DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Extracción de patrones

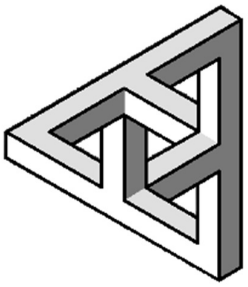
© Fernando Berzal, berzal@acm.org

Extracción de patrones

- Introducción
- Definiciones
  - Patrones frecuentes
  - Reglas de asociación
- Extracción de reglas de asociación
  - Identificación de patrones frecuentes: El algoritmo Apriori
  - Generación de reglas
- Visualización de reglas de asociación
- Evaluación de reglas de asociación
- Extensiones y variaciones
  - Atributos continuos
  - Reglas multinivel (a.k.a. generalizadas)
  - Tipos de patrones: secuencias, estructuras…

Introducción

Relaciones entre atributos: Asociación



Modelo

Descriptivo.

Objetivo

Descubrir "reglas" en lógica proposicional (pero cualificadas probabilísticamente) que involucren valores de atributos.

Datos

Categóricos & numéricos discretizados

Variantes y extensiones

Motif discovery, análisis de secuencias…

Introducción

Problema típico

Dado un conjunto de transacciones, encontrar reglas que describen tendencias en los datos:

Detectar cuándo la ocurrencia de un artículo está asociada a la ocurrencia de otros artículos en la misma transacción.

Introducción

"Market-basket analysis"

Transacciones

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

Reglas de asociación

{pañales} → {cerveza}

{leche, pan} → {huevos}

{cerveza, pan}
→ {leche, huevos}

¡OJO!

Asociación implica co-ocurrencia, no causalidad.

Introducción – Aplicaciones (1)

"Product placement": Colocación de productos en las estanterías de un supermercado

Objetivo

Identificar artículos que muchos clientes compran conjuntamente.

Solución

Procesar los datos de los terminales de punto de venta proporcionados por los escáneres de códigos de barras.

Ejemplo

Si un cliente compra pañales, es muy probable que compre cerveza (¡no se sorprenda si ve las cervezas colocadas al lado de los pañales en el súper!)

Introducción – Aplicaciones (2)

Promociones y ofertas

Si se identifica una regla del tipo: {impresora} → {tóner}

- Tóner en el consecuente
  => Puede determinarse cómo incrementar sus ventas.
- Impresora en el antecedente
  => Puede determinarse qué productos se verían afectados si dejamos de vender impresoras.
- Impresora en el antecedente y tóner en el consecuente
  => Puede utilizarse para ver qué productos deberían venderse con impresoras para promocionar las ventas de tóner.

Introducción – Aplicaciones (3)

Gestión de inventarios

Problema

Una empresa de reparación de electrodomésticos quiere anticipar la naturaleza de las reparaciones que tendrá que realizar y mantener a sus vehículos equipados con las piezas que permitan reducir el número de visitas a casa de sus clientes.

Solución

Procesar los datos sobre herramientas y piezas utilizadas en reparaciones previas para descubrir patrones de co-ocurrencia.

Definiciones

- **Itemset**
  Conjunto de uno o más items (artículos).
  p.ej. {pan, leche}
- **K-itemset**
  Itemset con k elementos.

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

- **Soporte de un itemset [support]**
  Fracción de las transacciones que contienen el itemset.
  p.ej. supp({pan,leche}) = 3/5
- **Itemset frecuente**
  Itemset con soporte igual o superior a un umbral de soporte establecido por el usuario (MinSupp).

Definiciones

Regla de asociación

Expresión de la forma
X → Y
donde X e Y son itemsets.

p.ej. {pañales} → {cerveza}
{cerveza} → {pañales}
{pan, leche} → {huevos}
{pan} → {leche, huevos}

Definiciones

Medidas de evaluación de las reglas de asociación

- **Soporte de la regla:** **supp (X → Y)**

  Fracción de las transacciones que contiene tanto a X como a Y; esto es, supp(XUY).

- **Confianza de la regla:** **conf (X → Y)**

  Fracción de las transacciones en las que aparece X que también incluyen a Y; esto es, la confianza mide con qué frecuencia aparece Y en las transacciones que incluyen X.

Definiciones

Medidas de evaluación de las reglas de asociación

- **Soporte de la regla**

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

- **Confianza de la regla**

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

Definiciones

Medidas de evaluación de las reglas de asociación

supp({pañales}) = 3/5 = 0.6

supp({cerveza}) = 4/5 = 0.8

supp({cerveza}→{pañales})
= supp({pañales, cerveza})
= 3/5 = 0.6 = 60%

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

conf ({cerveza}→{pañales})
= supp({pañales, cerveza}) / supp({cerveza})
= (3/5) / (4/5) = 3/4 = 0.75 = 75%

Extracción de reglas de asociación

Formulación del problema

Dado un conjunto de transacciones T,
encontrar todas las reglas de asociación...

- cuyo soporte sea mayor o igual que un umbral mínimo de soporte, MinSupp:
  supp (X → Y) ≥ MinSupp

- cuya confianza sea mayor o igual que un umbral mínimo de confianza, MinConf:
  conf (X → Y) ≥ MinConf

Formulación del problema

Solución por fuerza bruta

- Enumerar todas las reglas de asociación posibles.
- Calcular el soporte y la confianza de cada regla.
- Eliminar las reglas que no superen los umbrales de soporte y confianza (MinSupp y MinConf).

Computacionalmente prohibitivo…

Formulación del problema

Ejemplo

Reglas derivadas de
{pan, pañales, cerveza}

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

{pan} → {pañales, cerveza}, supp=0.4, conf=2/4=0.5
{pañales} → {pan, cerveza}, supp=0.4, conf=2/3=0.66
{cerveza} → {pan, pañales}, supp=0.4, conf=2/4=0.5
{pan, pañales} → {cerveza}, supp=0.4, conf=2/2=1
{pan, cerveza} → {pañales}, supp=0.4, conf=2/3=0.66
{pañales, cerveza} → {pan}, supp=0.4, conf=2/3=0.66

Formulación del problema

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

Ejemplo

Reglas derivadas de

{pan, pañales, cerveza}

Observaciones

- Todas las reglas anteriores son particiones binarias del mismo itemset ({pan, pañales, cerveza}).
- Todas las reglas que provienen del mismo itemset tienen el mismo soporte, aunque su confianza pueda variar.
- Por tanto, podemos separar la parte que depende del soporte de la que depende de la confianza.

Formulación del problema

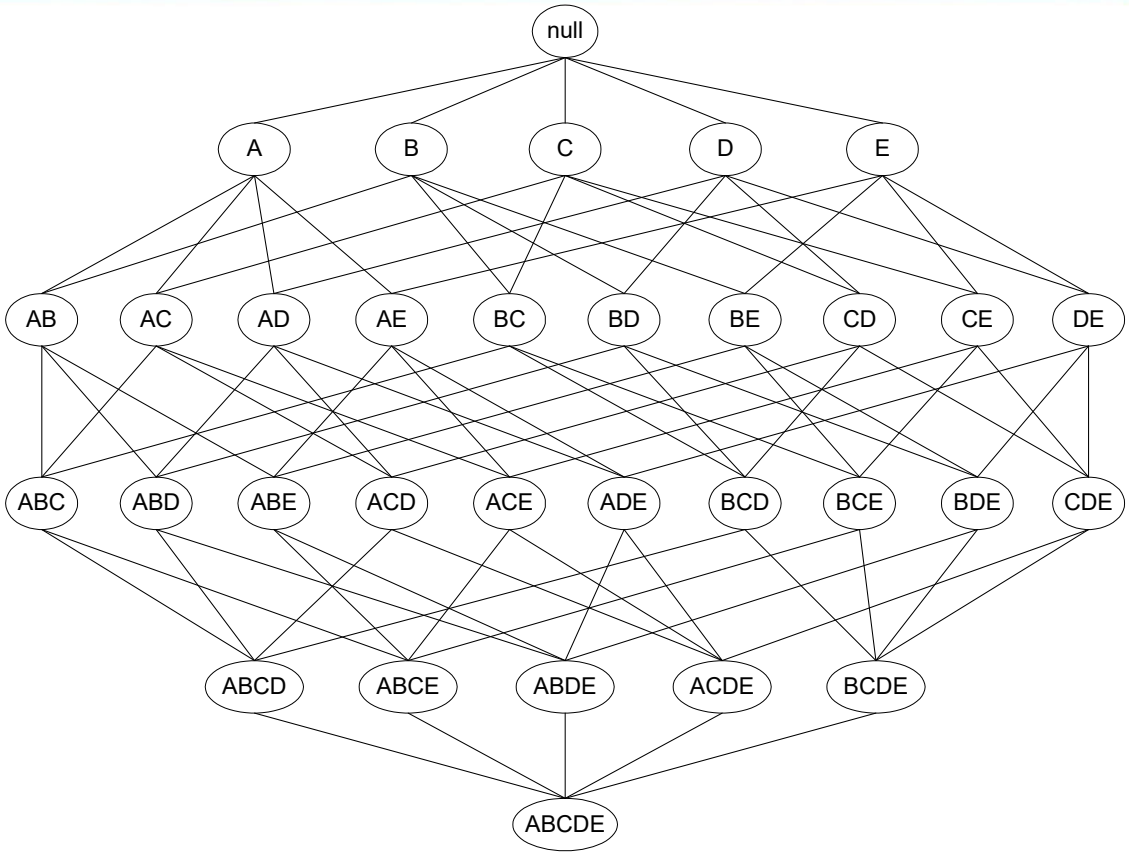
Solución en dos etapas

1. Generación de itemsets frecuentes:
   Identificar los itemsets con soporte **≥ MinSupp**.

2. Generación de reglas de asociación:
   Obtener reglas de asociación con una confianza elevada a partir de cada itemset frecuente, donde cada regla es una partición binaria del itemset.

Nota: La generación de itemsets frecuentes sigue siendo computacionalmente costosa.

Itemsets frecuentes



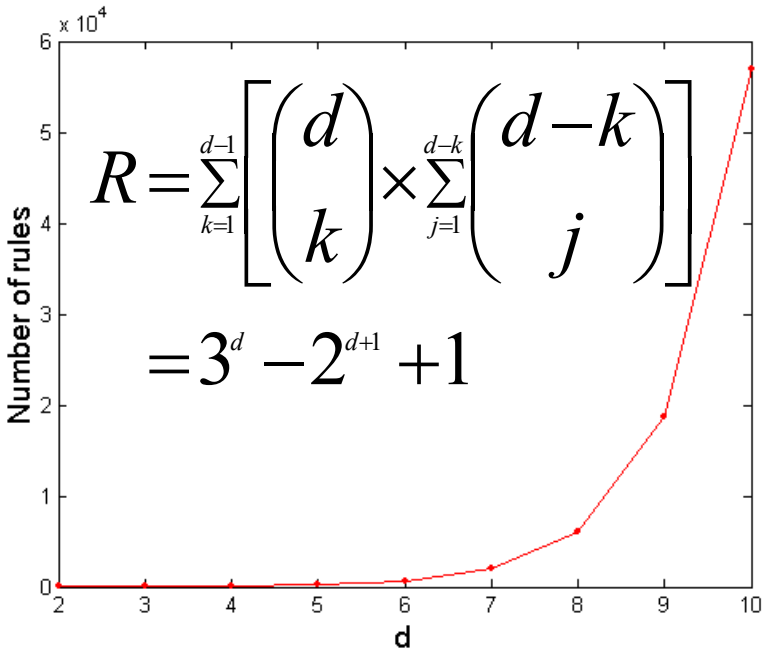
Itemsets frecuentes

Complejidad computacional

Dados d items, tenemos 2^d itemsets y R posibles reglas:

$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$

$$= 3^d - 2^{d+1} + 1$$



Itemsets frecuentes

Solución por fuerza bruta

Cada itemset del retículo es un candidato a ser frecuente:

- Contabilizar el soporte de cada candidato recorriendo la base de datos y emparejando cada transacción con cada posible candidato.

- Si tenemos N transacciones de W items (en media) y M candidatos, la complejidad del algoritmo resultante es de orden O(NMW)

=> **Muy costoso, ya que $M=2^d$!!!**
siendo d el número de items diferentes

Itemsets frecuentes

Estrategias

- **Reducir el número de candidatos (M)**
  Uso de técnicas de poda.
  Ejemplo: Algoritmos Apriori y DHP [Direct Hashing and Pruning]

- **Reducir el número de transacciones (N)**
  Reducir N conforme aumenta el tamaño del itemset.
  Ejemplo: Algoritmos AprioriTID y Eclat.

- **Reducir el número de comparaciones (NM)**
  Uso de estructuras de datos eficientes para almacenar los candidatos o las transacciones, de forma que no haya que comparar cada candidato con todas las transacciones.

Itemsets frecuentes

Reducción del número de candidatos

La propiedad Apriori

Si un itemset es frecuente,
también lo son todos sus subconjuntos

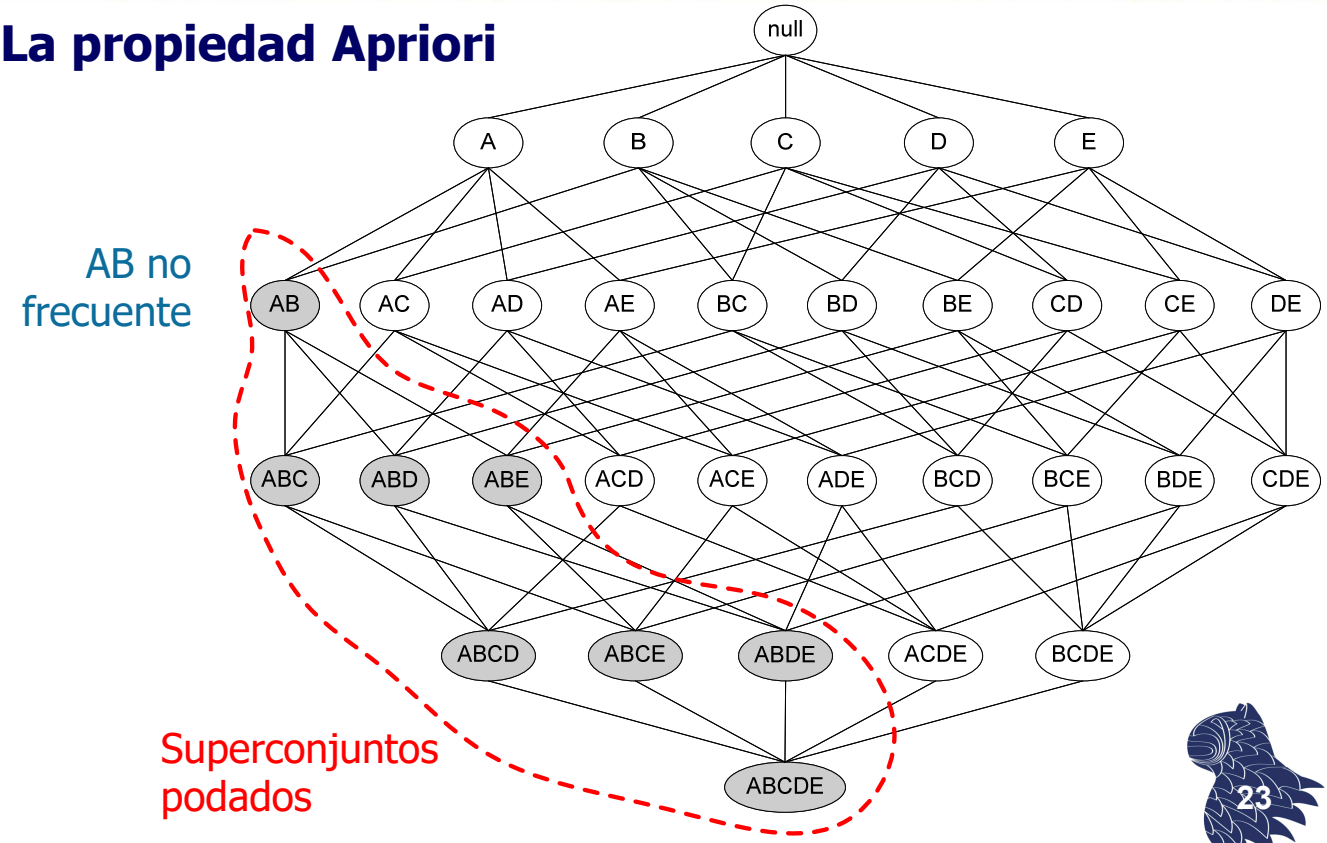
¿Por qué? Porque el soporte de un itemset nunca puede ser mayor que el de cualquiera de sus subconjuntos:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

Formalmente, esta propiedad se conoce con el nombre de anti-monotonía del soporte.

Itemsets frecuentes

La propiedad Apriori



Algoritmo Apriori

Tablas

L[k] = Conjunto de k-itemsets frecuentes

C[k] = Conjunto de k-itemsets potencialmente frecuentes.

Algoritmo

Generar L[1] (patrones frecuentes de tamaño 1, i.e. items)

Repetir mientras se descubran nuevos itemsets frecuentes:

a) Generar los candidatos C[k+1]
a partir de los patrones frecuentes L[k].

b) Contabilizar el soporte de cada candidato de C[k+1]
recorriendo la base de datos secencialmente.

c) Eliminar candidatos no frecuentes,
dejando en L[k+1] sólo aquéllos que son frecuentes.

Algoritmo Apriori

Item	supp
pan	4
vino	2
leche	4
cerveza	3
pañales	4
huevos	1

L1

MinSupp = 3

a) No hace falta generar candidatos
en los que intervengan items no frecuentes

C2

Itemset
{pan,leche}
{pan,cerveza}
{pan,pañales}
{leche,cerveza}
{leche,pañales}
{cerveza,pañales}

b) Conteo

Itemset	supp
{pan,leche}	3
{pan,cerveza}	2
{pan,pañales}	3
{leche,cerveza}	2
{leche,pañales}	3
{cerveza,pañales}	3

c) Filtrado de patrones
no frecuentes

L2

Itemset	supp
{pan,leche}	3
{pan,pañales}	3
{leche,pañales}	3
{cerveza,pañales}	3

a) No hay que generar
candidatos a partir de
{cerveza, pañales}

C3

Itemset
{pan,leche,pañales}

b) Conteo
c) Filtrado

L3

Itemset	supp
{pan,leche,pañales}	3

Generación de reglas

Dado un itemset frecuente L, se encuentran todos los subconjuntos no vacíos $f \subset L$ tales que $f \rightarrow L - f$ satisfaga el umbral de confianza mínima (MinConf).

Ejemplo

A partir del itemset frecuente {A,B,C,D}, se generan las siguiente reglas candidatas:

ABC →D,	ABD →C,	ACD →B,	BCD →A,
A →BCD,	B →ACD,	C →ABD,	D →ABC
AB →CD,	AC → BD,	AD → BC,	BC →AD,
BD →AC,	CD →AB,		

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

Generación de reglas

¿Cómo generar las reglas de forma eficiente?

- ¿Es la confianza anti-monótona como el soporte?
  NO: La confianza de ABC →D puede ser mayor o menor que la confianza de AB →D.

- Pero la confianza de las reglas generadas de un mismo itemset tienen una propiedad antimonótona:

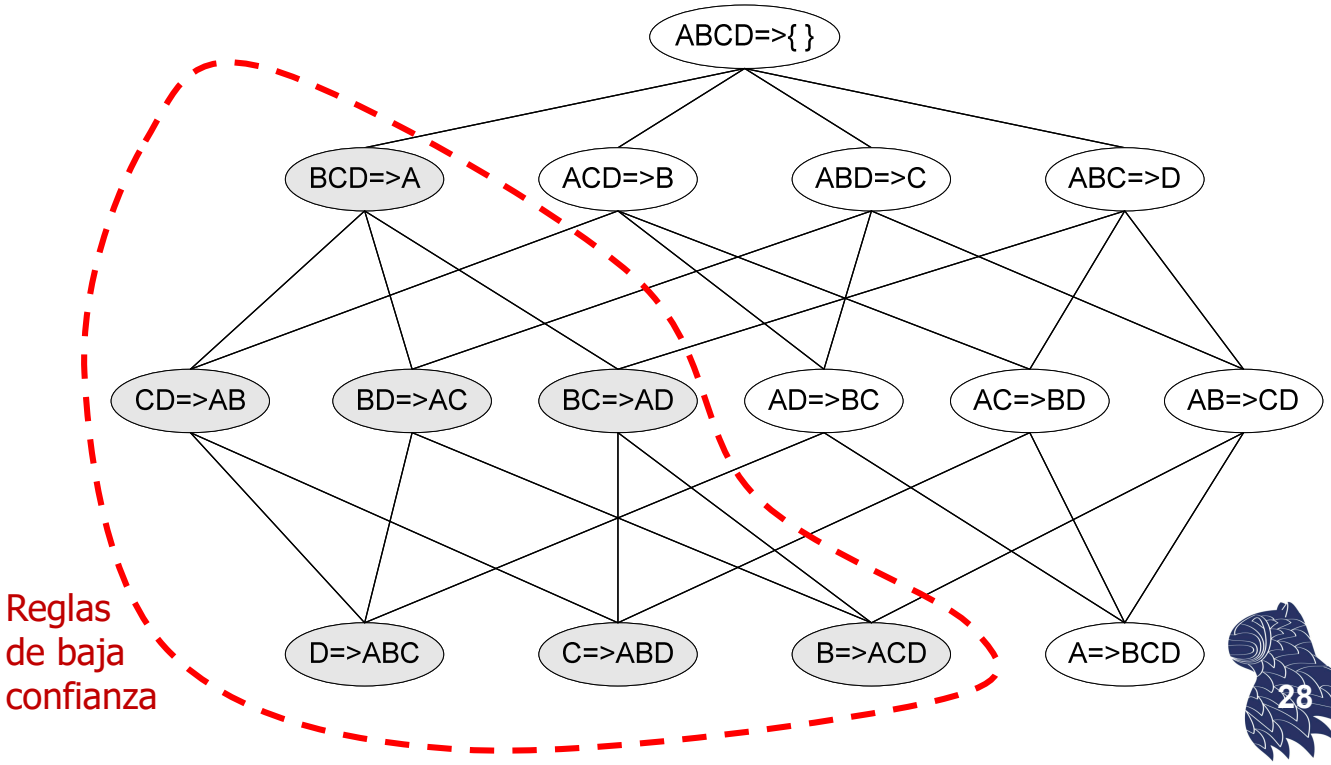
  p.ej. L = {A,B,C,D}

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

- La confianza es antimonótona con respecto al número de items en la parte derecha de la regla.

Generación de reglas

¿Cómo generar las reglas de forma eficiente?



Resumen

El proceso de extracción de reglas de asociación…

- Método descriptivo
  (si bien puede adaptarse como modelo predictivo).
- Para datos en formato transaccional (o relacional).
- Principalmente, para datos de tipo nominal (los atributos numéricos deben discretizarse previamente).
- Búsqueda exhaustiva (¡¡¡bastante lenta!!!).

Resumen

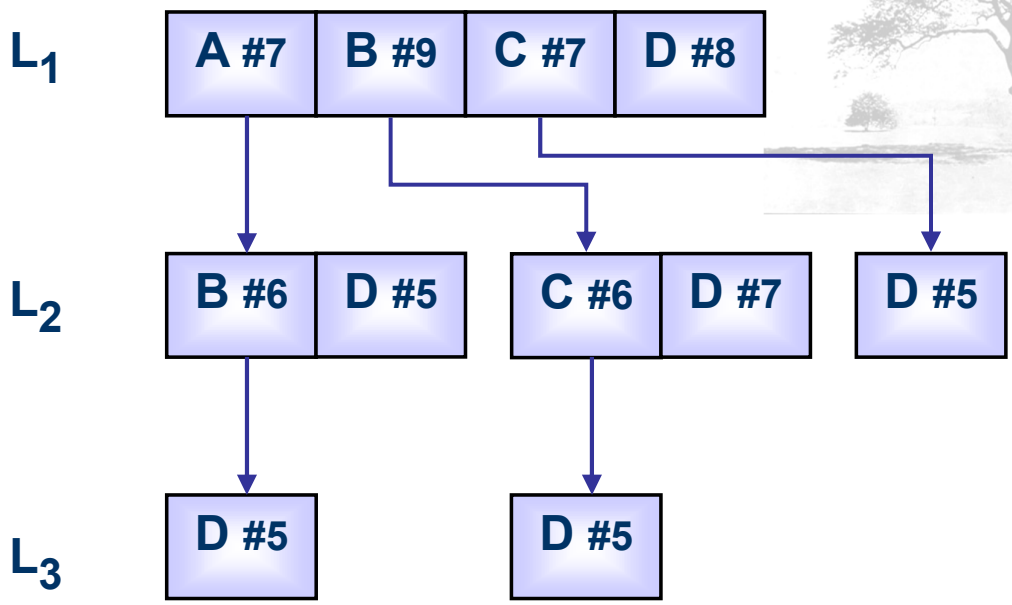
El proceso de extracción de reglas de asociación…

Solución en dos etapas
basada en el uso de umbrales de soporte y confianza:

- Identificación de los itemsets frecuentes (estrategias de poda usando el soporte).
- Obtención de reglas de asociación (estrategias de poda usando la confianza).

Optimizaciones

Árbol de enumeración de subconjuntos



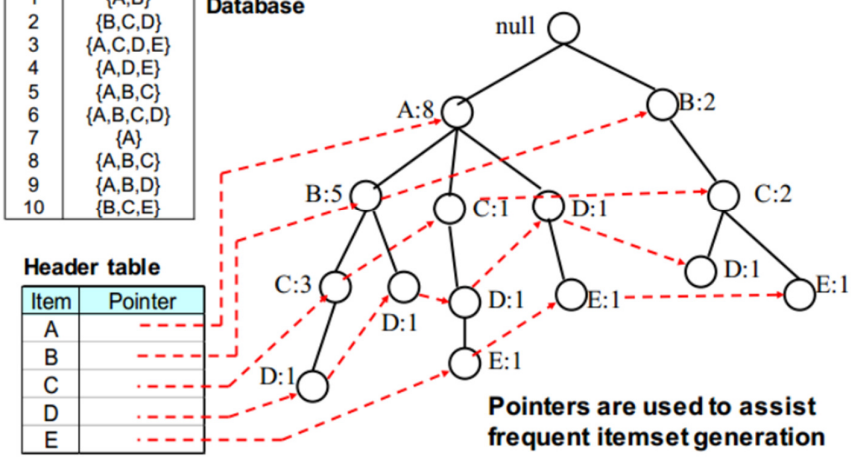
Optimizaciones

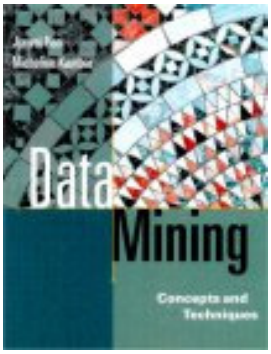
FP-Growth [Frequent Pattern Growth]

- Alternativa a los algoritmos basados en Apriori.
- Comprime la base de datos en un árbol (FP-Tree):

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{A}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Transaction Database



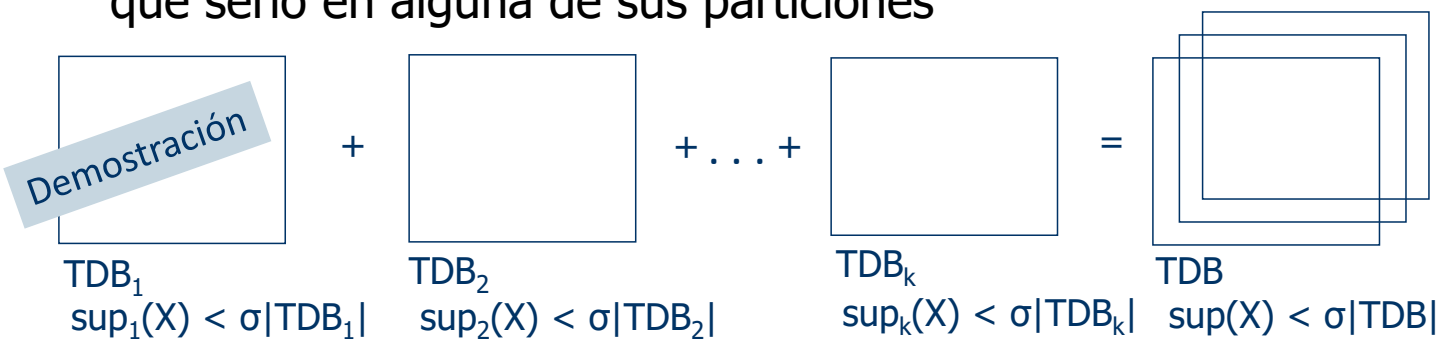


Han, Pei & Yin:
"Mining Frequent Patterns without Candidate Generation", SIGMOD'2000

Optimizaciones

Particionamiento

Cualquier patrón frecuente en una base de datos tiene que serlo en alguna de sus particiones



TDB_1 $sup_1(X) < \sigma|TDB_1|$

TDB_2 $sup_2(X) < \sigma|TDB_2|$

TDB_k $sup_k(X) < \sigma|TDB_k|$

TDB $sup(X) < \sigma|TDB|$

ALGORITMO

1. Se particiona la base de datos y se encuentran patrones frecuentes en las particiones.
2. Se consolidan los patrones frecuentes globales.

Optimizaciones

Un patrón frecuente de tamaño k contiene un número exponencial de subpatrones que también son frecuentes (2^k-1). ¿Cómo reducimos el número de patrones que hemos de mantener?

- Compresión sin pérdidas: **Patrones cerrados** C [closed patterns], cuando no existe un superconjunto S de C tal que support(C)=support(S).
- Compresión con pérdidas: **Patrones maximales** M [maximal (frequent) patterns], cuando M es frecuente y no existe un superconjunto S de M que sea frecuente.

Optimizaciones

Uso de restricciones [constraint-based mining]

En la práctica, la extracción de patrones es un proceso iterativo guiado por el usuario, que puede definir varios tipos de restricciones que se pueden incorporar en el algoritmo de extracción de patrones.

EJEMPLO

Una restricción c es antimonótona si, cuando un itemset S la viola, todos sus superconjuntos también lo hacen (se puede podar como en Apriori)

e.g. suma(S.precio)≤X, máximo(S.beneficio)≤X…

Optimizaciones

Uso de restricciones [constraint-based mining]

Pattern space pruning constraints	Data space pruning constraints
Anti-monotonic: If constraint c is violated, its further mining can be terminated.	**Data anti-monotonic:** If a transaction t does not satisfy c, then t can be pruned to reduce data processing effort.
Monotonic: If c is satisfied, no need to check c again.	
Convertible: c can be converted to monotonic or anti-monotonic if items can be properly ordered in processing.	
Succinct: If the constraint c can be enforced by directly manipulating the data.	**Data succinct:** Data space can be pruned at the initial pattern mining process.

Visualización de reglas

Es habitual que el número de reglas de asociación generadas sea excesivo, por lo que hay que filtrarlas:

Solución guiada por el usuario

- El usuario establece las reglas/atributos/valores que considera interesantes.
- El sistema las compara con las obtenidas.
- El usuario las visualiza y selecciona.

Sin ayuda del usuario

Se establece un ranking del grado de interés de cada regla (con medidas de calidad más restrictivas que la confianza).

Visualización de reglas

Técnicas de visualización integradas en herramientas de minería de datos para facilitar la interpretación de los resultados:

- Técnicas basadas en **tablas**
  p.ej. SAS Enterprise Miner, DBMiner…
- Técnicas basadas en **matrices 2D**
  p.ej SGI MineSet, DBMiner…
- Técnicas basadas en **grafos**
  p.ej. DBMiner ball graphs…
- Técnicas basadas en **coordenadas paralelas**
  p.ej. VisAR, TMiner…

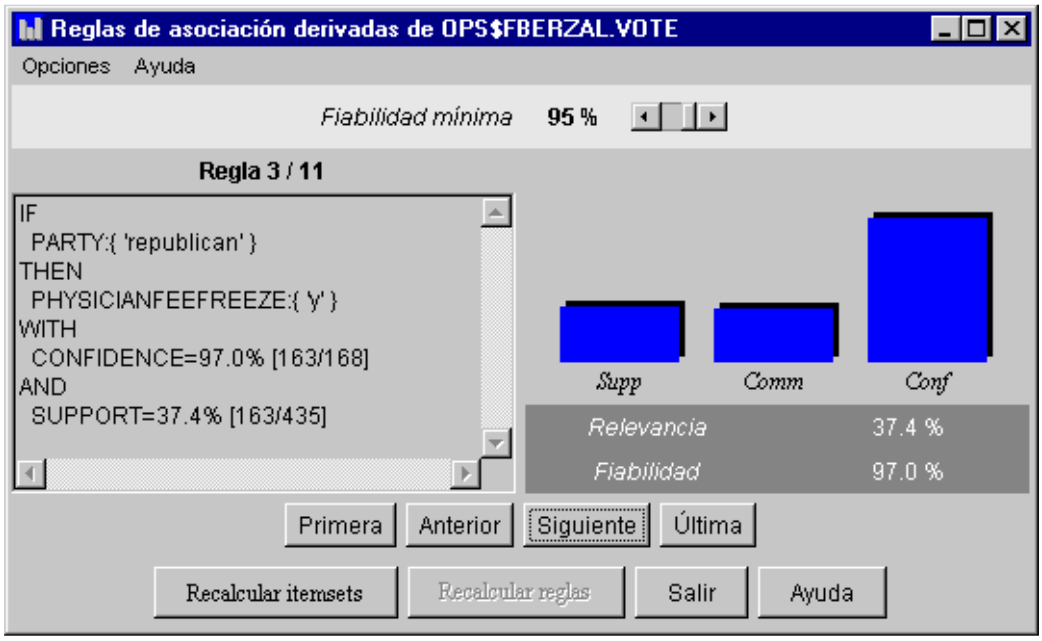
Visualización de reglas

	Body	Implies	Head	Supp (%)	Conf (%)
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) = '0.00~100.00'	==>	revenue(x) = '500.00~1000.00'	19.67	33.23

Sheet1

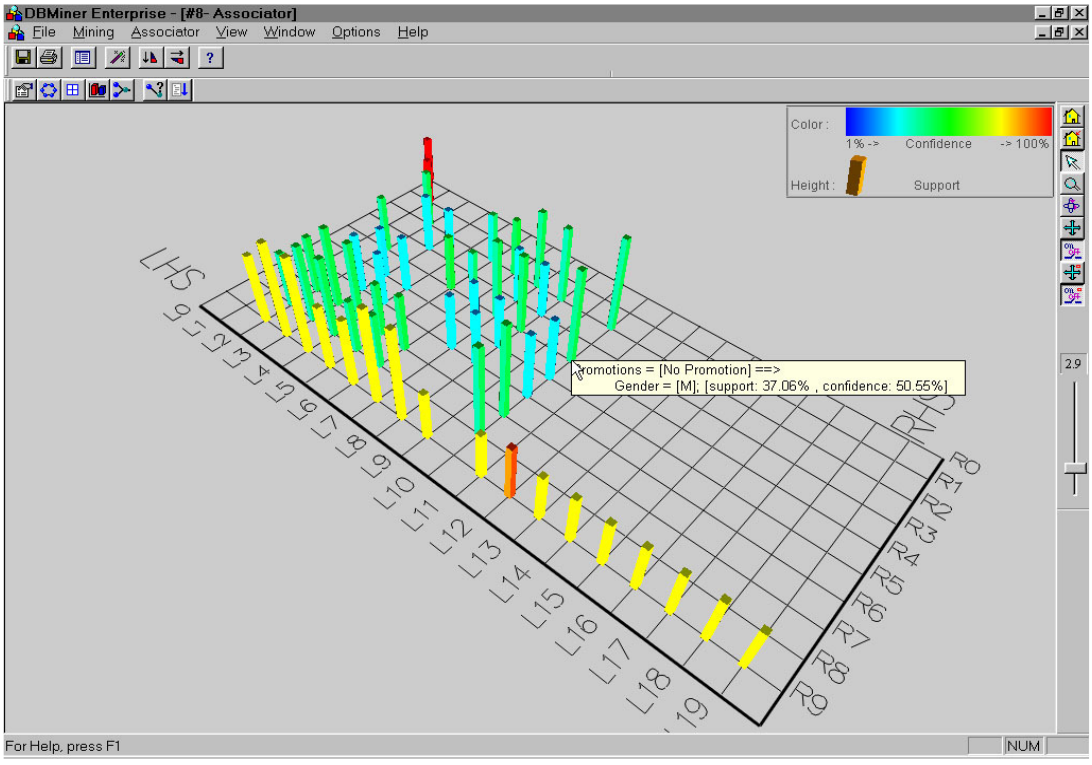
Representación tabular de un conjunto de reglas :-(

Visualización de reglas



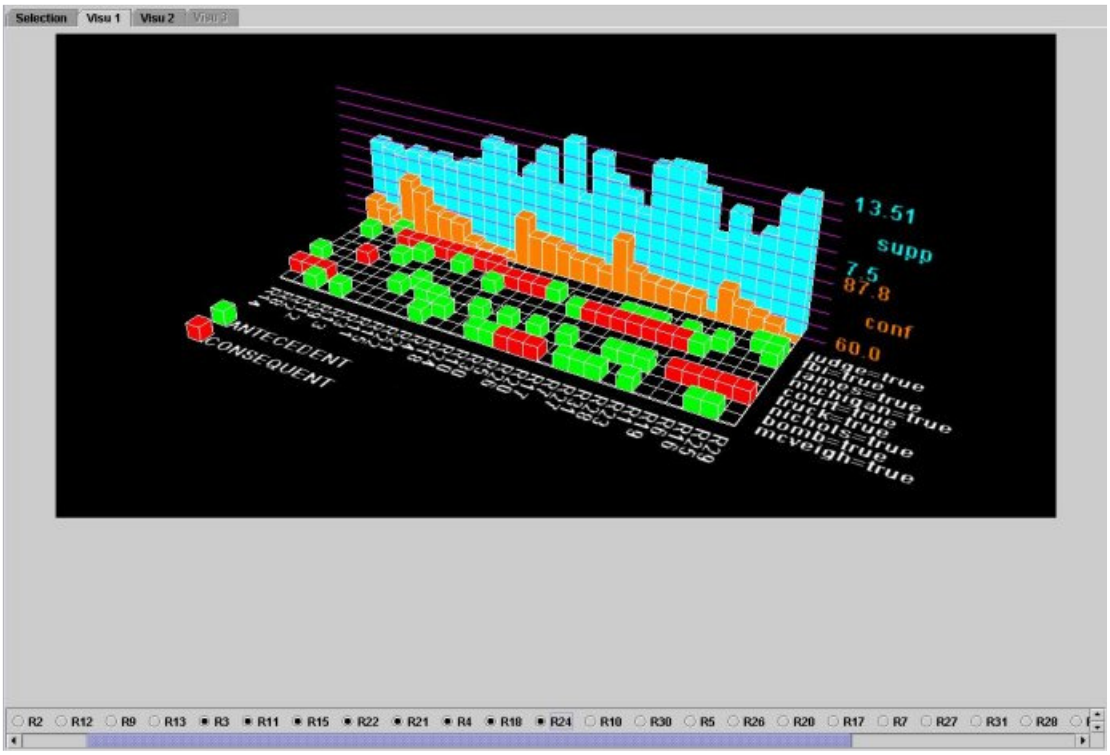
Uso de ayudas visuales :-|

Visualización de reglas



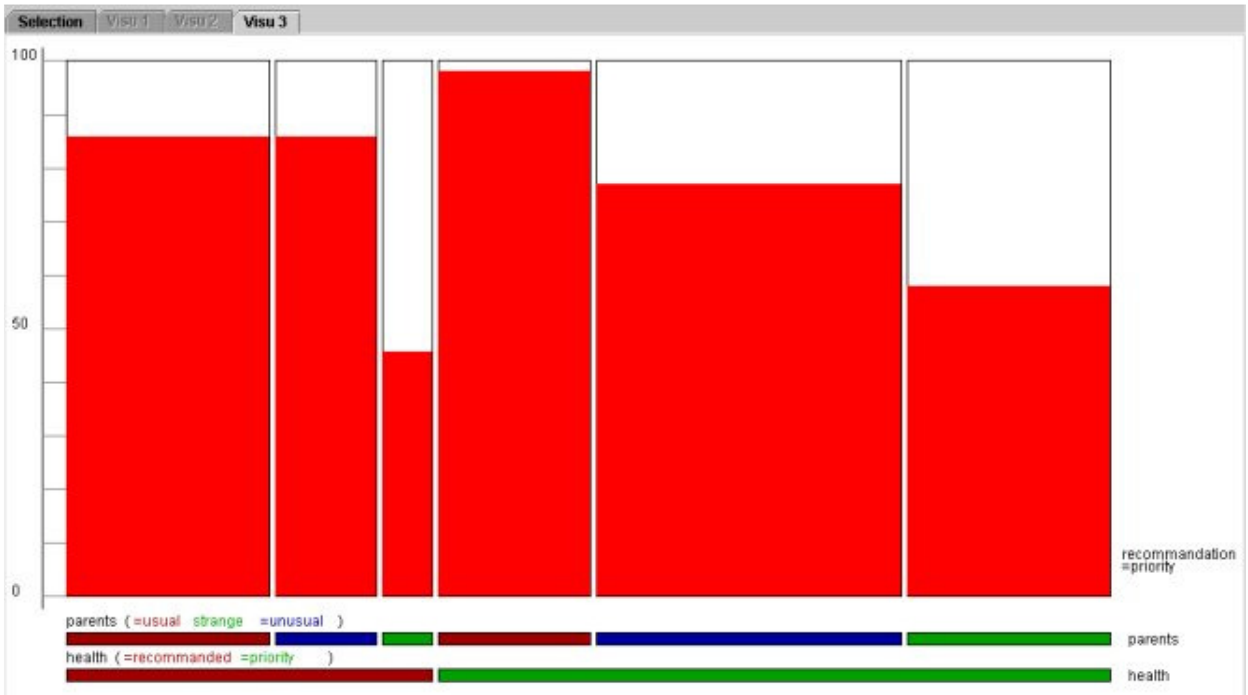
Representación 2D (DBMiner)

Visualización de reglas



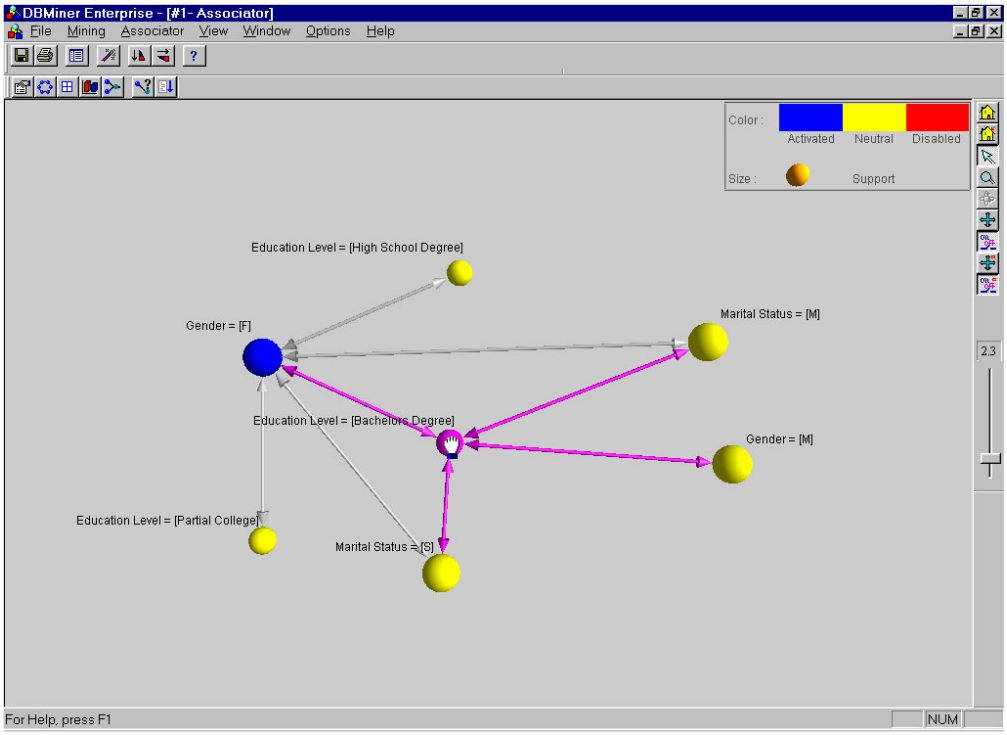
http://www2.lifl.fr/~jourdan/download/arv.html

Visualización de reglas



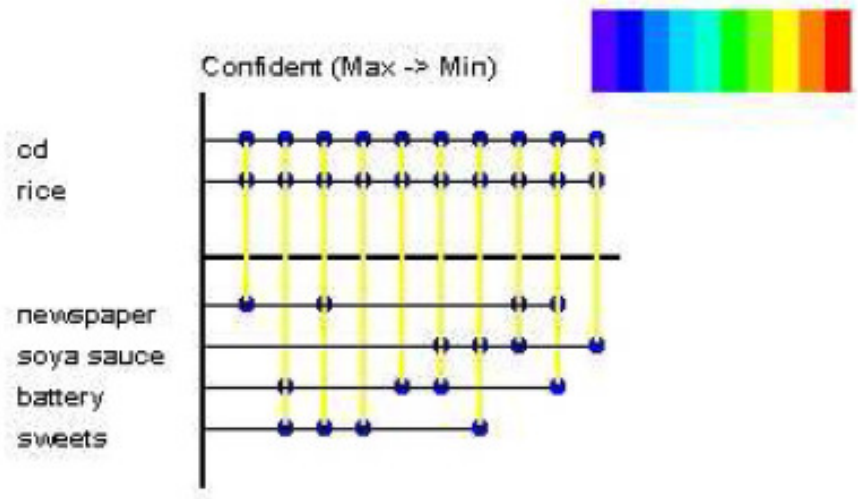
http://www2.lifl.fr/~jourdan/download/arv.html

Visualización de reglas



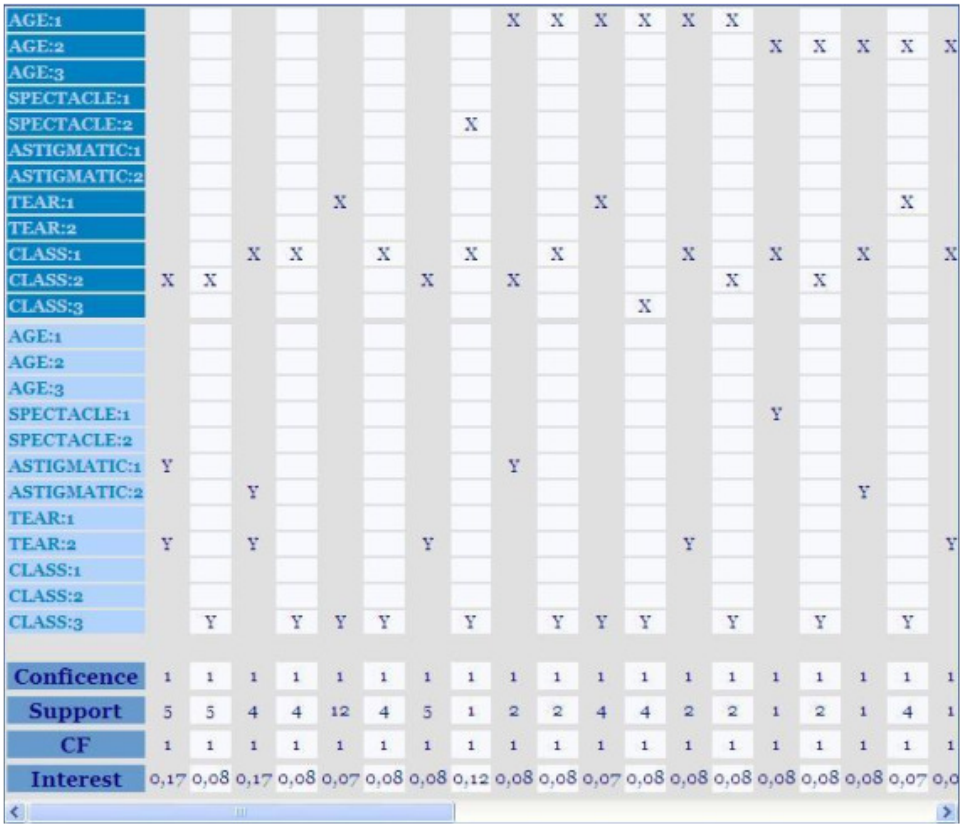
Visualización basada en grafos (DBMiner)

Visualización de reglas



VisAR (coordenadas paralelas)

Visualización de reglas



AGE:1								X	X	X	X	X	X					
AGE:2														X	X	X	X	X
AGE:3																		
SPECTACLE:1																		
SPECTACLE:2							X											
ASTIGMATIC:1																		
ASTIGMATIC:2																		
TEAR:1					X					X							X	
TEAR:2																		
CLASS:1			X	X		X		X		X			X		X		X		X
CLASS:2	X	X					X		X					X		X		
CLASS:3												X						
AGE:1																		
AGE:2																		
AGE:3																		
SPECTACLE:1															Y			
SPECTACLE:2																		
ASTIGMATIC:1	Y								Y									
ASTIGMATIC:2			Y														Y	
TEAR:1																		
TEAR:2	Y		Y				Y						Y					Y
CLASS:1																		
CLASS:2																		
CLASS:3		Y		Y	Y	Y		Y		Y	Y	Y		Y		Y		Y	
Confidence	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Support	5	5	4	4	12	4	5	1	2	2	4	4	2	2	1	2	1	4	1
CF	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Interest	0,17	0,08	0,17	0,08	0,07	0,08	0,08	0,12	0,08	0,08	0,07	0,08	0,08	0,08	0,08	0,08	0,08	0,07	0,0

TMiner

Evaluación de reglas

La confianza no es la mejor medida de interés posible para las reglas de asociación.

p.ej.
Un item muy frecuente aparecerá a menudo en el consecuente de muchas reglas, independientemente de su relación con el antecedente de la regla.

Evaluación de reglas

La confianza no es la mejor medida de interés posible para las reglas de asociación.

Dada la siguiente tabla de contingencia

	Café	noCafé	
Té	15	5	20
noTé	75	5	80
	90	10	100

La regla Té → Café tiene una confianza del 75% pese a que el hecho de comprar té disminuye la probabilidad inicial de comprar café (90% global).

Evaluación de reglas

- Las técnicas de extracción de reglas de asociación tienden a producir demasiadas reglas
  - Muchas de ellas son redundantes.
    {A,B}→{C} and {A}→{C} con el mismo soporte y confianza.
  - Muchas de ellas no son interesantes
    La regla Té → Café del ejemplo anterior.
- Se pueden definir medidas de interés alternativas que nos ayuden a podar/ordenar el conjunto de reglas obtenido…

Evaluación de reglas

Propiedades deseables de las medidas de interés I para reglas de asociación [Piatetsky-Shapiro, 1991]:

P1 I(A⇒C) = 0
cuando supp(A⇒C) = supp(A)supp(C).

P2 I(A⇒C) crece monótonamente con supp(A⇒C).

P3 I(A⇒C) decrece monótonamente con supp(A) (o supp(C)).

Evaluación de reglas

Propiedades deseables de las medidas de interés I para reglas de asociación [Piatetsky-Shapiro, 1991]:

P1 I(A⇒C) = 0
cuando supp(A⇒C) = supp(A)supp(C).

La confianza no verifica esta propiedad

P2 I(A⇒C) crece monótonamente con supp(A⇒C).

OK

P3 I(A⇒C) decrece monótonamente con supp(A) (o supp(C)).

La confíanza no la verifica para supp(C)

Evaluación de reglas

Principio clásico

Cuanto mayor sea el soporte, mejor es el itemset.

Problema

Dado un itemset C con soporte elevado,
cualquier otro itemset A parece un buen "predictor" de C.

- conf(A⇒C) > supp(C)
  OK
- conf(A⇒C) ≤ supp(C)
  Dependencia negativa o independencia
  (debería descartarse la regla)

Tendríamos que comprobar el soporte del consecuente...

Evaluación de reglas

Una medida de interés alternativa:

Lift / interés / implicación

$$Lift = Interest = \frac{conf(X \rightarrow Y)}{supp(Y)} = \frac{P(Y|X)}{P(Y)} = \frac{P(X,Y)}{P(X)P(Y)}$$

	Café	noCafé	
Té	15	5	20
noTé	75	5	80
	90	10	100

Regla **Té → Café**

- Confianza
  = P(Café|Té) = 0.75
- Lift = P(Café|Té) / P(Café) = 0.75/0.9= 0.8333
  (< 1, asociación negativa)

Evaluación de reglas

Un inconveniente del lift…

	Y	No Y	
X	10	0	10
no X	0	90	90
	10	90	100

	Y	No Y	
X	90	0	90
no X	0	10	10
	90	10	100

$$Lift = \frac{0.1}{(0.1)(0.1)} = 10$$

$$Lift = \frac{0.9}{(0.9)(0.9)} = 1.11$$

Valor no acotado.

Efecto de las transacciones nulas (sin X ni Y).

Independencia estadística: P(X,Y)=P(X)P(Y), lift=1

Evaluación de reglas

Otro inconveniente del lift…

lift (té → café) = lift (café → té)

(medida de interés simétrica)

Otra medida de interés alternativa:

Factores de certeza

[CF: Certainty factors]

- … satisfacen las propiedades de Piatetsky-Shapiro
- … se utilizan a menudo en sistemas expertos
- … no son simétricos (como el lift)

Evaluación de reglas

Factores de certeza [CF]

$$CF(X \rightarrow Y) = \begin{cases} \dfrac{conf(X \rightarrow Y) - conf(Y)}{1 - supp(Y)} & si\ conf(X \rightarrow Y) > supp(Y) \\ \dfrac{conf(X \rightarrow Y) - conf(Y)}{supp(Y)} & si\ conf(X \rightarrow Y) < supp(Y) \\ 0 & en\ otro\ caso \end{cases}$$

- Valores acotados entre -1 y +1.
- Medida de interés no simétrica.
- Verifica propiedades deseables (incluyendo PS) p.ej. $CF(X \rightarrow \neg Y) = -CF(X \rightarrow Y)$

Evaluación de reglas

Existen muchas más medidas de interés 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)	$\sqrt{P(A,B)}\max(P(B\|A) - P(B), P(A\|B) - P(A))$

Evaluación de reglas

... con diferentes conjuntos de propiedades:

Symbol	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

Evaluación de reglas

Ejemplo de propiedad interesante: **Null-Invariance**
Invarianza con respecto al número de transacciones nulas.

Measure	Definition	Range	Null-Invariant?
$\chi^2(A,B)$	$\sum_{i,j}\frac{(e(a_i,b_j)-o(a_i,b_j))^2}{e(a_i,b_j)}$	$[0, \infty]$	No
$Lift(A,B)$	$\frac{s(A\cup B)}{s(A)\times s(B)}$	$[0, \infty]$	No
$Allconf(A,B)$	$\frac{s(A\cup B)}{max\{s(A),s(B)\}}$	$[0, 1]$	Yes
$Jaccard(A,B)$	$\frac{s(A\cup B)}{s(A)+s(B)-s(A\cup B)}$	$[0, 1]$	Yes
$Cosine(A,B)$	$\frac{s(A\cup B)}{\sqrt{s(A)\times s(B)}}$	$[0, 1]$	Yes
$Kulczynski(A,B)$	$\frac{1}{2}(\frac{s(A\cup B)}{s(A)}+\frac{s(A\cup B)}{s(B)})$	$[0, 1]$	Yes
$MaxConf(A,B)$	$max\{\frac{s(A\cup B)}{s(A)},\frac{s(A\cup B)}{s(B)}\}$	$[0, 1]$	Yes

Esencialmente, agregaciones de p y q (min, max, mean)

$$p = \frac{s(A \cup B)}{s(A)} = P(B|A) \qquad q = \frac{s(A \cup B)}{s(B)} = P(A|B)$$

p y q son invariantes frente a transacciones nulas.

Evaluación de reglas

Medidas de interés alternativas…

- Algunas son buenas para ciertas aplicaciones, pero no para otras.
- Algunas poseen ciertas propiedades, otras no (lo que puede afectar a la eficiencia del algoritmo de extracción de reglas de asociación)

Extensiones y variaciones

- Reglas de asociación **cuantitativas** (atributos continuos)
- Reglas **multinivel** (a.k.a. reglas de asociación generalizadas)
- Variaciones en función del **tipo de patrones**:
  - Itemsets frecuentes (bases de datos transaccionales y relacionales)
  - Análisis de secuencias (secuencias, p.ej. Bioinformática, y series temporales)
  - Análisis de estructuras (datos estructurados, p.ej. grafos)

Atributos continuos

Reglas de asociación cuantitativas

Distintos tipos de reglas con atributos continuos

$$\text{Edad} \in [21,35) \wedge \text{Salario} \in [40k,60k) \rightarrow \text{Compra}$$

$$\text{Salario} \in [30k,60k) \wedge \text{Compra} \rightarrow \text{Edad: } \mu=28,\ \sigma=4$$

Métodos…

- … basados en técnicas de discretización
- … basados en técnicas estadísticas

Atributos continuos

Discretización

El tamaño de los intervalos afecta
al soporte y a la confianza de las reglas…

$$\{\text{Refund} = \text{No}, (\text{Income} = \$51{,}250)\} \rightarrow \{\text{Cheat} = \text{No}\}$$

$$\{\text{Refund} = \text{No}, (60K \leq \text{Income} \leq 80K)\} \rightarrow \{\text{Cheat} = \text{Yes}\}$$!!!

$$\{\text{Refund} = \text{No}, (0K \leq \text{Income} \leq 1B)\} \rightarrow \{\text{Cheat} = \text{No}\}$$

- Si los intervalos son muy pequeños…
  … las reglas pueden no tener suficiente soporte.

- Si los intervalos son demasiado grandes…
  … las reglas pueden no tener confianza suficiente.

Atributos continuos

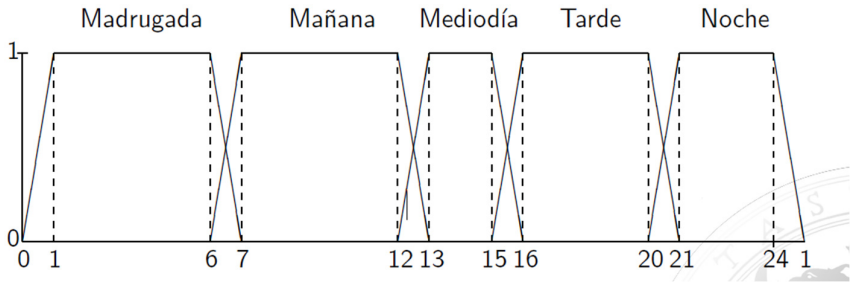
Reglas de asociación difusas

Intervalos difusos de los atributos continuos caracterizados por etiquetas lingüísticas…

Base de datos de urgencias hospitalarias

[CLASISTENCIA=YESOS] → [HINGRESO=TARDE]
supp=0.022 C=0.48

[HINGRESO=MAÑANA] → [CLASISTENCIA=OBSERVACIÓN]
supp=0.128 CF=0.43



Atributos continuos

Fenómenos extraordinarios

LHS → RHS

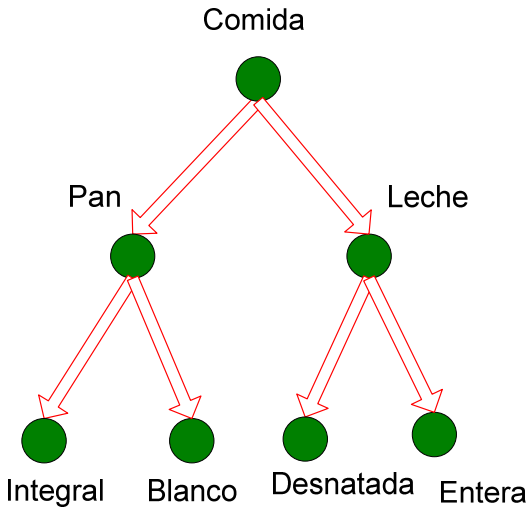
- LHS: Subconjunto de la población.
- RHS: Comportamiento extraordinario en ese subconjunto.

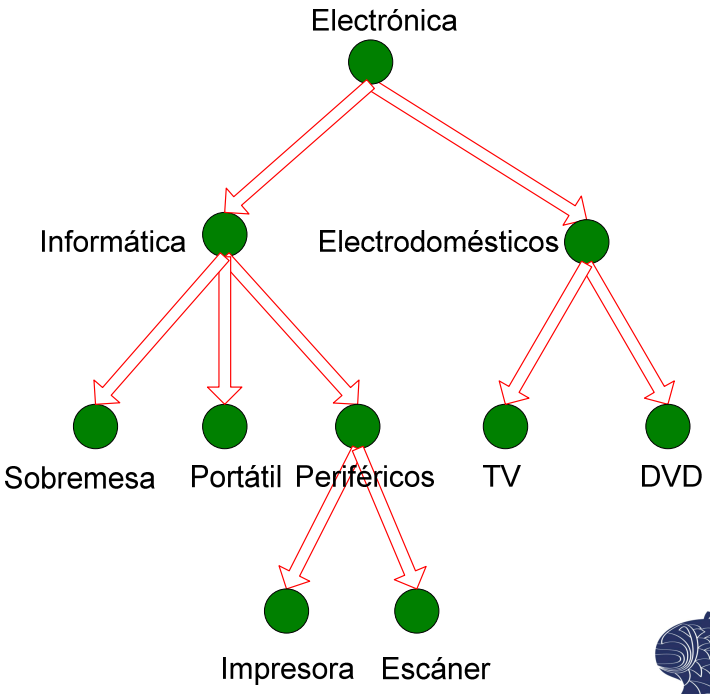
Estudios = graduado escolar → Sueldo medio = 50€/hora
[media global = 20€/hora]

La regla se acepta sólo si pasa un test estadístico (Z-test)

Yonatan Aumann & Yehuda Lindell: "A Statistical Theory for Quantitative Association Rules", KDD'99

Reglas multinivel





Reglas multinivel

¿Por qué utilizar jerarquías de conceptos?

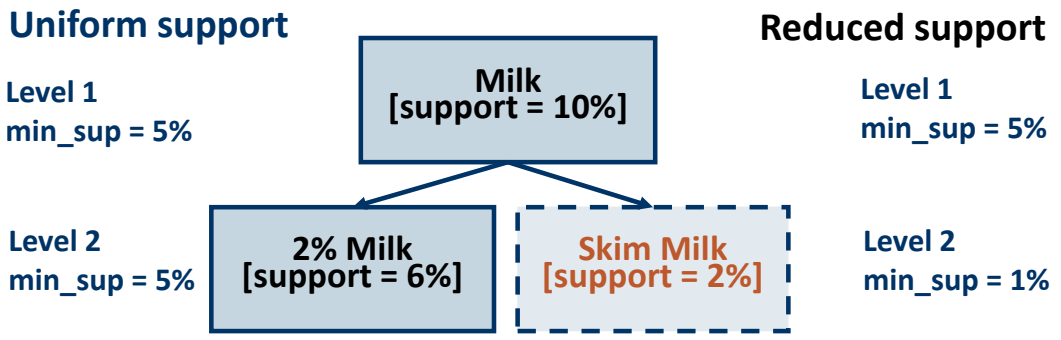
- Porque las reglas que involucran artículos en los niveles más bajos puede que no tengan soporte suficiente como para aparecer en algún patrón frecuente.

- Porque las reglas a niveles bajos de la jerarquía son demasiado específicas.

  p.ej. leche desnatada → pan blanco,
  leche entera → pan integral,
  leche desnatada → pan integral
  ...
  indican una asociación entre pan y leche.

Reglas multinivel

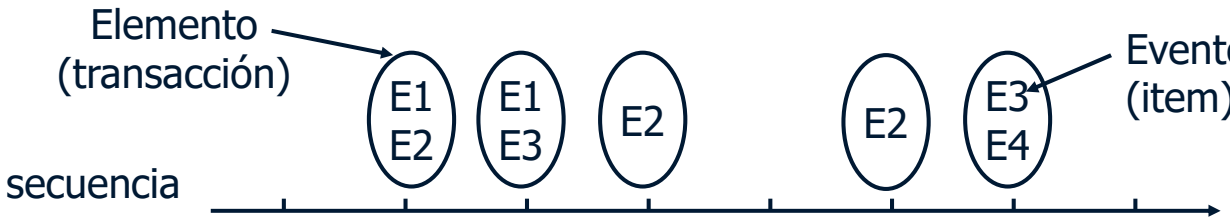
Estrategias



- Artículos en niveles bajos de la jerarquía se espera que tengan un menor soporte.
- "Shared multi-level mining": Algoritmos eficientes usan el nivel de soporte más bajo para extraer el conjunto de patrones candidatos y, a posteriori, filtran reglas redundantes.

Análisis de secuencias

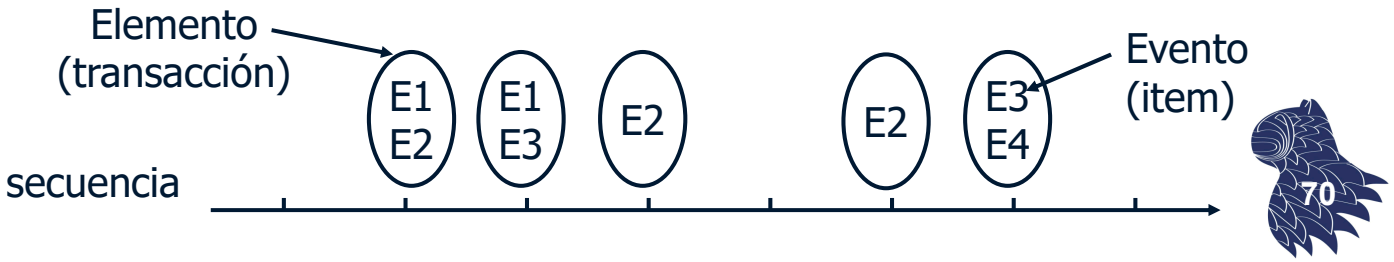
Base de datos	Secuencia	Elemento (Transacción)	Evento (Item)
Clientes	Historial de compras de un cliente determinado	Conjunto de artículos comprados por un cliente en un instante concreto	Libros, productos…
Web	Navegación de un visitante del sitio web	Colección de ficheros vistos por el visitante tras 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



Análisis de secuencias

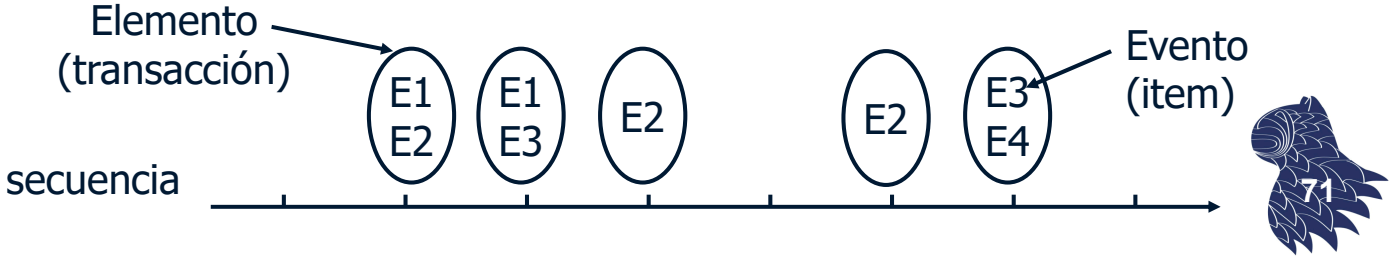
Una secuencia $<a_1\ a_2 \ldots a_n>$
está contenida en otra secuencia $<b_1\ b_2 \ldots b_m>$ $(m \geq n)$
si existe un conjunto de enteros $i_1 < i_2 < \ldots < i_n$
tales que $a_1 \subseteq b_{i1}$, $a_2 \subseteq b_{i1}$, …, $a_n \subseteq b_{in}$

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



Análisis de secuencias

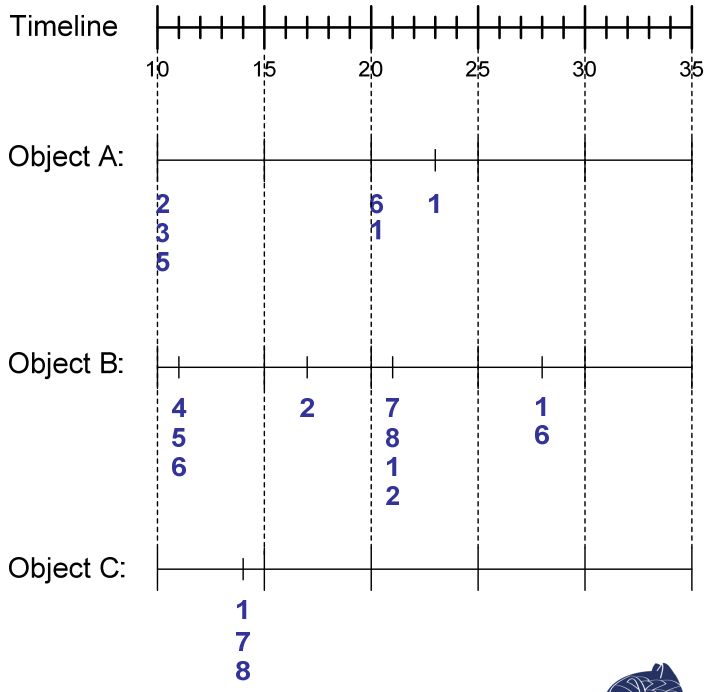
- El soporte de una subsecuencia S se define como la fracción de secuencias de la base de datos que incluyen la subsecuencia S.
- Un patrón secuencial es una subsecuencia frecuente (esto es, una subsecuencia con soporte ≥ MinSupp)



Análisis de secuencias

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



Análisis de secuencias

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%

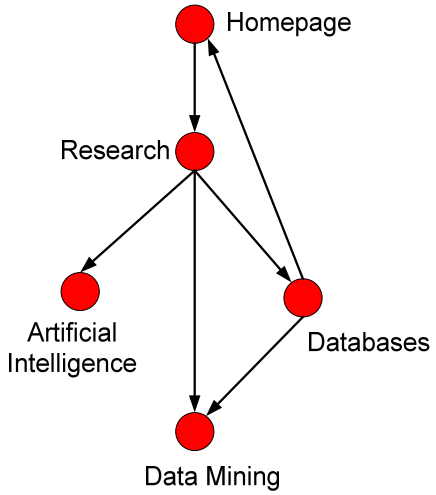
Ejemplos de subsecuencias frecuentes:

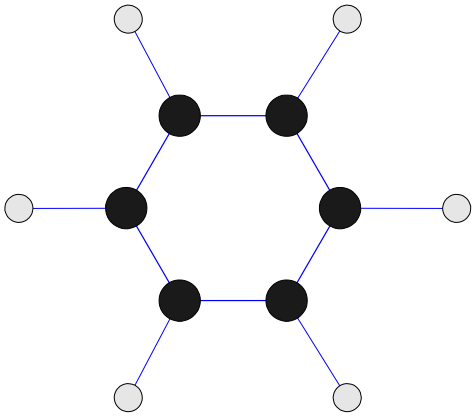
< {1,2} >	s=60%
< {2,3} >	s=60%
< {2,4}>	s=80%
< {3} {5}>	s=80%
< {1} {2} >	s=80%
< {2} {2} >	s=60%
< {1} {2,3} >	s=60%
< {2} {2,3} >	s=60%
< {1,2} {2,3} >	s=60%

Análisis de estructuras

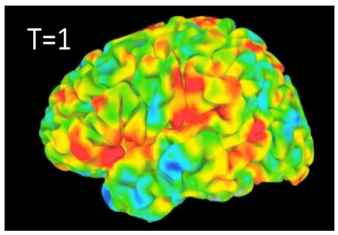
Identificación de patrones frecuentes en grafos

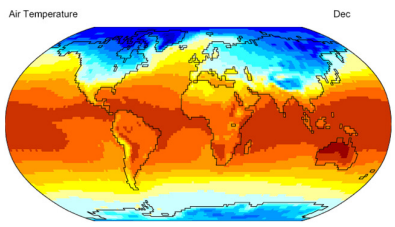
Aplicaciones: Web Mining, Bioinformática, redes sociales…

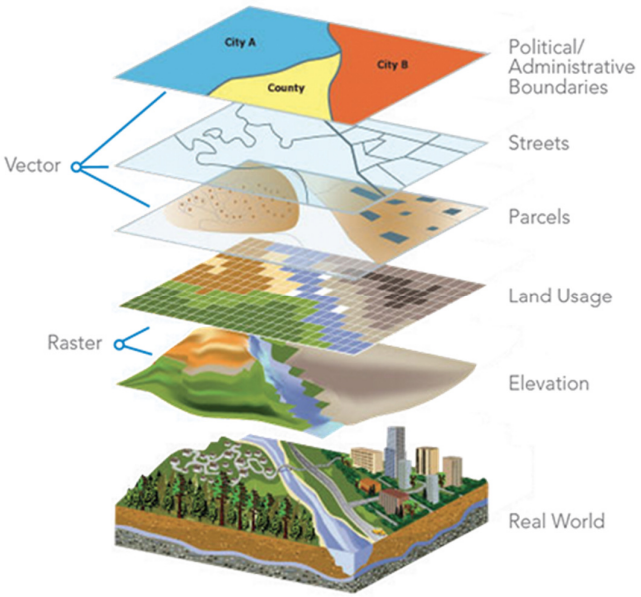




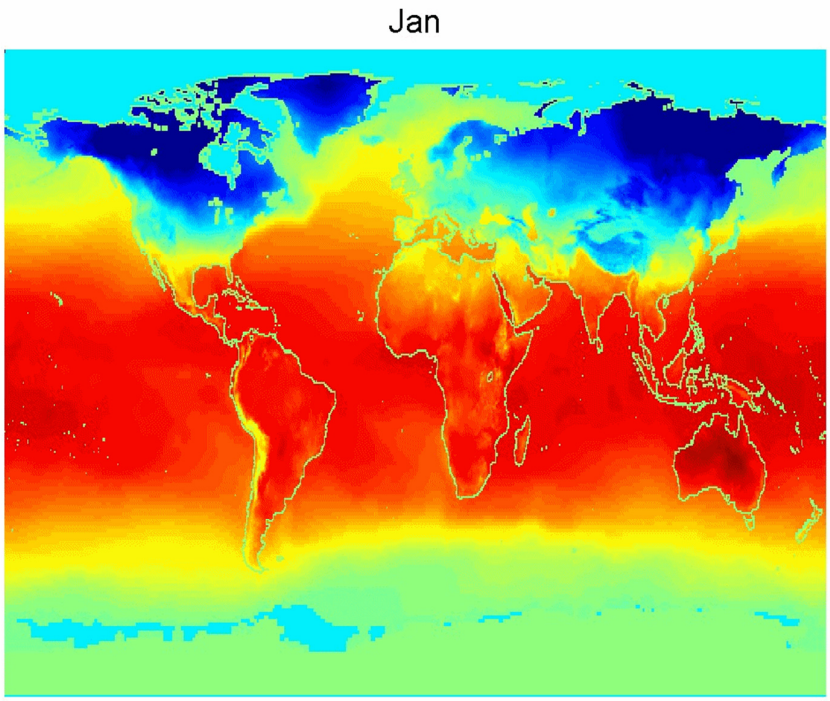
Datos espaciales







Datos espaciotemporales



Detección de anomalías

Regla de asociación anómala

Regla con una confianza elevada que representa una desviación homogénea del comportamiento habitual.

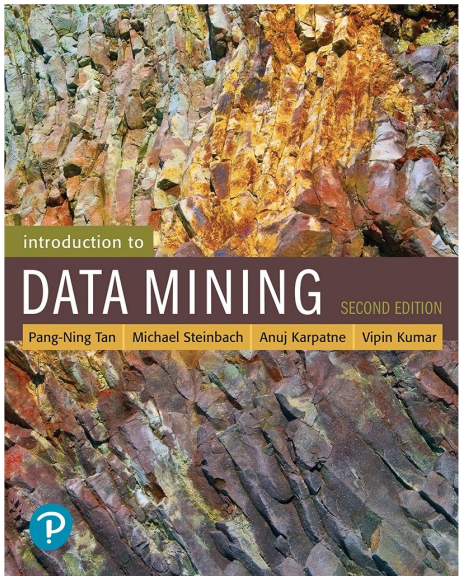
"Anomalía"

```
if WORKCLASS: Local-gov
then CAPGAIN: [99999.0, 99999.0]   (7 out of 7)
when not CAPGAIN: [0.0, 20051.0]
```

Consecuente habitual

Bibliografía

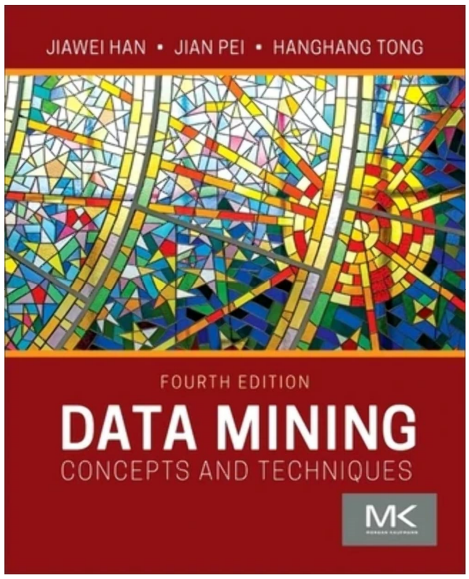
Pang-Ning Tan,
Michael Steinbach,
Vipin Kumar &
Anuj Karpatne:
Introduction to Data Mining,
2nd edition, Addison Wesley, 2018.
ISBN 0133128903



4 Association Analysis: Basic Concepts and Algorithms
7 Association Analysis: Advanced Concepts
10.4 Statistical Testing for Association Analysis

Bibliografía

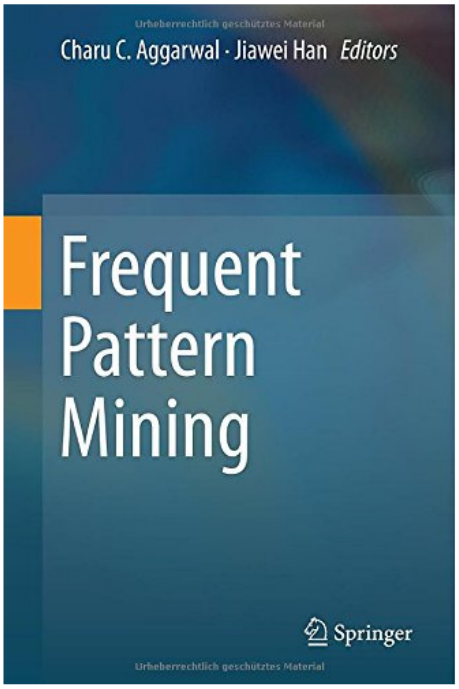
Jiawei Han,
Jian Pei &
Hanghang Tong:
Data Mining:
Concepts and Techniques,
4th edition, Morgan Kaufmann, 2022.
ISBN 0128117605



4 Pattern mining: Basic concepts and methods
5 Pattern mining: Advanced methods

Bibliografía

- Charu C. Aggarwal & Jiawei Han(editors): **Frequent Pattern Mining**. Springer, 2014. ISBN 3319078208.



Bibliografía

- Agrawal & Skirant: **Fast Algorithms for Mining Association Rules**, VLDB'94
- Park, Chen & Yu: **An Effective Hash-Based Algorithm for Mining Association Rules**, SIGMOD'95 (DHP)
- Toivonen: **Sampling Large Databases for Association Rules**, VLDB'96
- Park, Yu & Chen: **Mining Association Rules with Adjustable Accuracy**, CIKM'97
- Savasere, Omiecinski & Navathe: **An Efficient Algorithm for Mining Association Rules in Large Databases**, VLDB'95
- Brin, Motwani, Ullman & Tsur: **Dynamic Itemset Counting and Implication Rules for Market Basket Data**, SIGMOD'97 (DIC)
- Hidber: **Online Association Rule Mining**, SIGMOD'99 (CARMA)
- Berzal, Cubero, Sánchez & Serrano: **TBAR: An efficient method for association rule mining in relational databases**, Data & Knowledge Engineering, 2001
- Han, Pei & Yin: **Mining Frequent Patterns without Candidate Generation**, SIGMOD'2000 (FP-Growth)
- Berzal, Blanco, Sánchez & Vila: **Measuring the accuracy and interest of association rules: A new framework**, Intelligent Data Analysis, 2002
- Hilderman & Hamilton: **Evaluation of interestingness measures for ranking discovered knowledge**, PAKDD'2001
- Tan, Kumar & Srivastava: **Selecting the right objective measure for association analysis**. Information Systems, vol. 29, pp. 293-313, 2004