Problemas de relación entre atributos: asociación

Maria-Amparo Vila
vila@decsai.ugr.es

Grupo de Investigación en Bases de Datos y
Sistemas de Información Inteligentes
https://idbis.ugr.es/
Departamento de Ciencias de la Computación
e Inteligencia Artificial
Universidad de Granada

Octubre 2014

Reglas de asociación: ideas básicas

Problemas más importantes de DM

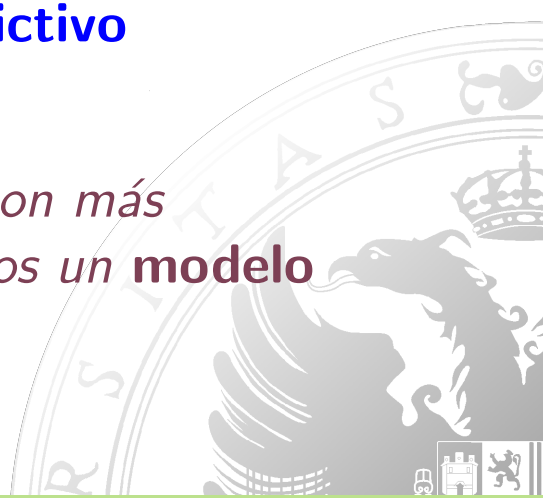
Modelización de dependencias

• *Objetivo:* Describir dependencias significativas entre las variables incluidas en la base de datos Los modelos de dependencias pueden ser:

- Cualitativas o cuantitativas (dependencias funcionales y análisis de regresión)
- Dependencias parciales o completas

Cuando se trata de variables cuantitativas, y se espera la existencia de una relación $y = f(x_1, ..x_n)$ tenemos un **modelo predictivo** *normalmente de análisis de regresión*

Cuando no se tiene conocimiento previo, las variables son más generales y se buscan asociaciones entre valores tenemos un **modelo descriptivo**



Reglas de asociación: ideas básicas

Planteamiento del problema

• *Objetivo*: Descubrir "reglas" en lógica proposicional (pero cualificadas probabilísticamente) que involucren algunos valores de ciertos atributos.

· Es un **modelo descriptivo**

· Los datos serán, en principio nominales o intervalares.

· El formato de los datos es más general que una base de datos relacional.

Se trabaja en una base de datos de *transacciones*

Reglas de asociación: ideas básicas

Planteamiento del problema

Ejemplo: "cesta de la compra"

Cliente1:	lechePascualE , azúcar1Kg , pepinos , cinta_8mm
Cliente2:	lechePascualE , azúcar1Kg , ternera

Se buscarán asociaciones del tipo:
El 90% de los clientes que compran azúcar, compran leche
El 80% de los clientes que compran algún disco de JL Guerra, compran un disco de Gloria Stefan

Reglas de asociación: ideas básicas

Definición formal de reglas de asociación (Agrawal y otros, 1993)

- I conjunto de items
- T conjunto de transacciones, $T \subseteq \wp(I)$
- **Regla de asociación en T:**

$$A \Rightarrow C \; \forall A, C \subset I \; A \cap C = \emptyset$$

- *Otro ejemplo interesante: Análisis de textos*
  - I: *términos en un documento*
  - T: *conjunto de textos*
  - $\{amor, dolor\} \Rightarrow \{muerte\}$

Asociación significa **coocurrencia**, *no causalidad*

Reglas de asociación: ideas básicas

Bases de datos Relacional y transaccional

Definición de Base de datos Transaccional

· Base de datos donde cada "fila" o registro tiene dos partes:

1. La identificación de registro
2. Un conjunto de items (itemset)

Si de partida tenemos una BDR, habrá que transformarla en una BDT, es inmediato si consideramos que los items posibles son **todos** los posibles valores de los atributos

Reglas de asociación: ideas básicas

Bases de datos Relacional y transaccional

Ejemplo de transformación BD relacional- BD transaccional

BDR:

DNI	Nombre	Altura	Peso	Dirección
5	JC	186	87	Gr
6	P	175	70	Ma

BDT:

5	NombreJC , Altura186 , Peso87 , DireccGr
6	NombreP , Altura175 ,Peso70 , DireccMa

Reglas de asociación: ideas básicas

Bases de datos Relacional y transaccional

Ejemplo de transformación BD transaccional- BD relacional

BDT:

Cliente1	lechePascualE , azúcar1Kg , pepinos , cinta_8mm
Cliente2	lechePascualE , azúcar1Kg , ternera

↓ Recodificación

BDT:

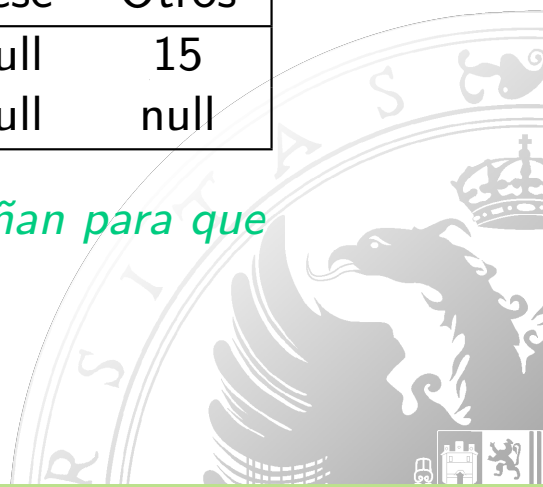
Cliente1:	Lact3 , Ultr5 , Fruta9 , Otros15
Cliente2:	Lact3 , Ultr5 , Carne2

↓ Transformación

BDR:

IDCl	Lact	Ultr	Fruta	Carne	Pesc	Otros
Cliente1:	3	5	9	null	null	15
Cliente2:	3	5	null	2	null	null

los algoritmos que extraen reglas de asociación siempre se diseñan para que trabajen directamente sobre BDT



Reglas de asociación: ideas básicas

Medidas de valoración de una regla de asociación

- **Soporte de un itemset I_0 en T**

$$supp(I_0) = \frac{|\{\tau \in T \mid I_0 \subseteq \tau\}|}{|T|}$$

Número de veces que ocurre en una base de datos/número total de transacciones

- **Soporte de una R.A.:**

$$Supp(A \Rightarrow C) = supp(A \cup C)$$

Soporte del conjunto total de items involucrados en ella

- **Confianza de una R.A.:**

$$Conf(A \Rightarrow C) = \frac{supp(A \cup C)}{supp(A)}$$

Proporción entre la frecuencia común, y la del consecuente

Reglas de asociación: ideas básicas

Medidas de valoración de una regla de asociación

Ejemplo de medidas

A	B	C	D
a0	b1	c1	d1
a1	b1	c1	d2
a2	b1	c1	d3
a3	b1	c1	d4
a4	b2	c3	d5
a5	b2	c3	d6
a6	b1	c1	d4

Soporte: $supp(b1, c1) = 5/7$

$supp(b1, c1, d4) = 2/7$

Soporte de una regla: $supp((b1, c1) \rightarrow d5) = 0$

$supp((b2, c3) \Rightarrow d5) = 1/7$

Confianza: $conf((b1, c1) \Rightarrow d4) = \frac{2/7}{5/7} = 2/5$

$conf((b2, c3) \Rightarrow d5) = \frac{1/7}{1/7} = 1$

Reglas de asociación: ideas básicas

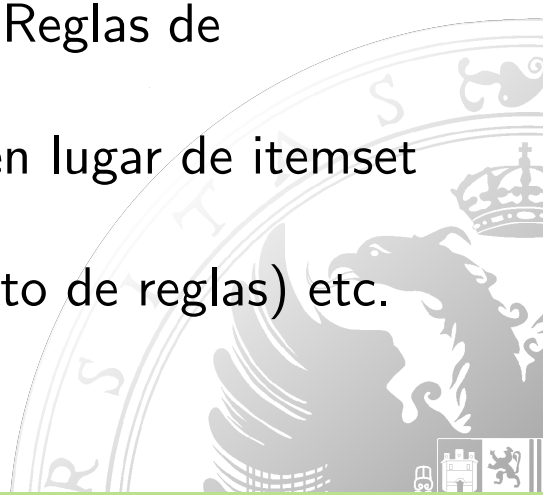
Medidas de valoración de una regla de asociación

- La confianza de una regla mide su *calidad*
- El soporte mide la *cantidad* de tuplas que soportan la inducción.
- Se suelen imponer dos umbrales:
  - **Umbral de Soporte** $minsup$ (5% p.e)
    Todo conjunto de items X tal que $supp(X) \geq minsup$ se denomina *itemset frecuente*
    Toda regla $A \Rightarrow C$ es *frecuente* si $supp(A \Rightarrow C) \geq minsup$
  - **Umbral de Confianza** $minconf$ (70% p.e)

Reglas de asociación: ideas básicas

Problemas asociados a las reglas de asociación

1. *Extracción de reglas con soporte y confianza mayores que los umbrales (minsupp y minconf)*
   - ◇ Algoritmos de Minería de reglas: A priori y variantes
2. *Interpretación de las reglas (puede haber una explosión combinatoria)*
   - ◇ Otras medidas de calidad: factor de certeza etc..
   - ◇ Otros tipos de reglas más complejas que impliquen causalidad
   - ◇ Mecanismos de agrupamiento de items en conceptos más complejos (P.E. no leche pascual, sino leche o producto lácteos) (Reglas de asociación difusas)
   - ◇ Uso de otros tipos de conjuntos para generar reglas en lugar de itemset frecuentes (item sets cerrados)
   - ◇ Mecanismo de Minería de segundo nivel (agrupamiento de reglas) etc.



Reglas de asociación: ideas básicas

Generación de reglas de asociación

Problema

Dada una base de datos transaccional obtener todas las reglas con soporte y confianza mayores que $minsupp$ y $minconf$

Enfoque inicial: fuerza bruta

- Listar todas las posibles reglas
- Calcular los soportes y confianza
- Eliminar los que no verifique los umbrales

El costo es computacionalmente prohibitivo incluso para conjuntos pequeños

Reglas de asociación: ideas básicas

Generación de reglas de asociación

Ejemplo

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Ejemplos de reglas:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Observaciones

- Todas las reglas son particiones binarias del mismo itemset $\{Milk, Diaper, Beer\}$
- Las reglas que se originan a partir del mismo itemset tiene igual soporte pero distinta confianza
- Se pueden separar las tareas de buscar itemset frecuentes y reglas fiables

Algoritmo Apriori

Reglas de asociación: generación

Algoritmo Apriori: principios básicos

Estrategia divide y vencerás:

1. En primer lugar, se calculan aquellas reglas 'frecuentes' (con un soporte mayor que un umbral $minsupp$).
2. En segundo lugar, se ve cuales de esas reglas tienen una confianza mayor que el umbral $minconf$.

El primer paso es, con diferencia, el más costoso. Se basa en:
La regla $A \Rightarrow B$ es *frecuente*, si y solo si par $A \cup B$ es *frecuente*
- El problema de encontrar reglas de asociación se divide en:

1. Encontrar itemsets frecuentes (soporte alto). Se plantean k-itemsets (itemsets con k items)
2. A partir de ellos, obtener reglas con confianza alta

Algoritmo Apriori: encontrar k-Itemsets frecuentes

Fuerza bruta

Ejemplo

100	A_1	B_2	C_0
200	A_2	B_1	C_1
300	A_2	B_2	C_1
400	A_3	B_1	C_2
500	A_4	B_1	C_2
600	A_2	B_1	C_2

$k = 1$

A_1	1
A_2	3
A_3	1
A_4	1
B_1	4
B_2	2
C_0	1
C_1	2
C_2	3

$k = 2$

A_1B_1	0	A_3C_1	0
A_1B_2	1	A_3C_2	1
A_1C_0	1	A_4B_1	1
A_1C_1	0	A_4B_2	0
A_1C_2	1	A_4C_0	0
A_2B_1	2	A_4C_1	0
A_2B_2	1	A_4C_2	1
A_2C_0	0	B_1C_0	0
A_2C_1	2	B_1C_1	1
A_2C_2	1	B_1C_2	3
A_3B_1	1	B_2C_0	1
A_3B_2	0	B_2C_1	1
A_3C_0	0	B_2C_2	0

$k = 3$

$A_1B_1C_0$	0	$A_3B_1C_0$	0
$A_1B_1C_1$	0	$A_3B_1C_1$	0
$A_1B_1C_2$	0	$A_3B_1C_2$	1
$A_1B_2C_0$	1	$A_3B_2C_0$	0
$A_1B_2C_1$	0	$A_3B_2C_1$	0
$A_1B_2C_2$	0	$A_3B_2C_2$	0
$A_2B_1C_0$	0	$A_4B_1C_0$	0
$A_2B_1C_1$	1	$A_4B_1C_1$	0
$A_2B_1C_2$	1	$A_4B_1C_2$	1
$A_2B_2C_0$	0	$A_4B_2C_0$	0
$A_2B_2C_1$	1	$A_4B_2C_1$	0
$A_2B_2C_2$	0	$A_4B_2C_2$	0

Reglas de asociación: generación

Algoritmo Apriori: encontrar k-Itemsets frecuentes

Solución

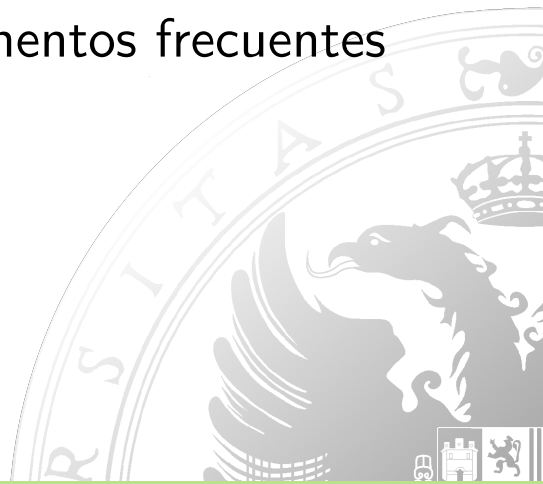
Usar la propiedad "a priori"

Todo subconjunto de un itemset frecuente es frecuente, luego si un itemset **no es frecuente**, *ningún conjunto que lo contenga lo será Si* I *no es frecuente, y* $I' \subseteq I$ *then* I' *no es frecuente*
Esto nos dá una estrategia de poda

Algoritmo Apriori: encontrar k-Itemsets frecuentes

Descripción del algoritmo

1. Hacer k=0, sean C_k el conjunto de k-itemsets candidatos de la base de datos, L_k conjunto de k-itemset frecuentes. $C_0 = \emptyset$ y $L_0 = \emptyset$
2. Hacer $k = k + 1$
3. Calcular C_k añadiendo ordenadamente los elementos de los itemsets de $L_{(k-1)}$ a los $(k-1)$-itemsets de $L_{(k-1)}$ para formar k-itemsets
4. Calcular el soporte de los elementos de C_k. Si hay elementos frecuentes generar L_k e ir a 2. En caso contrario parar



Reglas de asociación: generación

Algoritmo Apriori: encontrar k-Itemsets frecuentes

Aplicando la propiedad "a priori"

Ejemplo

r

100	A_1	B_2	C_0
200	A_2	B_1	C_1
300	A_2	B_2	C_1
400	A_3	B_1	C_2
500	A_4	B_1	C_2
600	A_2	B_1	C_2

C_1

A_1
A_2
.
.
.
C_1
C_2

L_1 (1-Itemsets frecuentes)

Itemset	Soporte
$\{A_2\}$	3
$\{B_1\}$	4
$\{B_2\}$	2
$\{C_1\}$	2
$\{C_2\}$	3

Reglas de asociación: generación

Algoritmo Apriori: encontrar k-Itemsets frecuentes

Aplicando la propiedad "a priori": Construimos C_2, cogiendo pares de items de L_1. En general:

$$\boxed{L_{k-1} \longrightarrow C_k}$$

C_2 = 2-Itemsets candidatos a ser frecuentes = apriori_gen(L_1)

$L_1 =$

Itemset	Soporte
$\{A_2\}$	3
$\{B_1\}$	4
$\{B_2\}$	2
$\{C_1\}$	2
$\{C_2\}$	3

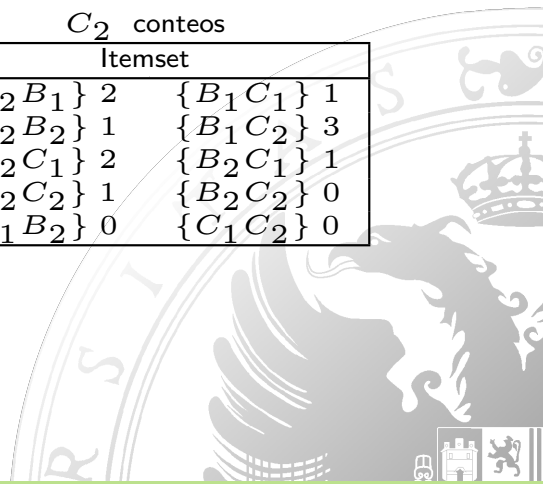
$C_2 =$

C_2

Itemset	
$\{A_2B_1\}$	$\{B_1C_1\}$
$\{A_2B_2\}$	$\{B_1C_2\}$
$\{A_2C_1\}$	$\{B_2C_1\}$
$\{A_2C_2\}$	$\{B_2C_2\}$
$\{B_1B_2\}$	$\{C_1C_2\}$

C_2 conteos

Itemset	
$\{A_2B_1\}$ 2	$\{B_1C_1\}$ 1
$\{A_2B_2\}$ 1	$\{B_1C_2\}$ 3
$\{A_2C_1\}$ 2	$\{B_2C_1\}$ 1
$\{A_2C_2\}$ 1	$\{B_2C_2\}$ 0
$\{B_1B_2\}$ 0	$\{C_1C_2\}$ 0



Reglas de asociación: generación

Algoritmo Apriori: encontrar k-Itemsets frecuentes

Aplicando la propiedad "a priori":
La obtención de L_k a partir de C_k (con conteos) es inmediata:

L_2 (2-Itemsets frec.)

Itemset	Soporte
$\{A_2 B_1\}$	2
$\{B_1 C_2\}$	3
$\{A_2 C_1\}$	2

L_2 contiene los 2-itemsets frecuentes de r

$$L_2 \longrightarrow C_3$$

L_2 (2-Itemsets frec.)

Itemset	Soporte
$\{A_2 B_1\}$	2
$\{B_1 C_2\}$	3
$\{A_2 C_1\}$	2

Algoritmo Apriori: encontrar k-Itemsets frecuentes

Aplicando la propiedad "a priori"

Ahora se construye C_3 = 3-Itemsets candidatos a ser frecuentes, a partir de L_2. La función se denomina *apriori_gen*

$$C_3 = \text{apriori\_gen}(L_2)$$

Estrategia de poda en apriori_gen:

No generar en C_k un k-itemset tal que algún (k-1)-itemset contenido en él, no pertenezca a L_{k-1}

Reglas de asociación: generación

Algoritmo Apriori: Generación de Reglas

Una vez obtenidos los itemsets frecuentes, se generan las reglas de asociación.
Supongamos que $A_1B_3C_2$ es frecuente.
También lo será $A_1B_3, \cdots, A_1, \cdots$
Por lo tanto, todas las reglas

$$A_1B_3 \Rightarrow C_2, C_2 \Rightarrow B_3, \cdots$$

tienen soporte alto.

Reglas de asociación: generación

Algoritmo Apriori: Generación de Reglas

Idea básica

Estrategia de poda (ahora entra en juego el umbral de confianza)

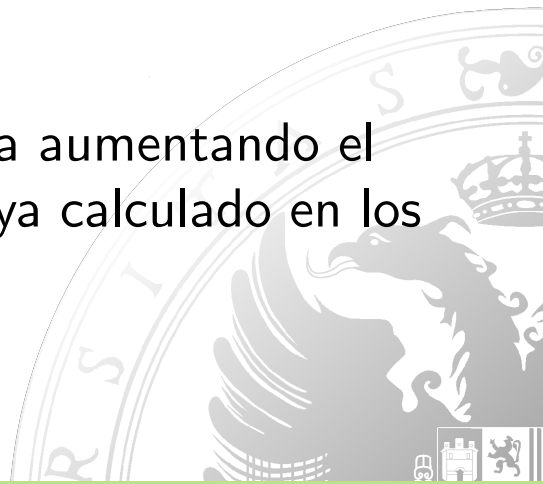
$$conf(A_1B_3 \Rightarrow C_2) = \frac{\text{num tuplas con } A_1B_3C_2}{\text{num tuplas con } A_1B_3} >$$

$$> \frac{\text{num tuplas con } A_1B_3C_2}{\text{num tuplas con } A_1} = conf(A_1 \Rightarrow B_3C_2) \Longrightarrow$$

$$\Longrightarrow \text{ Si } conf(A_1B_3 \Rightarrow C_2) < \epsilon \Longrightarrow conf(A_1 \Rightarrow C_2B_3) < \epsilon$$

$$\text{Si } A_1B_3 \not\Rightarrow C_2 \text{ entonces } A_1 \not\Rightarrow C_2B_3$$

Primero se generan reglas con un sólo consecuente, y se va aumentando el número de consecuentes. El soporte de cada itemset está ya calculado en los C_k



Reglas de asociación: generación

Algoritmo Apriori

Resumiendo

- Al principio, se imponen umbrales de soporte y confianza.
- Estrategia divide y vencerás:
  1. Cálculo de los itemsets frecuentes.
     Operación costosa.
     Estrategias de poda usando el soporte
  2. Cálculo de las reglas de asociación
     Apenas lleva tiempo
     Estrategias de poda usando la confianza

Reglas de asociación: generación

Algoritmo Apriori: Mejoras y alternativas

- Utilizar una relación auxiliar. En esta nueva relación se almacena el identificador de cada registro, junto con los itemsets frecuentes que contiene.
- Utilización de una tabla Hash
- Alternativas de exploración del retículo de subconjuntos
- Alternativas basadas en la exploración de árboles

Reglas de asociación: problemas

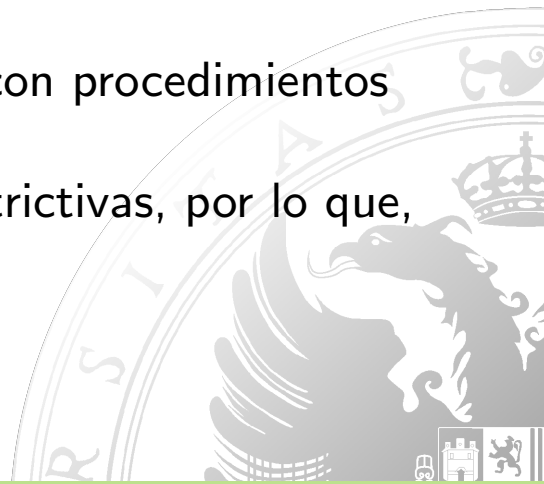
Filtraje de reglas no interesantes

Problema

Número de reglas generado excesivo: varios miles como mínimo.

Soluciones

- *Guiado por el usuario*.
  1. El usuario establece a priori las reglas que él considera interesantes y el sistema las compara con las obtenidas.
  2. El usuario visualiza las reglas y selecciona la parte de ellas que le interesa
- *Sin ayuda del usuario*.
  1. Ordenar las reglas según un grado de interés calculado con procedimientos estadísticos.
  2. Usando medidas de bondad alternativas mucho más restrictivas, por lo que, obviamente, salen menos reglas



Reglas de asociación: problemas

Visualización de reglas

Tipos de técnicas

- técnicas basadas en **tablas**
- Técnicas basadas en **matrices 2D**.
- Técnicas basadas en **grafos**
- Técnicas basadas en coordenadas **paralelas**

Reglas de asociación

Visualización de reglas

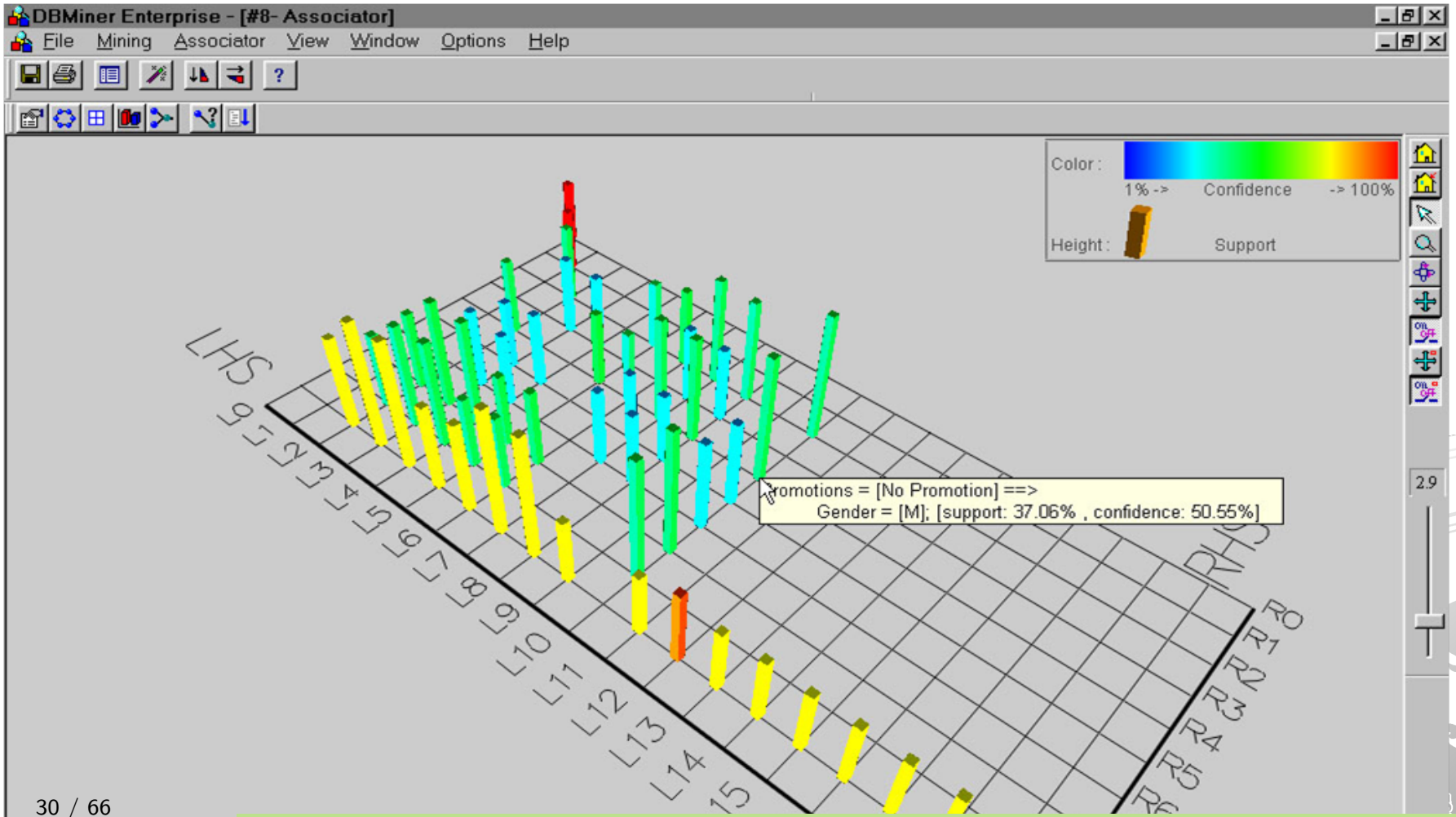
Ejemplo de tablas

	Body	Implies	Head	Supp (%)	Conf (%)	F	G	H	I
1	cost(x) = '0.00~1000.00'	==>	revenue(x) = '0.00~500.00'	28.45	40.4				
2	cost(x) = '0.00~1000.00'	==>	revenue(x) = '500.00~1000.00'	20.46	29.05				
3	cost(x) = '0.00~1000.00'	==>	order_qty(x) = '0.00~100.00'	59.17	84.04				
4	cost(x) = '0.00~1000.00'	==>	revenue(x) = '1000.00~1500.00'	10.45	14.84				
5	cost(x) = '0.00~1000.00'	==>	region(x) = 'United States'	22.56	32.04				
6	cost(x) = '1000.00~2000.00'	==>	order_qty(x) = '0.00~100.00'	12.91	69.34				
7	order_qty(x) = '0.00~100.00'	==>	revenue(x) = '0.00~500.00'	28.45	34.54				
8	order_qty(x) = '0.00~100.00'	==>	cost(x) = '1000.00~2000.00'	12.91	15.67				
9	order_qty(x) = '0.00~100.00'	==>	region(x) = 'United States'	25.9	31.45				
10	order_qty(x) = '0.00~100.00'	==>	cost(x) = '0.00~1000.00'	59.17	71.86				
11	order_qty(x) = '0.00~100.00'	==>	product_line(x) = 'Tents'	13.52	16.42				
12	order_qty(x) = '0.00~100.00'	==>	revenue(x) = '500.00~1000.00'	19.67	23.88				
13	product_line(x) = 'Tents'	==>	order_qty(x) = '0.00~100.00'	13.52	98.72				
14	region(x) = 'United States'	==>	order_qty(x) = '0.00~100.00'	25.9	81.94				
15	region(x) = 'United States'	==>	cost(x) = '0.00~1000.00'	22.56	71.39				
16	revenue(x) = '0.00~500.00'	==>	cost(x) = '0.00~1000.00'	28.45	100				
17	revenue(x) = '0.00~500.00'	==>	order_qty(x) = '0.00~100.00'	28.45	100				
18	revenue(x) = '1000.00~1500.00'	==>	cost(x) = '0.00~1000.00'	10.45	96.75				
19	revenue(x) = '500.00~1000.00'	==>	cost(x) = '0.00~1000.00'	20.46	100				
20	revenue(x) = '500.00~1000.00'	==>	order_qty(x) = '0.00~100.00'	19.67	96.14				
21									
22									
23	cost(x) = '0.00~1000.00'	==>	revenue(x) = '0.00~500.00' AND order_qty(x) = '0.00~100.00'	28.45	40.4				
24	cost(x) = '0.00~1000.00'	==>	revenue(x) = '0.00~500.00' AND order_qty(x) = '0.00~100.00'	28.45	40.4				
25	cost(x) = '0.00~1000.00'	==>	revenue(x) = '500.00~1000.00' AND order_qty(x) = '0.00~100.00'	19.67	27.93				
26	cost(x) = '0.00~1000.00'	==>	revenue(x) = '500.00~1000.00' AND order_qty(x) = '0.00~100.00'	19.67	27.93				
27	cost(x) = '0.00~1000.00' AND order_qty(x) = '0.00								

Reglas de asociación

Visualización de reglas

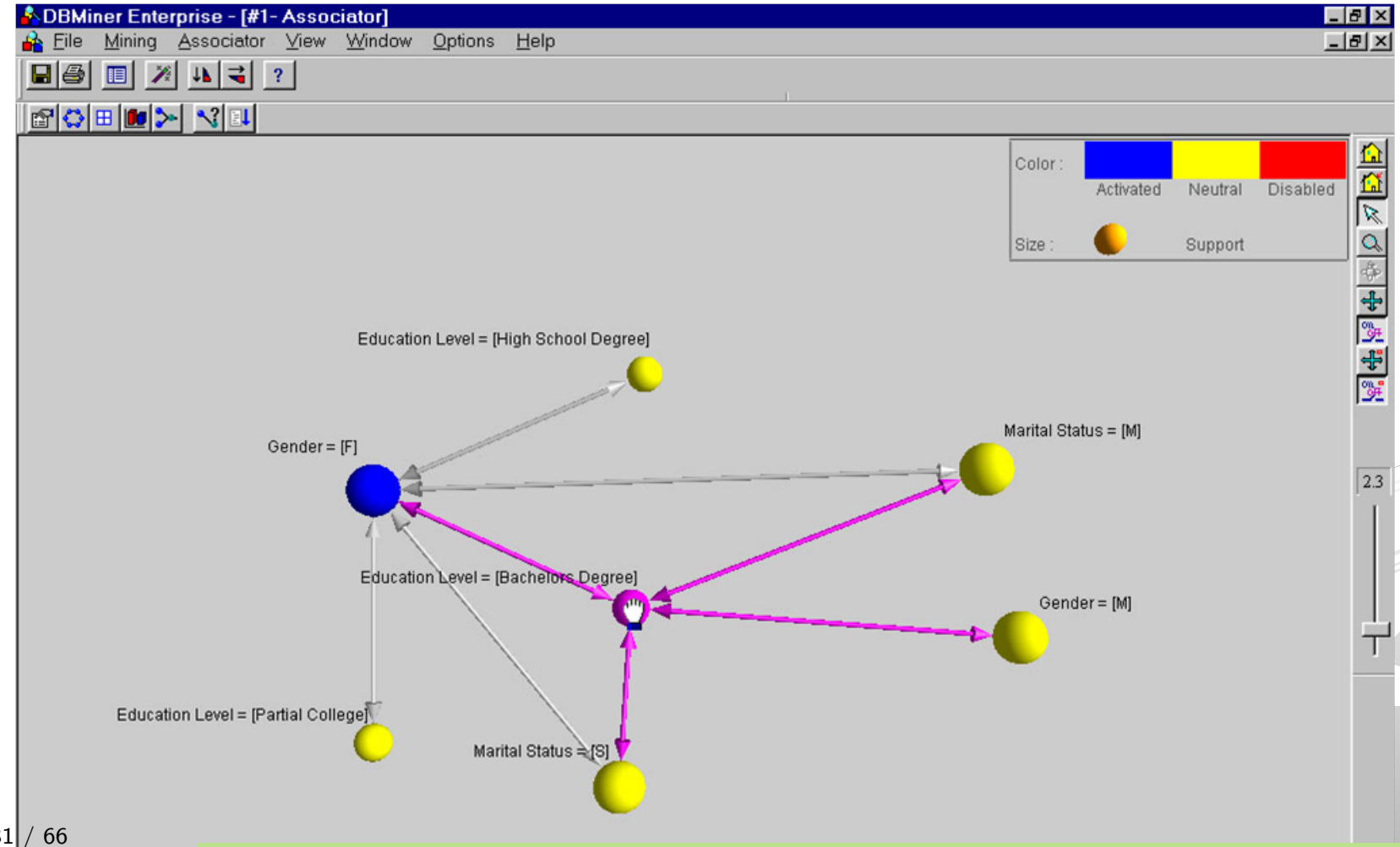
Ejemplo de gráfico 2D



Reglas de asociación

Visualización de reglas

Ejemplo de grafos



Reglas de asociación: otras medidas

Inconvenientes con la confianza

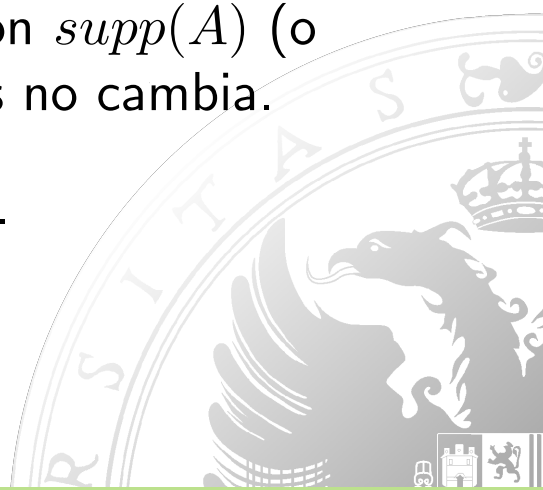
- Propiedades para medidas de cumplimiento (Piatetsky-Shapiro, 1991):

  P1 $ACC(A \Rightarrow C) = 0$ cuando $Supp(A \Rightarrow C) = supp(A)\ supp(C)$.

  P2 $ACC(A \Rightarrow C)$ monótono creciente con $Supp(A \Rightarrow C)$ cuando el resto de parámetros no cambia.

  P3 $ACC(A \Rightarrow C)$ monótono decreciente con $supp(A)$ (o $supp(C)$) cuando el resto de parámetros no cambia.

- Confianza no verifica **P1** y **P3** para el caso de supp(C).



Reglas de asociación: otras medidas

Inconvenientes con la confianza. Confianza y **P1**

i_1	i_2	i_3	i_4
1	0	1	0
0	0	0	0
0	1	1	0
0	1	1	1
1	1	1	1
1	1	1	1

Itemset	**Soporte**
$\{i_1\}$	$1/2$
$\{i_2\}$	$2/3$
$\{i_1, i_2\}$	$1/3$

$$Conf(\{i_1\} \Rightarrow \{i_2\}) = \frac{supp(\{i_1,i_2\})}{supp(\{i_1\})} = \frac{1/3}{1/2} = 2/3 \neq 0.$$

Reglas de asociación: otras medidas

Inconvenientes con la confianza. Confianza y **P3**

i_1	i_2	i_3	i_4
1	0	1	0
0	0	0	0
0	1	1	0
0	1	1	1
1	1	1	1
1	1	1	1

Itemset	**Soporte**
$\{i_1\}$	$1/2$
$\{i_4\}$	$1/2$
$\{i_1, i_4\}$	$1/3$

$supp(\{i_4\}) = p(\{i_4\}) = 1/2$
$Conf(\{i_1\} \Rightarrow \{i_4\}) = p(\{i_4\}|\{i_1\}) = 2/3 > 1/2.$

Reglas de asociación: otras medidas

Inconvenientes con el soporte

- Principio clásico:: "cuanto mayor el soporte, mejor el itemset".
- Sea C un itemset con soporte muy alto.
  - Cualquier otro itemset A parece ser un buen predictor de C.
  - Problema: falta de variabilidad en C.
- Casos:
  - $Conf(A \Rightarrow C) \leq supp(C)$ (dependencia negativa ó independencia) $\Rightarrow$ una medida adecuada de cumplimiento descartaría la regla.
  - $Conf(A \Rightarrow C) > supp(C)$ $\Rightarrow$ el cumplimiento puede ser muy alto, no se descartaría!!
- No se suele chequear supp(C)>>

Reglas de asociación: otras medidas

Consecuencias de utilizar soporte/confianza

- Se genera una gran cantidad de reglas dudosas $A \Rightarrow C$, incluyendo:
  - Reglas que verifican independencia/dependencia negativa.
  - Reglas con falta de variabilidad en C.
- Ejemplo real (Brin y otros, 1997):
  - "ha prestado servicio activo en el ejército $\Rightarrow$ no ha servido en Vietnam" se cumple en la base de datos del censo de los EEUU con confianza 0.9.
  - Claro que, supp("no ha servido en Vietnam")=0.95 $\Rightarrow$ Dependencia negativa.

Reglas de asociación: otras medidas

El LIFT

la medida de Implicación, también llamada Interés, o **Lift**

$$Lift(A,C) = \frac{conf(A \Rightarrow C)}{supp(C)} = \frac{supp(A \bigcup C}{Supp(A)supp(C)} = \frac{Prob(A,C)}{Prob(A)Prob(C)}$$

Ventajas

- $\forall A, C Lift(A,C) \in [0, +\infty]$
- Pondera la confianza con respecto a lo probable que es C. Si A y C fuesen perfectamente independientes, $prob(A,C) = prob(A)prob(C)$, luego $Lift(A,C) = 1$. Por lo tanto, lift nos mide hasta qué punto ocurren conjuntamente A y C más o menos de lo esperado si fuesen independientes.

Reglas de asociación: otras medidas

El LIFT

Ventajas

- Un valor mayor de 1 indica que A tiene un efecto positivo en la aparición de C.
- Un valor menor de 1 indica que A tiene un efecto negativo en la aparición de C.
- Un valor cercano a 1 indica que A apenas tiene efecto en la aparición de C.

Inconvenientes

- No está acotada y probabilidades muy pequeñas en los items set dan valores muy grandes
- Es una medida simétrica (Lift(A,C)=Lift(C,A))

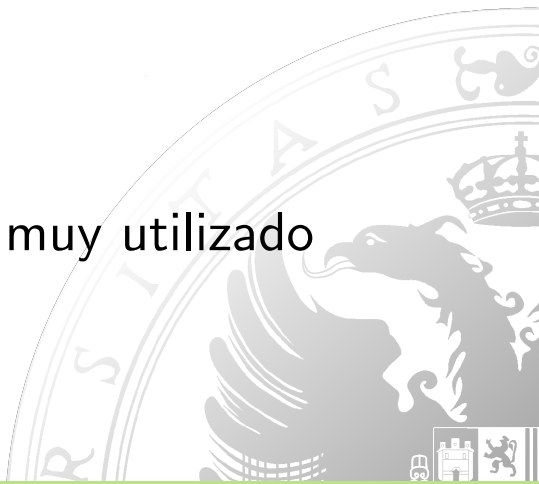
Reglas de asociación: otras medidas

El factor de certeza

- El factor de certeza de $A \Rightarrow C$ (Shortliffe and Buchanan, 1975):

$$CF(A \Rightarrow C) = \begin{cases} \frac{Conf(A \Rightarrow C) - supp(C)}{1 - supp(C)} & \text{si } Conf(A \Rightarrow C) > supp(C) \\ \frac{Conf(A \Rightarrow C) - supp(C)}{supp(C)} & \text{si } Conf(A \Rightarrow C) < supp(C) \\ 0 & \text{en otro caso} \end{cases}$$

- $CF(A \Rightarrow C) \in [-1, 1]$.
- Verifica **P1, P2, P3**.
- Verifica un buen conjunto de propiedades que lo hacen muy utilizado



Reglas de asociación: otras medidas

El factor de certeza

Propiedades del factor de certeza

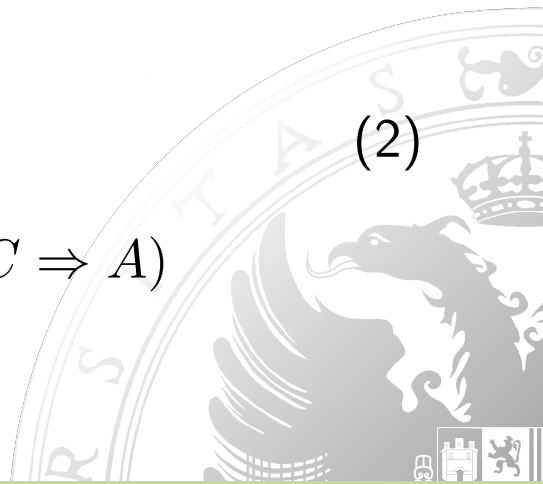
1. $CF(A \Rightarrow C) \leq Conf(A \Rightarrow C)$.
2. $CF(A \Rightarrow C) = Conf(A \Rightarrow C)$ iff $supp(C) < 1$ y $CF(A \Rightarrow C) = 1$.
3. Sea $CF(A \Rightarrow C) > 0$, $supp(C) < 1$ y $supp(A) > 0$. Entonces

$$CF(A \Rightarrow C) = 1 - \frac{1}{Conf(A \Rightarrow C)} \tag{1}$$

4. Sea $CF(A \Rightarrow C) < 0$ y $supp(C) > 0$. Entonces

$$CF(A \Rightarrow C) = Lift(A \Rightarrow C) - 1 \tag{2}$$

5. Sea $CF(A \Rightarrow C) < 0$. Entonces $CF(A \Rightarrow C) = CF(C \Rightarrow A)$
6. $CF(A \Rightarrow C)CF(C \Rightarrow A) > 0$



Reglas de asociación: otras medidas

El factor de certeza

Propiedades del factor de certeza

1. Si $CF(A \Rightarrow C) = -CF(A \Rightarrow \neg C)$.
2. Si $CF(A \Rightarrow C) > 0$ entonces $CF(A \Rightarrow C) = CF(\neg C \Rightarrow \neg A)$.
3. El factor de certeza de todas las reglas válidas que involucran a A, C, $\neg A$ y $\neg C$ (8 reglas) toma solo 4 valores distintos, que se quedan en 2 si tomamos valor absoluto.

Reglas de asociación: otras medidas

Resumen medidas alternativas

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\left(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),\right.$ $\left. P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\right)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2] + P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2] + P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A), P(A\|B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\left(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$

Reglas de asociación: otras medidas

Propiedades de otras medidas alternativas

Sym bol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
Φ	Correlation	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Lambda	0 ... 1	Yes	No	No	Yes	No	No*	Yes	No
α	Odds ratio	0 ... 1 ... ∞	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	0 ... 1	Yes	Yes	Yes	Yes	No	No*	Yes	No
J	J-Measure	0 ... 1	Yes	No	No	No	No	No	No	No
G	Gini Index	0 ... 1	Yes	No	No	No	No	No*	Yes	No
s	Support	0 ... 1	No	Yes	No	Yes	No	No	No	No
c	Confidence	0 ... 1	No	Yes	No	Yes	No	No	No	Yes
L	Laplace	0 ... 1	No	Yes	No	Yes	No	No	No	No
V	Conviction	0.5 ... 1 ... ∞	No	Yes	No	Yes**	No	No	Yes	No
I	Interest	0 ... 1 ... ∞	Yes*	Yes	Yes	Yes	No	No	No	No
IS	IS (cosine)	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	-0.25 ... 0 ... 0.25	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	-1 ... 0 ... 1	Yes	Yes	Yes	No	No	No	Yes	No
AV	Added value	0.5 ... 1 ... 1	Yes	Yes	Yes	No	No	No	No	No
S	Collective strength	0 ... 1 ... ∞	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$\left(\sqrt{\frac{2}{\sqrt{3}}-1}\right)\left(2-\sqrt{3}-\frac{1}{\sqrt{3}}\right)...0...\frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No	No	No	No	No

Reglas de asociación: variantes y extensiones

Tratamiento de atributos continuos

Los atributos continuos no son directamente utilizables en reglas de asociación

Es necesario discretizar previamente

- Distintos tipos de reglas con atributos continuos

$$Edad \in [21, 35) \wedge Salario \in [40k, 60k) \rightarrow Compra$$
$$Salario \in [30k, 60k) \wedge Compra \rightarrow Edad : \mu = 28, \sigma =$$

- Solución= dicretización
  - Equidistante
  - Equiprobable
  - Basado en clustering
  - Basado en Lógica difusa $\Longrightarrow$ **Reglas de asociación difusas**

Reglas de asociación: variantes y extensiones

Reglas de asociacion difusas

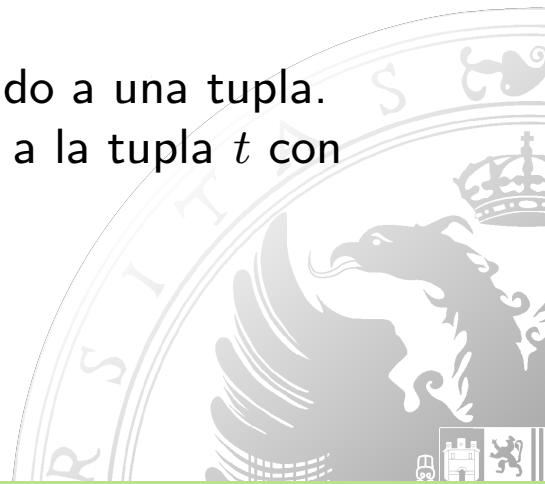
La idea básica es considerar intervalos difusos de los atributos continuos, caracterizados por etiquetas lingüísiticas

Ejemplos:

Edad= {*Niño, Joven, Maduro, Anciano*}

Salario ={ *Alto, Medio, Bajo* }

- *Nuevos conceptos:*
  - **Item:** par (Atributo,Etiqueta)
  - **Transacción difusa:** subconjunto difuso de items asociado a una tupla.
  - El item (At, L) está presente en la transacción asociada a la tupla t con grado $L(t[At])$.



Reglas de asociación: variantes y extensiones

Reglas de asociacion difusas

- Una **transacción difusa** $\tilde{\tau}$ es un subconjunto difuso de I.

$$\tilde{\tau} \in \tilde{\wp}(I)$$

- Para cada $i \in I$, $\tilde{\tau}(i)$ es el grado de pertenencia de i a $\tilde{\tau}$.
- Sea $I_0 \subseteq I$. Entonces $\tilde{\tau}(I_0) = \min_{i \in I_0} \tilde{\tau}(i)$.
- Un **FT-set** T es un conjunto (crisp) de transacciones difusas.

$$T \subseteq \tilde{\wp}(I)$$

Reglas de asociación: variantes y extensiones

Reglas de asociacion difusas

EJEMPLO: Transacciones difusas en bases de datos relacionales

I={ (atributo, etiqueta difusa) }

Tupla $t_i \rightarrow$ Transacción difusa $\tilde{\tau}_i$
$\tilde{\tau}_i((Atr, Lab)) = Lab(t_i[Atr])$

Label(Altura) = {bajo, medio, alto, muy alto}
Label(Peso) = {ligero, medio, pesado, muy pesado}

	Peso	Altura
t_1	70	170
t_2	68	180
t_3	60	175
t_4	90	175
t_5	50	195

$\Rightarrow$

	$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}_3$	$\tilde{\tau}_4$	$\tilde{\tau}_5$
(Altura,bajo)	0	1	0	0	0
(Altura,medio)	0	0.8	1	0	1
(Altura,alto)	1	0	1	0.5	0.5
(Altura,m.alto)	0.5	0	0	1	0
(Peso,ligero)	0	0	0	0	0
(Peso,medio)	1	1	0.5	0.5	0.25
(Peso,pesado)	0.5	1	1	1	1
(Peso,m.pesado)	0	0	0	0	0.25

Reglas de asociación: variantes y extensiones

Reglas de asociacion difusas

- **Una regla de asociación difusa es una regla de asociación en un FT-set T**
- La R.A.D. $I_1 \Rightarrow I_2$ tiene máximo cumplimiento en T sii

$$\forall \tilde{\tau} \in T \quad \tilde{\tau}(I_1) \leq \tilde{\tau}(I_2)$$

- Una **R.A.** es un caso particular de **R.A.D.**

EJEMPLOS

- (Altura,bajo) $\Rightarrow$ (Peso,medio) presenta total cumplimiento en T
- (Altura,m.alto) $\Rightarrow$ (Peso,m.pesado)) no.

	$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}_3$	$\tilde{\tau}_4$	$\tilde{\tau}_5$
(Altura,bajo)	0	1	0	0	0
(Altura,medio)	0	0.8	1	0	1
(Altura,alto)	1	0	1	0.5	0.5
(Altura,m.alto)	0.5	0	0	1	0
(Peso,ligero)	0	0	0	0	0
(Peso,medio)	1	1	0.5	0.5	0.25
(Peso,pesado)	0.5	1	1	1	1
(Peso,m.pesado)	0	0	0	0	0.25

Medidas de calidad de una R.A.D.

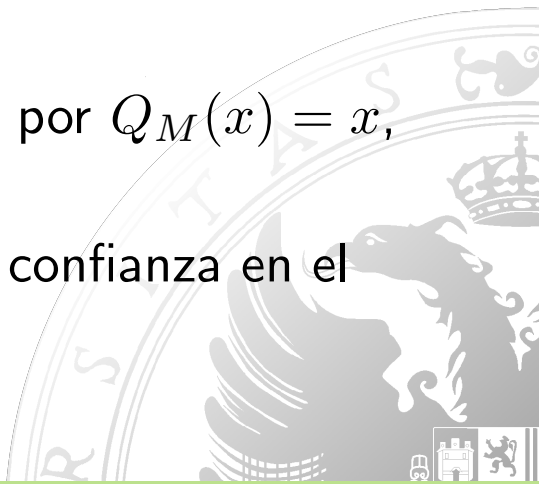
- Sea Q un cuantificador relativo difuso.
- **El soporte de la R.A.D.** $I_1 \Rightarrow I_2$ es la evaluación de la sentencia cuantificada

$$Q \text{ de los } T \text{ son } \tilde{\Gamma}_{I_1} \cap \tilde{\Gamma}_{I_2}$$

- **Confianza:** evaluación de la sentencia

$$Q \text{ de los } \tilde{\Gamma}_{I_1} \text{ son } \tilde{\Gamma}_{I_2}$$

- En general usaremos el cuantificador Q_M caracterizado por $Q_M(x) = x$, $\forall x \in [0, 1]$.
- Propiedad: generaliza las medidas usuales de soporte y confianza en el caso crisp.



Reglas de asociación: variantes y extensiones

Reglas de asociacion difusas: ejemplos de uso de B.D. reales

- Bases de datos reales
  - INTERVENCIONES en el Hospital Universitario S. Cecilio de Granada, entre Agosto del 97 y Agosto del 98. 15766 tuplas.
  - URGENCIAS en el mismo hospital. 81368 tuplas.

Reglas de asociación: variantes y extensiones

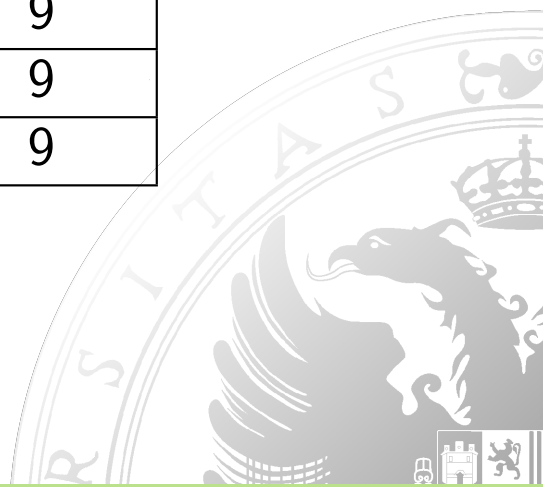
Reglas de asociacion difusas: ejemplos de uso de B.D. reales

R.A. DIFUSAS EN LA BASE DE DATOS URGENCIAS

Extracción de asociaciones entre valores de los atributos "Hora de Ingreso" y "Tipo de Asistencia".

Conjuntos de 2 items más frecuentes

2-Itemsets	**Tuplas**
[HINGRESO=22:45][CLASISTENCIA=CONSULTA]	11
[HINGRESO=11:00][CLASISTENCIA=CONSULTA]	9
[HINGRESO=20:15][CLASISTENCIA=CONSULTA]	9
[HINGRESO=21:10][CLASISTENCIA=CONSULTA]	9

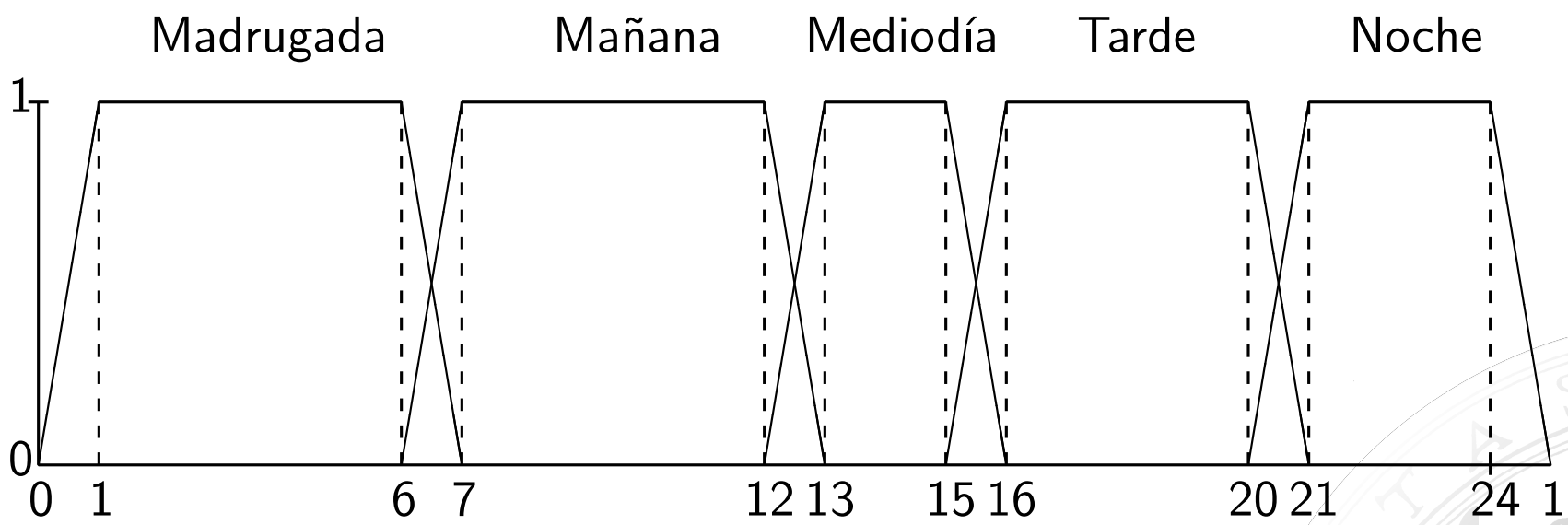


Reglas de asociación: variantes y extensiones

Reglas de asociacion difusas: ejemplos de uso de B.D. reales

R.A. DIFUSAS EN LA BASE DE DATOS URGENCIAS

Etiquetas difusas para Hora:



Reglas de asociación: variantes y extensiones

Reglas de asociacion difusas: ejemplos de uso de B.D. reales

R.A. DIFUSAS EN LA BASE DE DATOS URGENCIAS Items para el atributo "Hora de Ingreso"

1-Itemsets	**Soporte**
[HINGRESO=MAÑANA]	0.267
[HINGRESO=MEDIODIA]	0.160
[HINGRESO=TARDE]	0.315
[HINGRESO=NOCHE]	0.179
[HINGRESO=MADRUGADA]	0.074

Reglas de asociación: variantes y extensiones

Reglas de asociacion difusas: ejemplos de uso de B.D. reales

R.A. DIFUSAS EN LA BASE DE DATOS URGENCIAS Comparación de las metodologías clásica (sin agrupamiento) y difusa (agrupamiento mediante etiquetas).

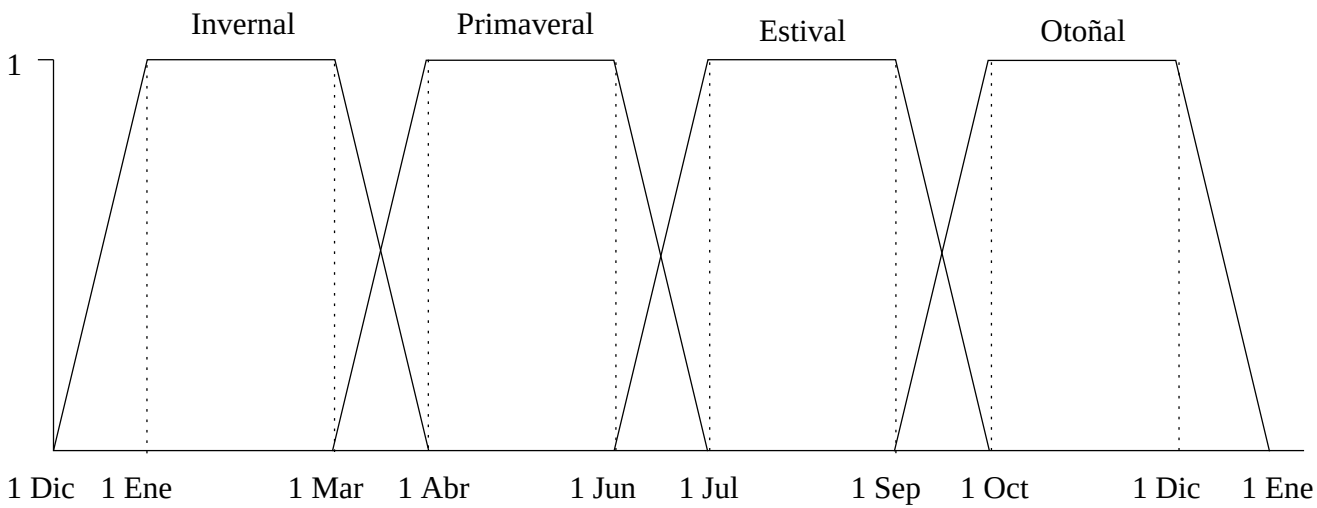
	Clásico	**Difuso**
Tuplas	81368	81368
Items	47543	13
Itemsets en tabla	101958	53
Reglas potenciales	203916	106
Reglas con sop>0.02	0	12
Tiempo empleado	Interrumpido tras 1 hora	2m16s
Memoria empleada	>250Mb	350Kb

R.A. DIFUSAS EN LA BASE DE DATOS URGENCIAS

- Algunas reglas descubiertas:
  - [CLASISTENCIA=YESOS] ⇒ [HINGRESO=TARDE] sop: 0.022 FC: 0.48
  - [HINGRESO=MAÑANA] ⇒ [CLASISTENCIA=OBSERVACIÓN] sop: 0.128 FC: 0.43

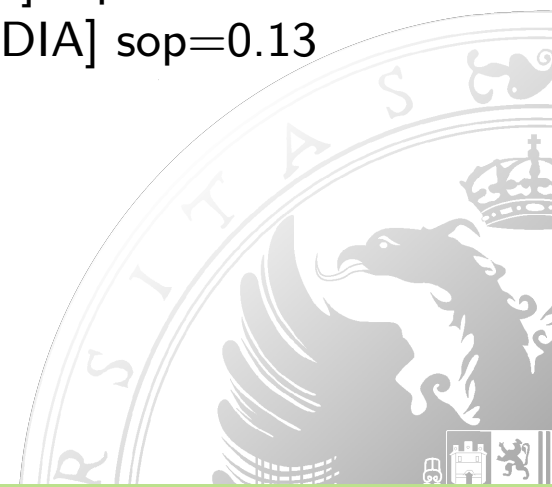
R.A. DIFUSAS EN LA BASE DE DATOS INTERVENCIONES

Etiquetas difusas para Fecha:



R.A. DIFUSAS EN LA BASE DE DATOS INTERVENCIONES

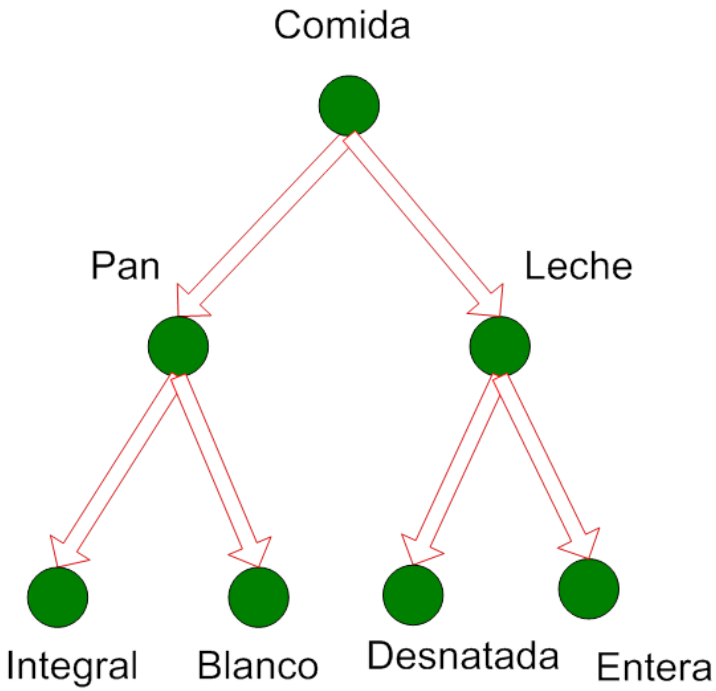
- Algunas reglas descubiertas:
  - [FECHA=ESTIVAL] ⇒ [SUSPENDIDA=N] sop=0.15 FC=0.319
  - [HTERMINO=MAÑANA] ⇒ [HCOMIENZO=MAÑANA] sop=0.41 FC=0.99
  - [HCOMIENZO=MAÑANA] ⇒ [HTERMINO=MAÑANA] sop=0.41 FC=0.5
  - [HCOMIENZO=MEDIODIA] ⇒ [HTERMINO=MEDIODIA] sop=0.13 FC=0.91

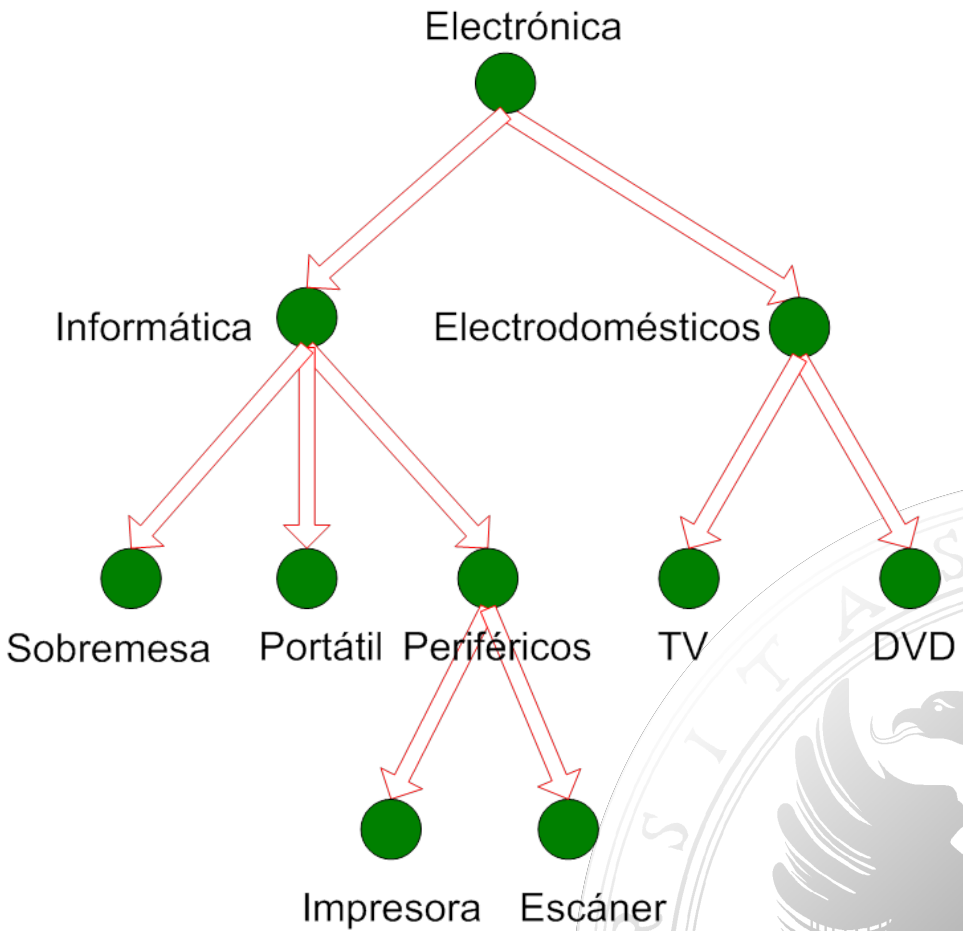


Reglas de asociación: variantes y extensiones

Reglas multinivel

Ejemplo de atributos jerárquicos





Reglas de asociación: variantes y extensiones

Reglas multinivel

Por qué utilizar jerarquias de conceptos?

- Porque las reglas que involucran artículos en los niveles más bajos puede que no tengan soporte suficiente como para aparecer en algún patrón frecuente.
- Porque las reglas a niveles bajos de la jerarquía son demasiado especificas.

$$leche\ desnatada \rightarrow pan\ blanco$$

$$leche\ entera \rightarrow pan\ integral$$

$$leche\ desnatada \rightarrow pan\ integral$$

indican una asociación entre pan y leche.

Reglas de asociacion: variantes y extensiones

Reglas multinivel

Soluciones para el tratamiento multinivel

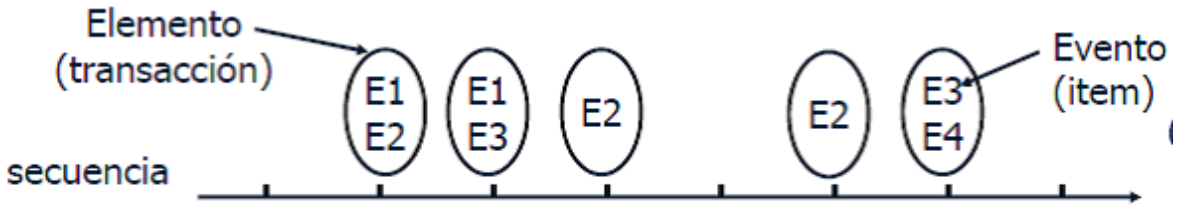
- Realizar un análisis previo y agregar los atributos de forma heurística
- Trabajar con un cubo de datos como datos de partida, siendo las dimensiones los atributos del problema:
  1. Se calculan las reglas a nivel más bajo
  2. Se establecen medidas de especificidad adicionales para dar una medida calidad combinada
  3. Se eliminan las reglas muy específicas y se agregan los atributos involucrados mediante operaciones en el cubo de datos
  4. Se recalculan las reglas relativas a los atributos agregados
  5. La salida final es un conjunto de reglas a varios niveles

El proceso descrito se conoce como **Olap Mining**

Reglas de asociacion: variantes y extensiones

Analisis de secuencias. Obtencion de patrones frecuentes

Problema



Ejemplo

Base de datos	Secuencia	Elemento (Transacción)	Evento (Item)
Clientes	Historial de compras de un cliente determinado	Conjunto de artículos comprados por un cliente en un instante concreto	Libros, productos...
Web	Navegación de un visitante del sitio web	Colección de ficheros vistos por el visitante tras un único click de ratón	Página inicial, información de contacto, fotografía...
Eventos	Eventos generados por un sensor	Eventos generador por un sensor en un instante t	Tipos de alarmas generadas
Genoma	Secuencia de ADN	Elemento de la secuencia de ADN	Bases A,T,G,C

Reglas de asociacion: variantes y extensiones

Definicion formal

- Sea I conjunto de items, sean $A, B, ..$ subconjuntos de I, definimos **secuencia** $S =< A_1, A_2, ..A_n >$. Pueden existir distintas secuencias en el sistema.
- Dada una secuencia $S =< A_1, A_2, ..A_n >$ decimos que es **subsecuencia** de otra secuencia $T =< B_1, B_2, ..B_m >$ existe una sucesion de enteros $i_1 < i_2, .. < i_n$, tal que:

$$A_1 \subseteq B_{i_1} ... A_n \subseteq B_{i_n}$$

Ejemplo

Secuencia	Subsecuencia	¿incluida?
< {2,4} {3,5,6} {8} >	< {2} {3,5} >	Sí
< {1,2} {3,4} >	< {1} {2} >	No
< {2,4} {2,4} {2,5} >	< {2} {4} >	Sí

Reglas de asociación: variantes y extensiones

Analisis de secuencias. Obtención de patrones frecuentes

- **El soporte** de una subsecuencia es la fraccion de esta subsecuencia que aparece en la base de datos
- Buscar patrones secuenciales es encontrar subsecuencias con un soporte $\geqslant$ que $MINSUPP$
- Se pueden adaptar los algoritmos de búsqueda de itemsets frecuentes

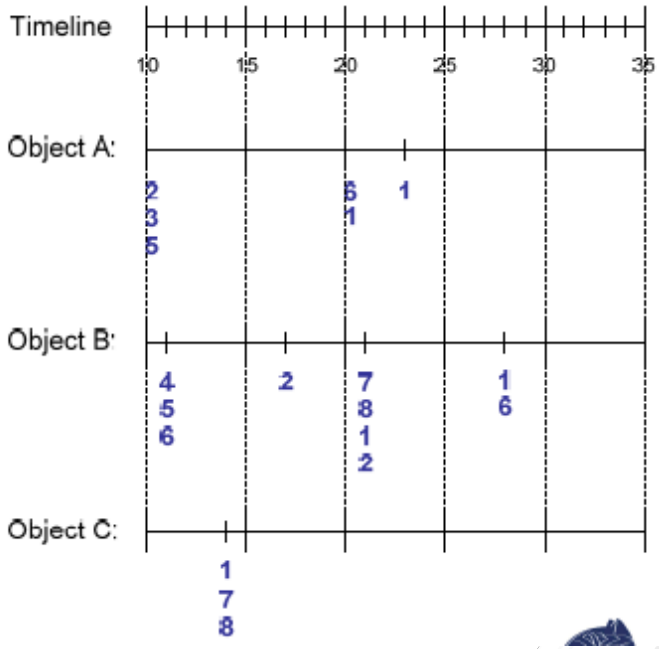
Reglas de asociación: variantes y extensiones

Analisis de secuencias. Obtención de patrones frecuentes

Ejemplo

Object	Timestamp	Events
A	10	2, 3, 5
A	20	6, 1
A	23	1
B	11	4, 5, 6
B	17	2
B	21	7, 8, 1, 2
B	28	1, 6
C	14	1, 7, 8

Base de datos de secuencias



Reglas de asociación: variantes y extensiones

Analisis de secuencias. Obtención de patrones frecuentes

Ejemplo

Object	Timestamp	Events
A	1	1,2,4
A	2	2,3
A	3	5
B	1	1,2
B	2	2,3,4
C	1	1, 2
C	2	2,3,4
C	3	2,4,5
D	1	2
D	2	3, 4
D	3	4, 5
E	1	1, 3
E	2	2, 4, 5

MinSupp = 50%

Ejemplos de subsecuencias frecuentes:

< {1,2} >	s=60%
< {2,3} >	s=60%
< {2,4}>	s=80%
< {3} {5}>	s=80%
< {1} {2} >	s=80%
< {2} {2} >	s=60%
< {1} {2,3} >	s=60%
< {2} {2,3} >	s=60%
< {1,2} {2,3} >	s=60%