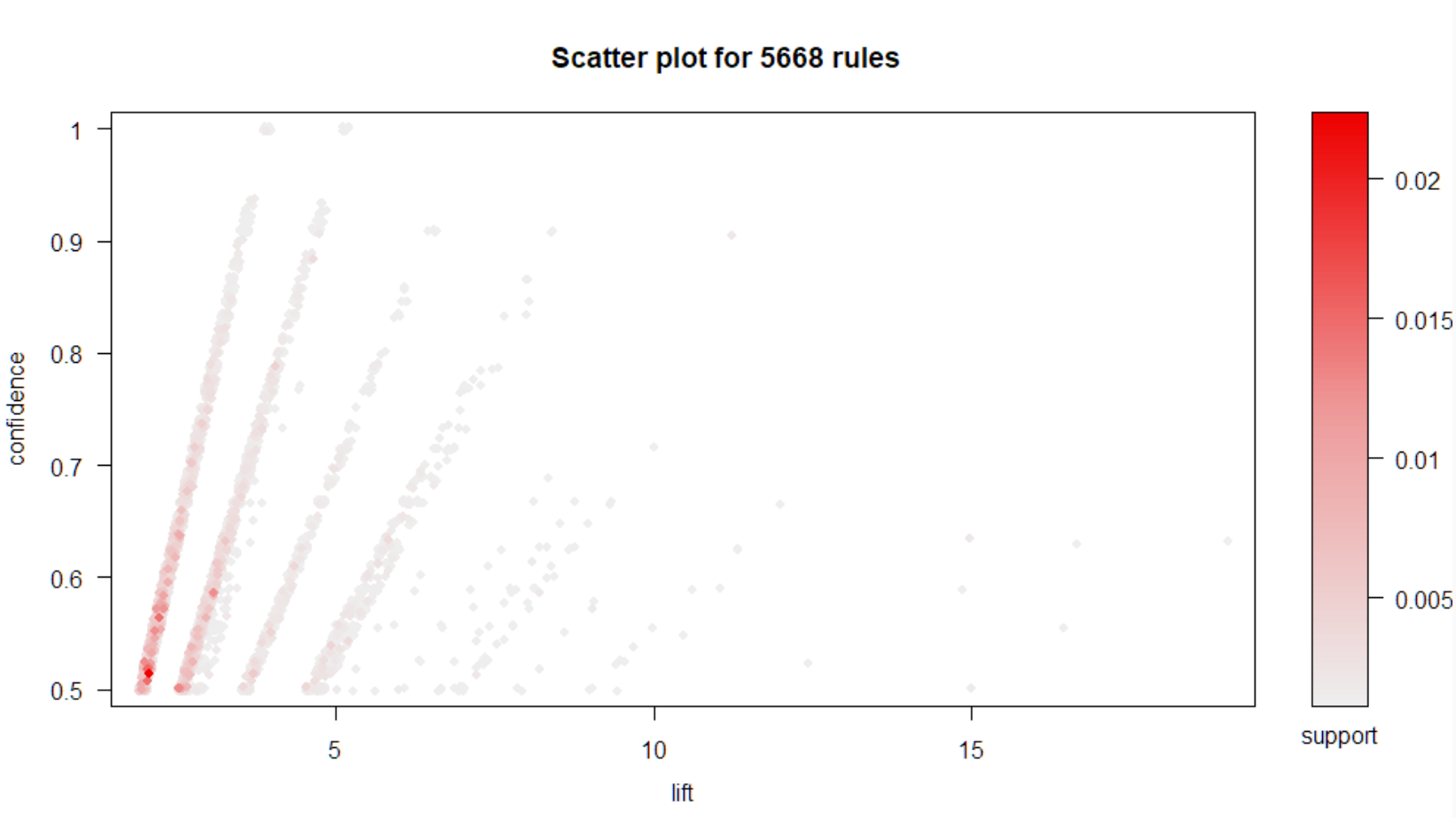
Otras medidas de evaluación

Confianza frente al soporte y lift



Otras medidas de evaluación

Soporte frente a lift y confidence



BIBLIOGRAFÍA

Bibliografía

Aggarwal, C. C.
Data Mining. The Textbook.
Springer, 2015.

Agrawal, R., Imieliński, T., and Swami, A.
Mining association rules between sets of items in large databases.
In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 1993), SIGMOD '93, ACM, pp. 207–216.

Tan, P.-N., Steinbach, M., and Kumar, V.
Introduction to Data Mining.
Addison-Wesley, 2006.

Licencia