REGLAS DE ASOCIACIÓN: ASPECTOS AVANZADOS

Daniel Sánchez Fernández

Grupo: Bases de Datos y Sistemas de Información Inteligentes

Departamento de Ciencias de la Computación e I.A.

Universidad de Granada

Contenidos

- Problemas de Interpretabilidad
- Evaluación: Medidas de Interés
- Interpretaciones
- Reglas de Asociación Difusas
- Aspectos algorítmicos. El caso de las reglas de asociación jerárquicas
- Evaluación de reglas por grupos

Problemas de Interpretabilidad

- Problemas derivados de los datos
- Problemas derivados de los usuarios
- Problemas derivados de las medidas

Problemas derivados de los datos

- Las reglas de asociación, como hemos visto, establecen asociaciones de tipo implicación (regla) entre la presencia conjunta de ítems en transacciones.
- Expresan tendencias.
- En ocasiones la semántica de estas asociaciones se malinterpreta de diversas formas, lo cual repercute negativamente en el uso que podemos hacer del conocimiento obtenido.
- En este apartado veremos varios aspectos que pueden dar lugar a una interpretación errónea de las reglas.

Problemas derivados de los datos

- Si los datos no son apropiados, las reglas pueden ser dudosas e inútiles.

- Algunos problemas:
  - Falta de variabilidad: items muy poco frecuentes o demasiado frecuentes.
  - Representatividad: los datos deben comprender todos los casos que se pretenden estudiar con un número suficiente de casos.
  - Sesgos muestrales.
  - Factores ocultos: estacionalidad, items no considerados.
  - Valores perdidos en los datos.

- Es necesario estudiar y preparar previamente los datos.

Problemas derivados del usuario

- No disponibilidad de expertos en los datos para la valoración de las reglas
- Confusión de semánticas (dependencia simétrica, implicación, causalidad)
  - Confusión con dependencias simétricas (por influencia de la estadística).
  - Diversos tipos de implicación lógica. Las reglas de asociación con medida de confianza expresan tendencias en la presencia conjunta de items.
  - Causalidad. Tendencia no implica necesariamente causalidad (ej: pañales → cerveza).

$$A \Rightarrow C \qquad D \Rightarrow A \qquad D \Rightarrow C$$

- Se necesita análisis semántico del problema.

Problemas derivados de las medidas

- Las medidas clásicas de soporte y confianza presentan problemas cuando las utilizamos para guiar la búsqueda de las reglas de asociación:
  - Soportes alto: dan lugar a reglas con soportes altos en el consecuente → reglas poco útiles
  - Confianza: al estar basada en frecuencias, no detecta cuando el soporte del consecuente es muy alto. Ej:

Confianza (A→B) = 1.0

Confianza (C→D) = 0.84

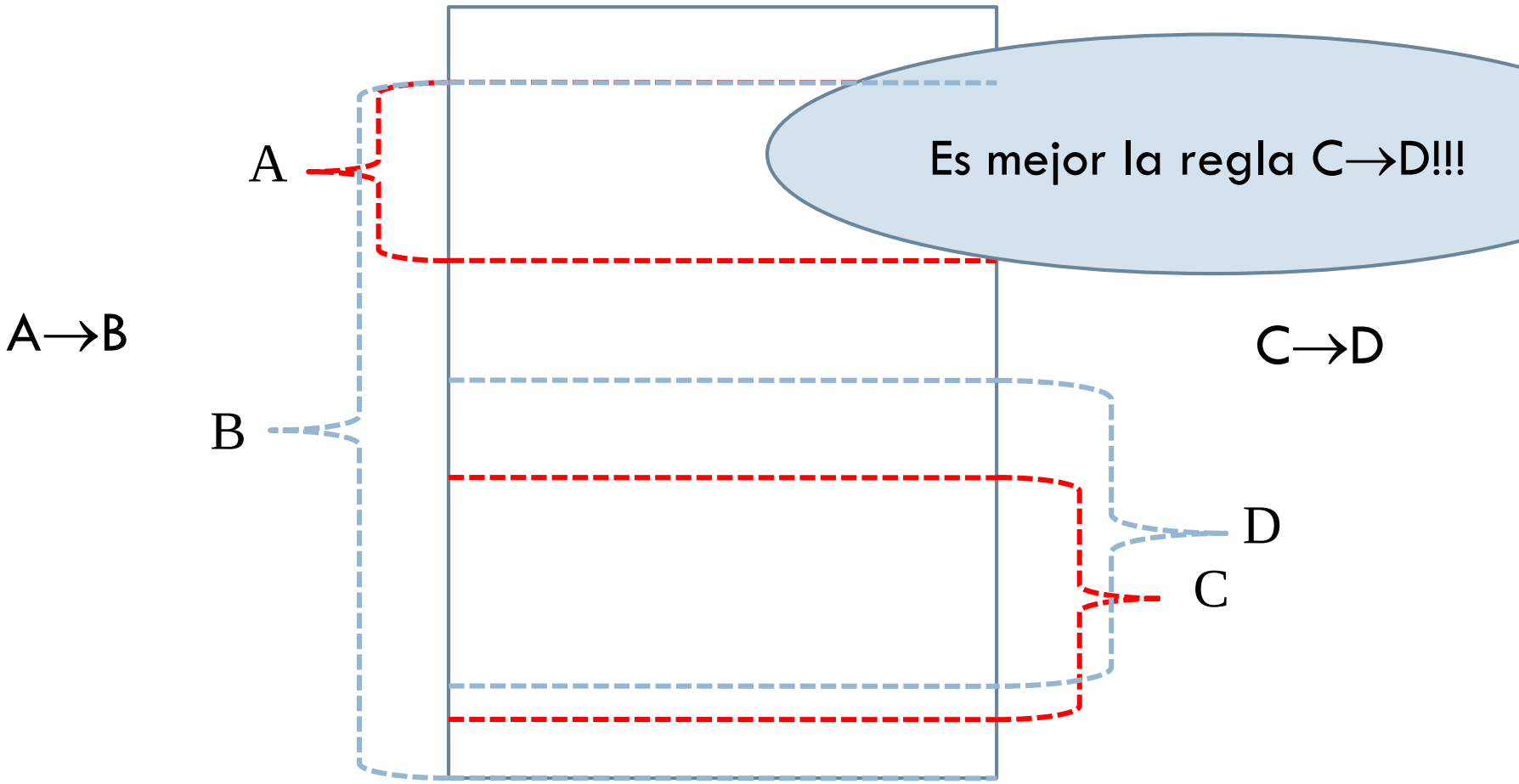
¿Qué regla es mejor?

Problemas derivados de las medidas

Sop(A) = 0.2; Sop(B) = 0.9

Sop(C) = 0.3; Sop(D) = 0.4

Base de datos



Problemas derivados de las medidas

- **Solución**: introducir nuevas medidas de calidad/interés que junto a las clásicas permitan evitar estos problemas.

- **Problema**: No existe la medida de calidad/interés que no tenga algún problema, por lo que tenemos que utilizar varias medidas que se complementen: lift, factor de certeza, etc.

Contenidos

- Problemas de Interpretabilidad
- Evaluación: Medidas de Interés
- Interpretaciones
- Reglas de Asociación Difusas
- Aspectos algorítmicos. El caso de las reglas de asociación jerárquicas
- Evaluación de reglas por grupos

Evaluación: Medidas de Interés

- Medidas objetivas
- Medidas subjetivas

Evaluación: Medidas Objetivas

- Las medidas objetivas de cumplimiento, como la confianza, suelen estar basadas en el cálculo de frecuencias y tener un significado fundamentalmente estadístico.

- Existen decenas de ellas, cada una con sus características.

Medidas Objetivas - Propiedades

- Propiedades deseables de las medidas de interés *I* para reglas de asociación según Piatetsky-Shapiro:
  - P1 $I(A \Rightarrow C) = 0$ cuando son independientes ($supp(A \Rightarrow C) = supp(A)supp(C)$).
  - P2 $I(A \Rightarrow C)$ crece monotonamente con $supp(A \Rightarrow C)$ cuando se mantiene el resto de valores.
  - P3 $I(A \Rightarrow C)$ decrece monotonamente con $supp(A)$ (o $supp(C)$) cuando se mantiene el resto de valores.

G. Piatetsky-Shapiro, Discovery, Analysis, and Presentation of Strong Rules, Knowledge Discovery in Databases (1991) 229-248.

Confianza Confirmada

- Se define para A ⇒ C como:

$$Conf(A \Rightarrow C) - Conf(A \Rightarrow \neg C)$$

- Tiene semántica de predicción: ¿hasta qué punto es útil A para predecir la presencia de C?

- Su rango es [-1; 1] donde 0 significa imposible predecir (independencia), 1 significa que A predice C, -1 significa que A predice ¬C.

Lift / Interés

- Se define para A→C como:

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

- Al considerar sop(C) permite resolver el problema de itemsets muy frecuentes.
- Es una medida simétrica (asociación, no implicación).
- Su rango es [0;+∞]. Valor 1 significa independencia estadística. Valores negativos, dependencia negativa.
- Valores no comparables entre tablas.

Convicción

- Se define para A→C como:

$$\frac{sop(A)sop(\neg C)}{sop(A \Rightarrow \neg C)}$$

- Al considerar sop(¬C) permite resolver el problema de itemsets muy frecuentes.

- Su rango es]0;+ ∞]. Valor 1 significa independencia estadística. Valores negativos, dependencia negativa.

- Valores entre 1.01 y 5 se consideran interesantes. Valores superiores a 5, reglas obvias (umbrales empíricos).

Factor de Certeza

- Se define para A→C como:

$$FC(A \rightarrow C) = \frac{Conf(A \rightarrow C) - Sop(C)}{1 - sop(C)} \; si \; Conf(A \rightarrow C) \geq Sop(C)$$

$$FC(A \rightarrow C) = \frac{Conf(A \rightarrow C) - Sop(C)}{sop(C)} \; si \; Conf(A \rightarrow C) < Sop(C)$$

- Proviene del ámbito de los sistemas expertos.
- Al considerar sop(C) permite resolver el problema de itemsets muy frecuentes.
- Es una medida de implicación. Mide la variación de nuestra creencia en C cuando se cumple A con respecto a la creencia a priori en C.
- Su rango es [-1; 1]. Valor 0 significa independencia estadística.
- Propiedades muy interesantes. Relación con interés y convicción.

Yule's Q

- Se define para A→C como:

$$\frac{Sop(AC) * Sop(\neg A \neg C) - Sop(A \neg C) * Sop(\neg AC)}{Sop(AC) * Sop(\neg A \neg C) + Sop(A \neg C) * Sop(\neg AC)}$$

- Esta medida representa la correlación entre dos eventos dicotómicos relacionados positivamente.
- Su rango es [-1; 1]. Valor 0 significa independencia estadística; valores negativos dependecia negativa; y valores positivos dependencia positiva.
- Cumple la mayoría de las propiedades de la literatura para las medidas de interés.

Diferencia absoluta de confianza

- Se define para una regla A → C como

$$Conf(A \Rightarrow C) - sop(C)$$

- Al considerar sop(C) permite resolver el problema de itemsets muy frecuentes.

- Es una medida de implicación.

- Su rango es [-1; 1]. Valor 0 significa independencia estadística.

Ratio de confianza

- Se define para una regla A → C como

$$1 - \frac{Conf(A \Rightarrow C)}{sop(C)} \text{ ó } 1 - \frac{sop(C)}{Conf(A \Rightarrow C)}$$

- Al considerar sop(C) permite resolver el problema de itemsets muy frecuentes.
- Es una medida de implicación.
- Su rango es [-1; 1]. Valor 0 significa independencia estadística.
- Especialmente adecuada para descubrir reglas que corresponden a itemsets poco frecuentes.

Diferencia de información

- Se define para una regla A → C como

$$(-sop(C)log_2 sop(C) - sop(\neg C)log_2 sop(\neg C)) -$$
$$(-conf(A \Rightarrow C)log_2 conf(A \Rightarrow C) - conf(A \Rightarrow \neg C)log_2 conf(A \Rightarrow \neg C)$$

- Cada parte es una medida de información basada en entropía. La primera considera la información dada solo por C, la segunda la dada por C en presencia de A. Se mide la ganancia (o pérdida) de información sobre C al conocer A.
- Es una medida de implicación.
- Su rango es [-∞;+ ∞].
- Se ve afectada por el soporte.

Chi-cuadrado normalizado

- Difícil de interpretar.
- Se ve afectada por el soporte.
- Escala poco intuitiva. Valor del estadístico no es significativo sin el test con la distribución.

Medidas de Interés

- Existen otras muchas 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)	$\sqrt{P(A,B)}\max(P(B\|A) - P(B), P(A\|B) - P(A))$

Medidas de Interés – Otras Propiedades

- Y otras propiedades propuestas por Tan et al. (O1,O2,O3 y O4):

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	$\left(\sqrt{\frac{2}{\sqrt{3}}-1}\right)\left(2-\sqrt{3}-\frac{1}{\sqrt{3}}\right)\ldots 0 \ldots \frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No	No	No	No	No

P. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," in *Proc. 8th Int. Conf. KDD*, 2002, pp. 32–41.

Evaluación: Medidas Subjetivas

- Las medidas objetivas tienen en cuenta tan solo los datos.
- Por el contrario, las medidas subjetivas de cumplimiento miden el interés de las reglas no en términos de cumplimiento, sino del resto de características que se pretenden: conocimiento no trivial, novedoso y potencialmente útil.
- Deben tener en cuenta el conocimiento previo del usuario (creencias, necesidades ...)
- Existen diversos enfoques, cada uno con sus características.
- Difíciles de encontrar en productos comerciales, ya que es difícil predecir la utilidad de un patrón y definir una medida para ello.

Utilidad

- Silberschatz y Tuzhilin (1996) propusieron dividir el espacio de los patrones encontrados en clases de equivalencia, asociando a cada clase un tipo de acción.
- A la hora de medir la Utilidad, hay que tener en cuenta:
  - Restricciones: ¿qué condiciones o qué contexto es necesario para que el patrón se cumpla?
  - Tiempo de vida: ¿durante cuánto tiempo será útil la información dada por el patrón?
  - Esfuerzo: ¿qué debemos hacer para actuar según nos muestre el patrón?
  - Efectos laterales: ¿se puede prever algún efecto lateral?
  - Impacto: desde la obtención del patrón, ¿se han producido cambios en la actualidad?
  - Prontitud: ¿cuándo podemos actuar y utilizar la información que nos brinda el patrón?

Reglas inesperadas

- Aquellas reglas que contradigan las creencias del usuario son reglas inesperadas y, como tales, interesantes.
- Hay que especificar cómo representar las creencias del usuario y cómo medir lo inesperado.
- Habitualmente mediante estadísticos y distancias entre los esperados por el usuario y los reales.

Reglas inesperadas

- Enfoques principales:
  - Medidas probabilísticas: se han usado redes bayesianas para poder usar probabilidades condicionadas y determinar su coherencia.
  - Medidas de la distancia sintáctica: Este enfoque se basa en la distancia entre las nuevas reglas y el conjunto de creencias. Por ejemplo, si los consecuentes de una regla son los mismos que los esperados por el usuario, pero los antecedentes son muy distintos, entonces esta regla se consideraría interesante.
  - Contradicción lógica: Usa una medida objetiva para indicar lo que el usuario espera, y después se analiza si hay alguna diferencia con los grados esperados por el usuario de dichas medidas.

Reglas inesperadas

- Un tipo de contradicción son las paradojas, como la de Simpson:
  - supongamos que tenemos que E representa un efecto beneficioso, C representa la causa por la que se toma un determinado medicamento, y F y ¬F representan ser mujer o no, respectivamente. La paradoja puede formularse como sigue:

$$
\begin{aligned}
P(E|C) &> P(E|\neg C), && (1)\\
P(E|C,F) &< P(E|\neg C,F), && (2)\\
P(E|C,\neg F) &< P(E|\neg C,\neg F) && (3)
\end{aligned}
$$

  - La primera ecuación dice que el medicamento es beneficioso para todos los pacientes, la segunda dice que es perjudicial para mujeres y la última indica que es perjudicial para hombres.

Contenidos

- Problemas de Interpretabilidad
- Evaluación: Medidas de Interés
- Interpretaciones
- Reglas de Asociación Difusas
- Aspectos algorítmicos. El caso de las reglas de asociación jerárquicas
- Evaluación de reglas por grupos

Marco formal de Reglas de Asociación

- Corresponde a la formalización abstracta del problema de búsqueda de Reglas de Asociación tal y como la hemos visto hasta ahora. **Independiente de los datos!**
- Conjunto I de Items
  - Representan elementos relevantes a relacionar
- Multiconjunto T de Transacciones. $\forall t \in T, t \subseteq I$.
  - Significado de una transacción: depende del criterio utilizado para agrupar ítems (compra conjunta, datos de una misma entidad, etc.)
- Reglas de asociación $A \rightarrow C$, donde A e C son conjuntos de items (**itemsets**), $A,C \subseteq I$, que cumplen $A \cap C = \varnothing$
  - Significado: la presencia de A está asociada a la presencia de C en las transacciones de t.
  - Significado matizado por el significado de ítems y transacciones
  - Semántica de la asociación matizada por las medidas empleadas

Concepto de Interpretación

- Objetivo: encontrar reglas de asociación en conjuntos de datos con una cierta estructura, que representen conocimiento novedoso, potencialmente útil, etc.
- Para ello, lo primero es determinar qué elementos de la estructura de los datos van a corresponder con los elementos del marco formal (ítems y transacciones)
- Una **INTERPRETACIÓN** es una correspondencia que se establece entre elementos de la estructura de los datos y elementos del marco formal. Hay muchas más interpretaciones posibles de las que podemos contar aquí.
- Va a definir qué entendemos por ítems y transacciones en el contexto de los datos que tenemos, y con ello van a contribuir a definir la semántica de las reglas.

N. MARÍN, M.D. RUIZ, D. SÁNCHEZ, **Fuzzy frameworks for mining data associations: fuzzy association rules and beyond**, WIRES Data Mining & Knowledge Discovery 6(2), 2016, pp. 50-69.

Interpretación: tabular común

- Supongamos que los datos tienen estructura tabular.
- Interpretación habitual:
  - Items: Parejas (atributo,valor)
  - Transacciones: Registros

DNI	Puesto	Sueldo	Estudios
111111111	Administrativo	Bajo	Medios
222222222	Programador	Medio	Medios
333333333	Analista	Medio	Superiores
444444444	Gerente	Alto	Superiores

	Transacción
1	(Puesto,Administrativo), (Sueldo,Bajo), (Estudios,Medios)
2	(Puesto,Programador), (Sueldo, Medio), (Estudios,Medios)
3	(Puesto,Analista), (Sueldo,Medio), (Estudios,Superiores)
4	(Puesto,Gerente), (Sueldo,Alto), (Estudios,Superiores)

- Ejemplo de regla: (Salario,alto) $\rightarrow$ (Estudios,Superiores)
  - Todo el que tiene un salario alto tiene estudios superiores, ó
  - Salario alto implica estudios superiores

Interpretación: ítems negados

Datos binarios:

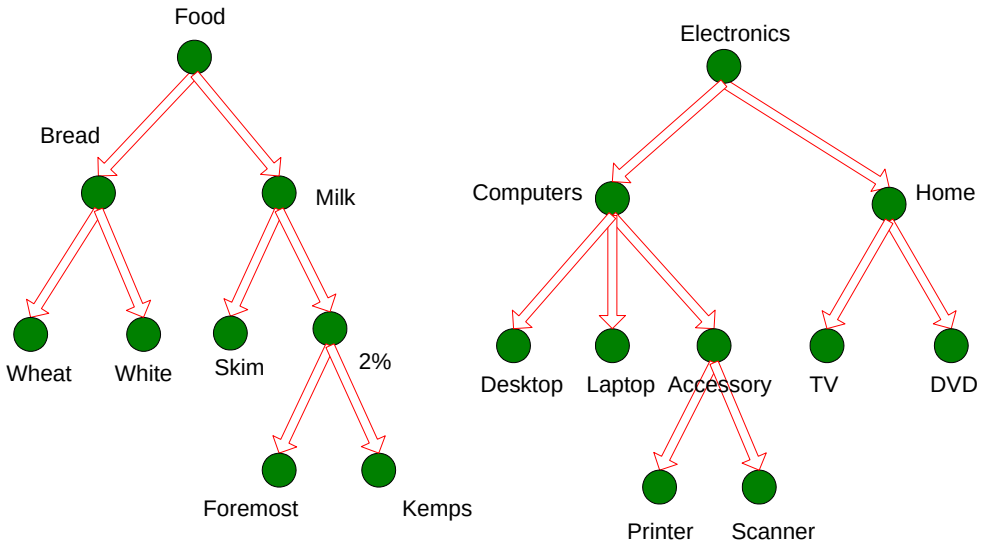
BD:

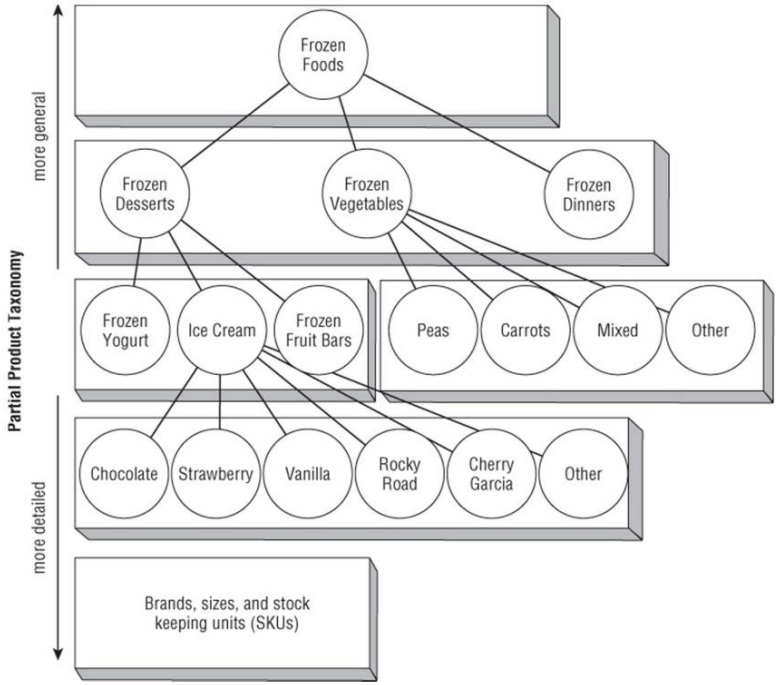
i_1	i_2	i_3	i_4
1	0	1	0
0	0	0	0
0	1	1	0
0	1	1	1
1	1	1	1
1	1	1	1

- Interpretación habitual:
  - Items: columnas i
  - Transacciones: filas. Item está en transacción si el valor es 1.
- Interpretación con ítems negados
  - Items: dos ítems por columna, i y ¬i
  - Transacciones: filas. Item i está en transacción si el valor es 1. Item ¬ i está en transacción si valor es 0.

Interpretación: Reglas Jerárquicas

- En esta extensión se considera que disponemos de una o varias jerarquías de items.



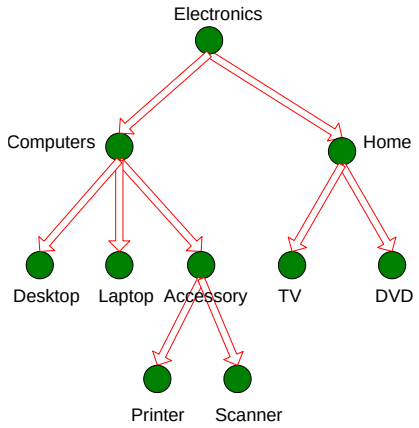


Interpretación: Reglas Jerárquicas

- Los datos consisten en:
  - un conjunto de transacciones que contienen ítems básicos (ej: cestas de compra que contienen productos identificados mediante código de barras), y
  - la jerarquía de categorías que los agrupa en distintos niveles (como la que hemos visto, para todos los productos).
- Interpretación:
  - Items son la unión de los ítems básicos y las categorías de la jerarquía
  - Transacciones se forman tomando cada transacción de ítems básicos y añadiendo los ancestros en la jerarquía de todos los items presentes. Por ejemplo:

Transacción básica: {Desktop, Printer}

Interpretación: {Desktop , Printer, Computer, Electronics, Accessory}



Reglas de Asociación Jerárquicas

- **Utilidad:** si items de niveles bajos no tienen soporte suficiente para generar reglas, podemos tomar el item inmediatamente superior en la jerarquía. Por el contrario, si el soporte es excesivo, podemos disminuir en la jerarquía.

- **Por ejemplo:** mínimo soporte 0,25
  - 2% milk → white bread Sop = 0,2 ⟸ **Infrecuente**
  - milk → bread Sop = 0,3

Interpretación: Reglas Secuenciales

- Un patrón secuencial es una secuencia de itemsets básicos que tienden a aparecer en un orden prefijado, por ejemplo en tiempo, o cualquier otro tipo de criterio de ordenación total.
- Asumimos que la aparición de ítems básicos en los datos está ordenada. Ejemplo sobre cestas de compra: cuando la compra de cada artículo tiene asociada una fecha y hora. Ejemplo en texto: ítems básicos son palabras, que aparecen ordenadas.
  - Items: secuencias ordenadas de itemsets básicos
  - Transacciones: conjuntos de estos ítems
- Las reglas tienen una sola secuencia tanto en la parte izquierda como en la derecha. Ambas pueden estar formadas por varios ítemsets básicos ordenados.

Interpretación: Reglas Secuenciales

- Ejemplo de reglas secuenciales: si un cliente compra congelado y luego cerveza, más tarde comprará congelado

  {Congelado} {Cerveza} → {Congelado}

  {Congelado y Cerveza } → {Congelado} (distinta!)

- Ejemplo en Minería de Textos:

  {Minería}{de} → {Datos}

- Se evalua confianza a partir de las secuencias {Minería}{de} y {Minería}{de}{Datos}. Secuencia consecuente debe ser posterior!
- OJO: Secuencias respetan orden, pero pueden saltar ítems
  - {Minería}{Datos} es una secuencia válida a partir del texto "Minería de Datos". Sin embargo, {Datos}{Minería} no.

Interpretación: Reglas Cuantitativas

- Utilizadas cuando tenemos datos estructurados con variables que tengan dominios numéricos con muchos valores, particularmente dominios continuos (reales, etc.)
- Si intentamos extraer reglas a partir de estas BDs usando interpretaciones que definan ítems como pares (Atributo,valor) tenemos dos problemas:
  - El soporte de la mayoría de los items es muy bajo por lo general
  - Reglas pobres semánticamente

Interpretación: Reglas Cuantitativas

- **Solución**: dividir el dominio de estos atributos en intervalos y determinar los ítems como pares (atributo, intervalo).
- Dos enfoques para determinar los intervalos:
  - Definir unos intervalos a priori (conocimiento experto):
  - Utilizar un método automático. Dos subopciones:
    - Aplicar un método que aprenda los intervalos y a partir de ellos extraer las reglas de asociación.
    - Búsqueda de reglas cuantitativas: dejar que el algoritmo busque los mejores intervalos para proporcionar reglas con buenos valores de las medidas de soporte y cumplimiento.

R. Srikant, R. Agrawal, **Mining quantitative association in large relational tables**, Proc. SIGMOD, New York, 1996, 1-12.

J.J. Mata, J.Riquelme, **An evolutionary algorithm to discover numeric association rules**, in: ACM Symposium on Applied Computing, Madrid, Spain, 2002.

Interpretación: Reglas Cuantitativas

Ejemplo: Base de datos tabular con 3 variables

- Dominio de las variables:
  - Edad → [0, 120]
  - Peso → [0, 200]
  - Altura → [0, 220]

Edad	Peso	Altura
10	40	130
50	90	185
2	10	85
70	70	165

- Intervalos definidos por un experto dado:
  - Edad → [0, 18] ,]18, 30],]30, 65],]65, 120]
  - Peso → [0, 25] ,]25, 70],]70, 100],]100, 200]
  - Altura → [0, 100],]100, 170],]170, 220]

Edad	Peso	Altura
[0, 18]	]25, 70]	]100, 170]
]30, 65]	]70, 100]	]170, 220]
[0, 18]	[0, 25]	[0, 100]
]65, 120]	]25, 70]	]100, 170]

Interpretación: Reglas Cuantitativas

- Definir unos intervalos a priori (conocimiento experto):
  - Riqueza semántica.
  - Podemos aplicar cualquiera de los métodos clásicos de extracción reglas de asociación.
  - Problemas: quizá los intervalos no sean los más adecuados para obtener buenas reglas ya que: Si hay muchos intervalos pequeños, podemos tener bajo soporte de cada uno de ellos. Por el contrario, los intervalos grandes dan lugar a reglas muy generales y a reglas inútiles cuando aparecen en el consecuente de las reglas.

Interpretación: Reglas Cuantitativas

- Utilizar un método automático. Posibles problemas:
  - Intervalos de semántica pobre.
  - Complejidad computacional (orden cuadrático en el número de valores del atributo).
  - No podemos utilizar medidas de información porque los ejemplos no tienen asociada una clase (aprendizaje no supervisado).
  - Muchas reglas.

Interpretación: Dependencias aproximadas.

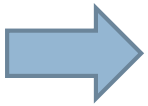
- Patrones en bases de datos relacionales. Corresponden a dependencias funcionales con excepciones.
- Dependencia funcional entre dos atributos V y W se cumple en una tabla r si el valor de V determina el valor de W. Puede formularse como sigue: $\forall t,s \in r$ Si $t[V]=s[V]$ entonces $t[W]=s[W]$
- Excepciones: pares de tuplas en las que $t[V]=s[V]$ y $t[W]\neq s[W]$
- Interpretación:
  - Items: atributos
  - Transacciones: asociadas a pares de tuplas de r.
  - El ítem asociado al atributo V está en la transacción asociado al par de tuplas (t,s) si y solo si $t[V]=s[V]$

D. Sánchez, J.M. Serrano, I. Blanco, M.J. Martín-Bautista, M.A. Vila, **Using association rules to mine for strong approximate dependencies,** Data Mining & Knowledge Discovery 16, 2008, pp. 313-348.

Interpretación: Dependencias aproximadas.

- Ejemplo:

Tupla	Edad	Peso	Altura
1	18	60	170
2	18	90	185
3	20	60	185
4	70	60	185



Par	Item-Edad	Item-Peso	Item-Altura
(1,2)	1	0	0
(1,3)	0	1	0
(1,4)	0	1	0
(2,3)	0	0	1
(2,4)	0	0	1
(3,4)	0	1	1

- Las reglas de asociación en el conjunto de transacciones de la derecha son dependencias aproximadas en la tabla original.
- Las medidas de estas reglas son válidas para valorar la calidad de las dependencias.
- Puede verse que hay distintas interpretaciones para los mismos datos según el tipo de patrón buscado!

Interpretación: Dependencias graduales.

- Representan asociaciones entre la variación (incrementos o decrementos) en valores de atributos. Representan correlaciones pos. ó negativas
- Dependencia gradual entre dos atributos V y W se cumple en una tabla r si se dan reglas de estos tipos:

  $\forall t,s \in r$ Si $t[V]>s[V]$ entonces $t[W]>s[W]$

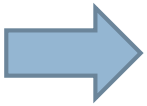
  $\forall t,s \in r$ Si $t[V]>s[V]$ entonces $t[W]<s[W]$ … (4 combinaciones)
- Interpretación:
  - Items: pares (atributo, variación) donde variación puede ser < ó >
  - Transacciones: asociadas a pares de tuplas de r. El ítem asociado al par (V,>) está en la transacción asociado al par de tuplas (t,s) si y solo si $t[V]>s[V]$. La presencia del par (V,<) se da sii $t[V]<s[V]$.

F. Berzal, J.C. Cubero, D. Sánchez, M.A. Vila, J.M. Serrano, **An alternative approach to discover gradual dependencies,** Int. J. of Uncertainty, Fuzziness, and Knowledge-Based Systems 15(5), 2007, pp. 559-570.

Interpretación: Dependencias graduales.

- Ejemplo:

Tupla	Edad	Peso	Altura
1	18	60	170
2	19	90	185
3	20	65	190
4	70	57	185



Par	(Edad,>)	(Peso,>)	(Altura,>)	(Edad,<)	(Peso,<)	(Altura,<)
(1,2)	0	0	0	1	1	1
(1,3)	0	0	0	1	1	1
(1,4)	0	1	0	1	0	1
(2,3)	0	1	0	1	0	1
(2,4)	1	0	0	0	1	0
(3,4)	0	1	1	1	0	0

- Las reglas de asociación en el conjunto de transacciones de la derecha son dependencias graduales en la tabla original.
- Las medidas de estas reglas son válidas para valorar la calidad de las dependencias graduales. Admiten excepciones.

Contenidos

- Problemas de Interpretabilidad
- Evaluación: Medidas de Interés
- Interpretaciones
- Reglas de Asociación Difusas
- Aspectos algorítmicos. El caso de las reglas de asociación jerárquicas
- Evaluación de reglas por grupos

Reglas de Asociación Difusas

- El enfoque clásico es ver las reglas de asociación difusas como un caso particular de regla de asociación cuantitativa por:
  - Problemas en los limites de los intervalos cuando coinciden con valores relativamente frecuentes.
  - Semántica imprecisa de la división del dominio en muchas ocasiones, que no encaja bien con el uso de intervalos.
- Enfoque clásico para reglas de asociación difusas.
  - Definir partición difusa (intervalos difusos). Son intervalos con bordes imprecisos.
  - Ventaja: la transición entre un intervalo y otro es gradual, no brusca.

T. Hong and Y. Lee, "An overview of mining fuzzy association rules," in Studies in Fuzziness and Soft Computing, vol. 220, H. Bustince, F. Herrera, and J.Montero, Eds. Berlin, Germany: Springer-Verlag, 2008, pp. 397–410.

Reglas de Asociación Difusas

- Definición de las medidas de soporte y confianza para las reglas de asociación difusas en el enfoque clásico:
  - Soporte
    - De un itemset X

      Soporte (X) = $\sum_i^{Total\ ejemplos} \mu_X(i)/Total\ ejemplos$
    - De la regla (X $\rightarrow$Y)

      Soporte (X $\rightarrow$Y) = $\sum_i^{Total\ ejemplos} \mu_{XY}(i)/Total\ ejemplos$

Donde $\mu_X(i)$ es el grado con el que el itemset X cubre el ejemplo i.

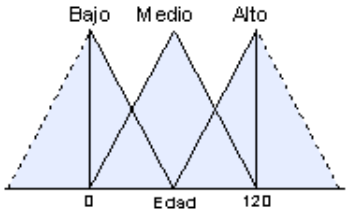
- Confianza de la regla (X $\rightarrow$Y)

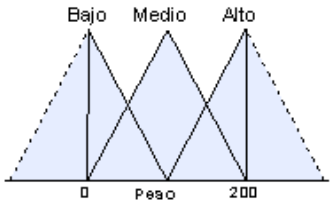
  Soporte (X $\rightarrow$ Y) / Soporte (X)

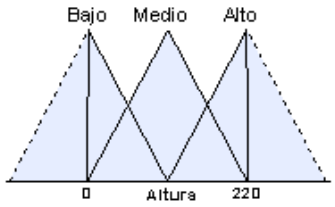
Reglas de Asociación Difusas

Ejemplo: Base de datos con 3 variables

- Dominio de las variables:
  - Edad → [0, 120]
  - Peso → [0, 200]
  - Altura → [0, 220]
- Particiones







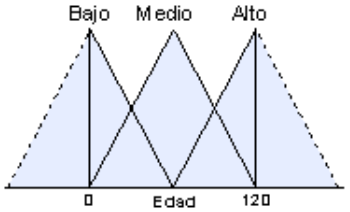
Edad	Peso	Altura
10	40	130
50	90	185
2	10	85
70	70	165

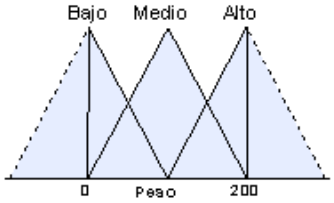
Edad	Peso	Altura
(Bajo,0.9) (Medio, 0.1) (Alto, 0.0)	(Bajo,0.6) (Medio, 0.4) (Alto, 0.0)	(Bajo,0.0) (Medio, 0.9) (Alto, 0.1)
(Bajo,0.1) (Medio, 0.9) (Alto, 0.0)	(Bajo,0.1) (Medio, 0.9) (Alto, 0.0)	(Bajo,0.0) (Medio, 0.45) (Alto, 0.65)
(Bajo,0.98) (Medio, 0.02) (Alto, 0.0)	(Bajo,0.9) (Medio, 0.1) (Alto, 0.0)	(Bajo,0.8) (Medio, 0.2) (Alto, 0.0)
(Bajo,0.0) (Medio, 0.85) (Alto, 0.15)	(Bajo,0.7) (Medio, 0.3) (Alto, 0.0)	(Bajo,0.0) (Medio, 0.5) (Alto, 0.5)

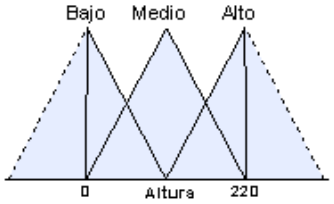
Reglas de Asociación Difusas

Ejemplo: Cálculo de soporte y confianza

- Particiones







Edad	Peso	Altura
(Bajo,0.9) (Medio, 0.1) (Alto, 0.0)	(Bajo,0.6) (Medio, 0.4) (Alto, 0.0)	(Bajo,0.0) (Medio, 0.9) (Alto, 0.1)
(Bajo,0.1) (Medio, 0.9) (Alto, 0.0)	(Bajo,0.1) (Medio, 0.9) (Alto, 0.0)	(Bajo,0.0) (Medio, 0.45) (Alto, 0.65)
(Bajo,0.98) (Medio, 0.02) (Alto, 0.0)	(Bajo,0.9) (Medio, 0.1) (Alto, 0.0)	(Bajo,0.8) (Medio, 0.2) (Alto, 0.0)
(Bajo,0.0) (Medio, 0.85) (Alto, 0.15)	(Bajo,0.7) (Medio, 0.3) (Alto, 0.0)	(Bajo,0.0) (Medio, 0.5) (Alto, 0.5)

- Soporte (Edad,Bajo) = (0.9+0.10.98+0.0) / 4 = 0.495
- Soporte (Edad,Bajo → Peso,Bajo) =

  (min(0.9,0.6)+min(0.1,0.1)+min(0.98,0.9)+min(0.0,0.7)) / 4 = 0.4

- Confianza (Edad,Bajo → Peso,Bajo) = 0.4 / 0.495 = 0.8

Reglas de Asociación Difusas

- Tres enfoques para extraer las reglas:
  - Realizar las particiones difusas a priori y extraer a partir de ellas las reglas que cumplan los umbrales de mínimo soporte y mínima confianza. Las particiones pueden ser proporcionadas por el experto o aprendidas mediante un método automático.

Ch. Kuok, A. Fu, M. Wong, Mining fuzzy association rules in databases, ACM SIGMOD Record 27:1 (1998) 41-6
T.P. Hong, C. Kuo, S. Chi, Trade-off between time complexity and number of rules for fuzzy mining from quantitative data, Internat. J. Uncertain Fuzziness Knowledge-Based Systems 9 (5) (2001) 587–604.

  - Aprender las reglas y particiones difusas las mismo tiempo → Problemas de interpretabilidad

T.P. Hong, C. Chen, Y. Lee, and Y. Wu, "Genetic-fuzzy data mining with divide-and-conquer strategy," IEEE Trans. Evol. Comput., vol. 12, no. 2, pp. 252–265, 2008
J. Alcalá-Fdez, R. Alcalá, M.J. Gacto, F. Herrera. Learning the Membership Function Contexts for Mining Fuzzy Association Rules by Using Genetic Algorithms. Fuzzy Sets and Systems 160:7 (2009) 905-921

Reglas de Asociación Difusas

- El enfoque clásico es muy útil y se utiliza de forma muy extendida. Sin embargo, constituye una simplificación excesiva del problema, ya que ha consistido básicamente en:
  - Considerar que las reglas de asociación difusas consisten en reglas difusas, de las utilizadas habitualmente en control.
  - No parten del marco abstracto formal de las reglas de asociación para proporcionar extensiones con semántica del mismo.
  - La mayoría de las propuestas utilizan enfoques muy simplistas para el cardinal, basados en medidas de energía.
  - No distinguen el modelo formal de la interpretación. La estructura y medidas son ad-hoc.

Reglas de Asociación Difusas

- Enfoques más avanzados sí proponen modelos más generales, que extienden con semántica clara el marco formal, estudiando entre otros:
  - La semántica de los conjuntos difusos empleados en la definición de ítems, transacciones y reglas, y las interpretaciones que permiten obtenerlos a partir de los datos
  - La determinación del grado de inclusión de itemsets y el uso de medidas de cardinal avanzadas y la definición de umbrales
- La visión clásica de reglas de asociación difusa se obtiene mediante una interpretación concreta del marco formal basado en transacciones difusas conjuntivas.
- Pueden combinarse con las distintas interpretaciones estudiadas!

M. DELGADO, N. MARÍN, D. SÁNCHEZ, M.A. VILA **Fuzzy associations rules: general model and applications.** IEEE Transactions on Fuzzy Systems 11(2), 2003, pp. 214-225.

N. MARÍN, M.D. RUIZ, D. SÁNCHEZ, **Fuzzy frameworks for mining data associations: fuzzy association rules and beyond**, WIRES Data Mining & Knowledge Discovery 6(2), 2016, pp. 50-69.

Contenidos

- Problemas de Interpretabilidad
- Evaluación: Medidas de Interés
- Interpretaciones
- Reglas de Asociación Difusas
- Aspectos algorítmicos. El caso de las reglas de asociación jerárquicas
- Evaluación de reglas por grupos

Aspectos algorítmicos

- La búsqueda de reglas de asociación tiene alta complejidad, por eso se han desarrollado muchos algoritmos buscando eficiencia.
- En principio, la extracción de reglas de asociación para todas las interpretaciones que hemos visto puede realizarse utilizando un mismo algoritmo general eficiente.
- Sin embargo, este enfoque plantea problemas importantes:
  - La complejidad computacional de algunas interpretaciones es muy alta (gran aumento del número de ítems y transacciones). Se necesita mucho tiempo y espacio.
  - Aumenta de manera importante el número de reglas, muchas de ellas triviales.
- En la práctica, algoritmos específicos para cada caso.

Ejemplo: Reglas Jerárquicas

Implementaciones

- **Enfoque Original:**
  - Añadir a cada transacción los ancestros en la jerarquía de todos los items presentes, y buscar reglas ahí. Por ejemplo:

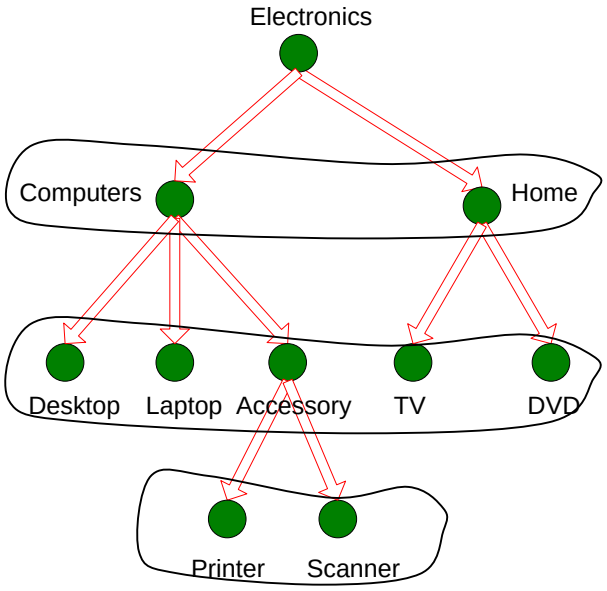
{Desktop, Printer} → {Desktop , Printer, Computer, Electronics, Accessory}

Encontrar ahora, en esta base de datos nueva, los itemsets frecuentes

  - Problema: más lento, muchas reglas obvias.

Ejemplo: Reglas Jerárquicas

- **Enfoque Habitual:**
  - Encontrar los itemsets frecuentes combinando entre sí los ítems pertenecientes a un mismo nivel.



  - Los itemsets frecuentes obtenidos de cada nivel son combinados para construir las reglas de asociación.
  - Problema: más lento, muchas reglas obvias.

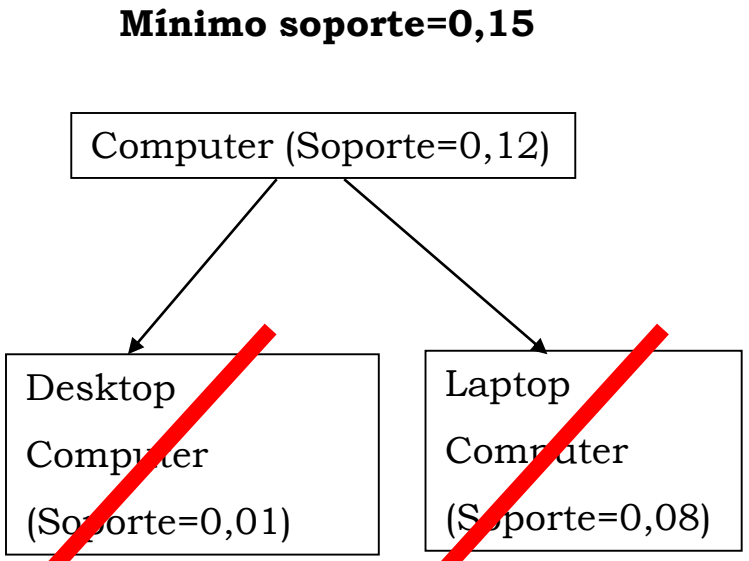
Ejemplo: Reglas Jerárquicas

- **Solución**: Utilizar la información del soporte de los itemsets de cada nivel para podar la búsqueda de itemsets frecuentes.
- **Dos alternativas:**

Primera: Utilizar el mismo mínimo soporte para todos los niveles de la jerarquía. Si en un nivel un ítem no supera el umbral de mínimo soporte no se comprueban sus descendientes (poda similar al método Apriori):

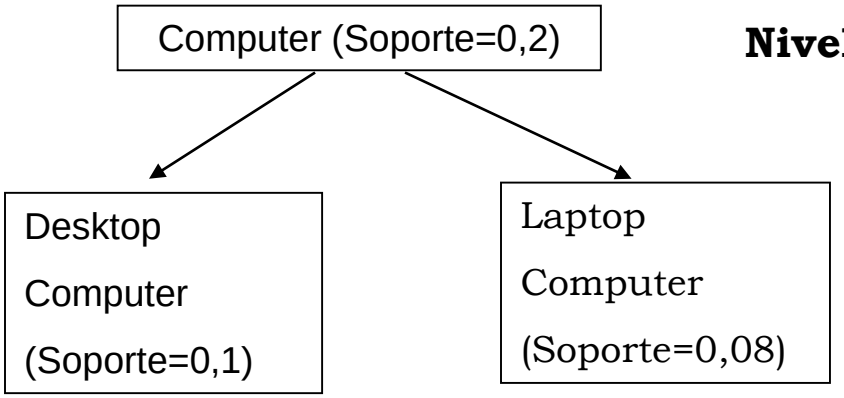
Desvetaja:

- Mínimos soportes elevados pueden provocar la perdida de asociaciones interesantes en nivel bajos.
- Mínimos soportes bajos pueden generar muchas asociaciones no interesantes de los niveles altos.



Ejemplo: Reglas Jerárquicas

Segunda (Reduced Support): Utilizar mínimos soportes menores a medida que bajamos de nivel en la jerarquía.



Nivel 1: min_sop=0,15

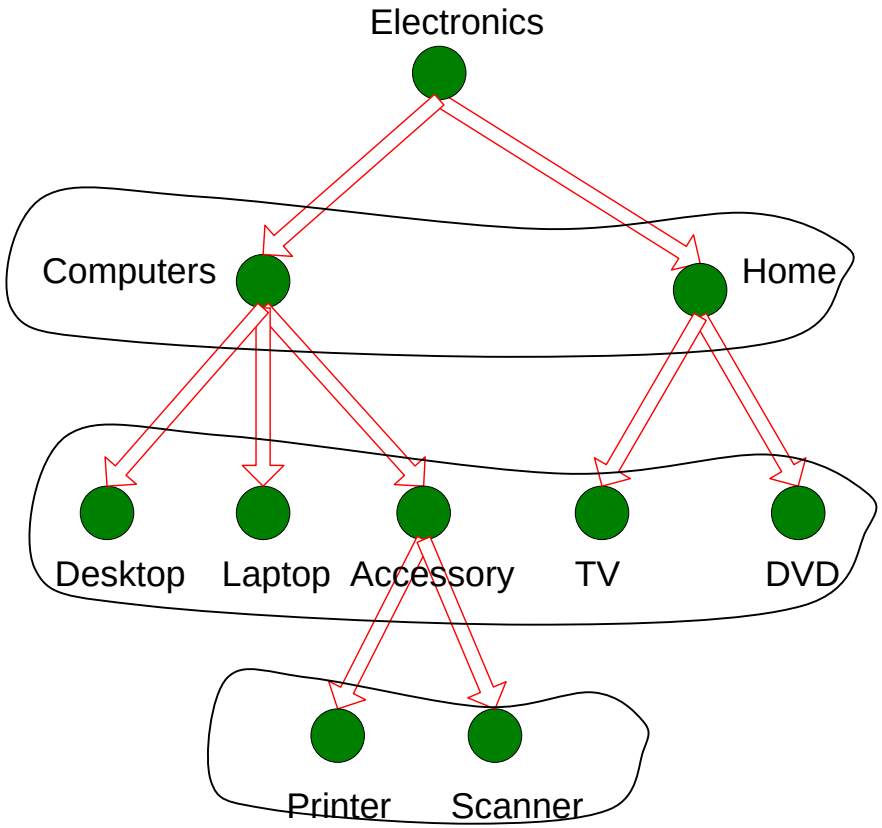
Nivel 2: min_sop=0,07

Existen 4 estrategías de búsqueda basada en este enfoque:

- Level by level independent
- Level-cross filtering by k-itemset
- Level-cross filtering by single item
- Controlled level-cross filtering by single item

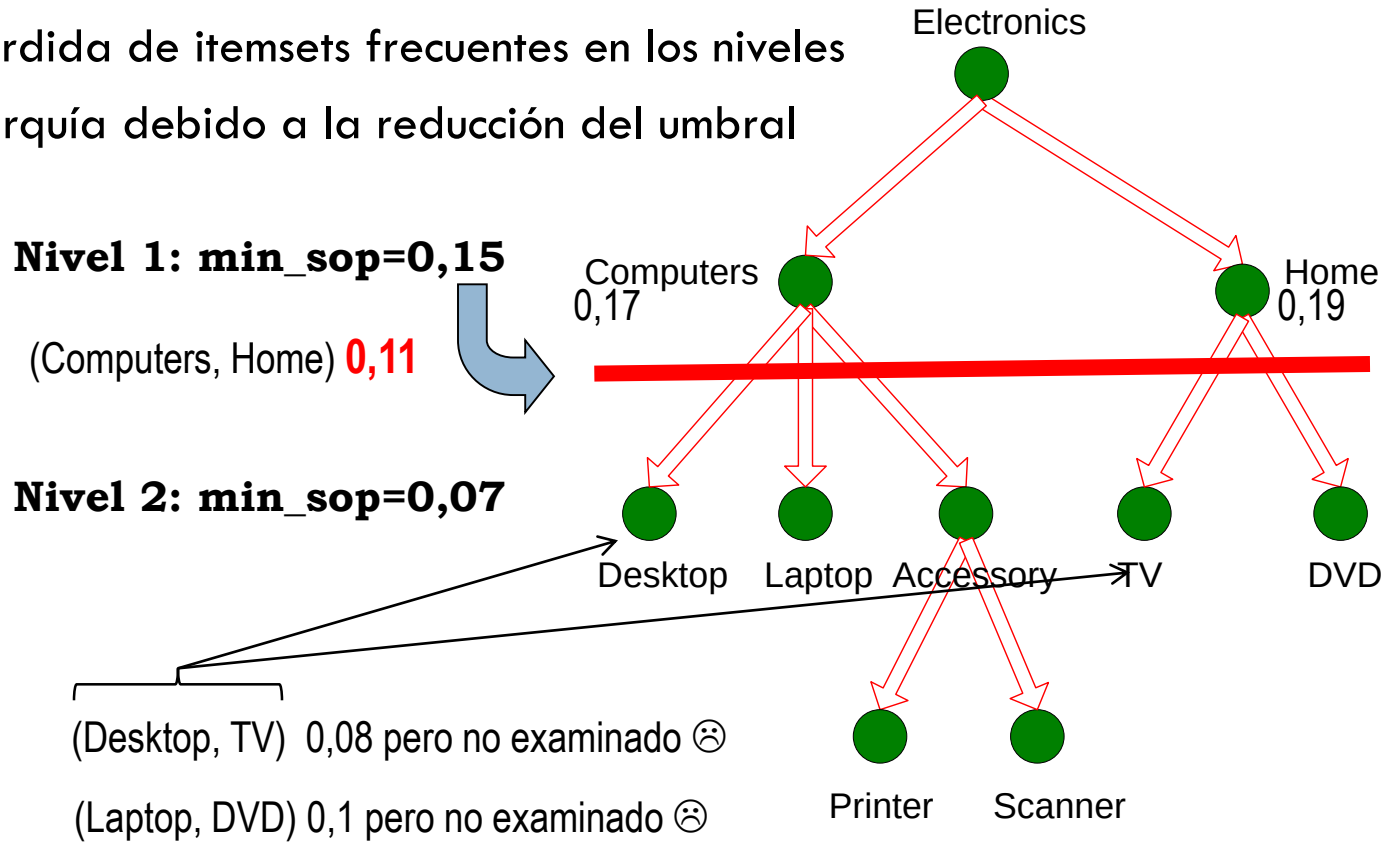
Ejemplo: Reglas Jerárquicas

- **Level by level independent**
  - Reduce el soporte mínimo en cada nivel y examina todos los niveles siempre que sus nodos padres sean frecuentes.



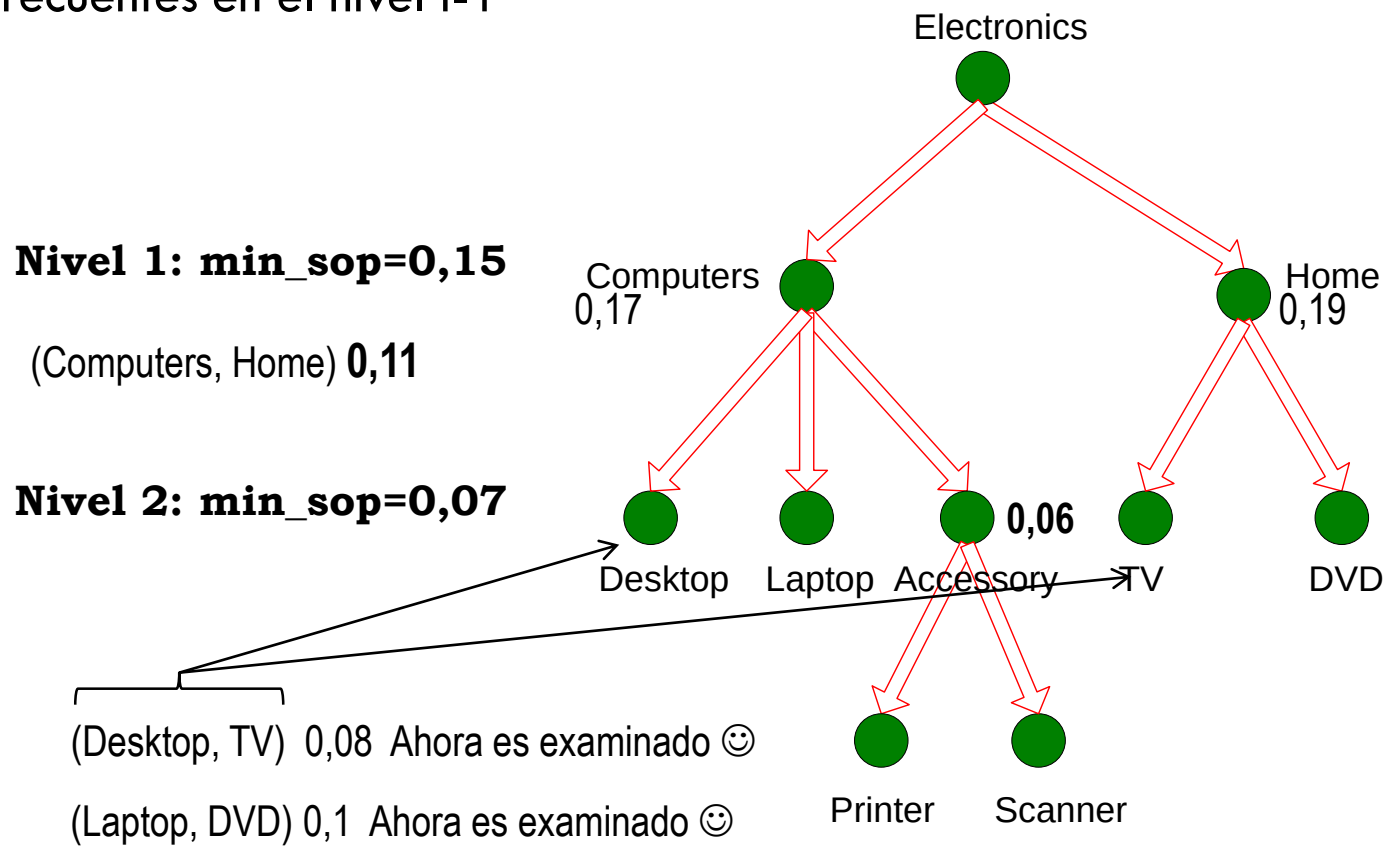
Ejemplo: Reglas Jerárquicas

- **Level-cross filtering by k-itemset**
  - Reduce el soporte mínimo en cada nivel. Un k-itemset del nivel i es examinado si y solo si su correspondiente k-itemset padre del nivel i-1 es frecuente.
  - Problema: Perdida de itemsets frecuentes en los niveles bajos de la jerarquía debido a la reducción del umbral



Ejemplo: Reglas Jerárquicas

- **Level-cross filtering by single item**
  - Un k-itemset del nivel i es examinado si todos sus ítems tienen padres frecuentes en el nivel i-1



Ejemplo: Reglas Jerárquicas

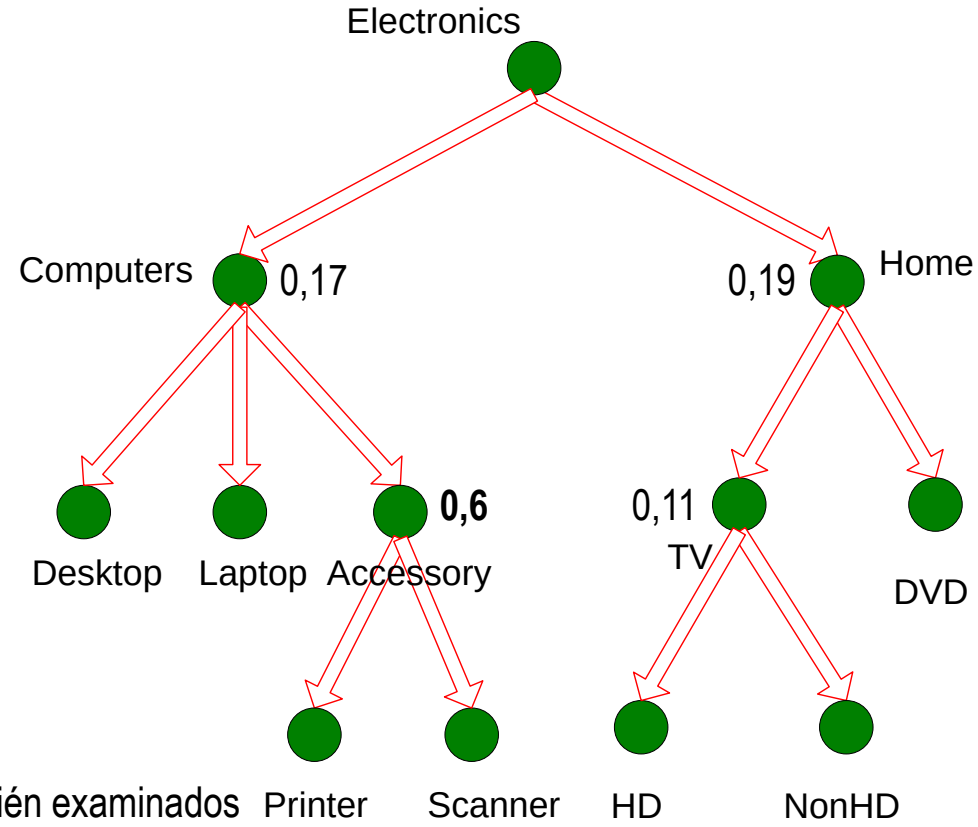
- **Controlled level-cross filtering by single item**
  - Se añade a cada nivel un umbral llamado level passage. Si el itemset tiene un soporte mayor que este umbral es examinado aunque tenga algún antecesor que no supere el mínimo soporte de su nivel

Nivel 1: min_sop=0,15

Level passage 1: min_sop=0,07

Nivel 2: min_sop=0,07

Level passage 2: min_sop=0,05



(Printer, HD) o (Printer, Scanner) son también examinados

Ejemplo: Reglas Jerárquicas

- Una vez localizados los itemsets frecuentes, generamos todas las reglas de asociación mezclando los itemsets frecuentes obtenidos de todos los niveles. Por ejemplo:

  Computer → Home (Itemsets del mismo nivel)
  Computer → Scanner (Itemsets de distinto nivel)

- Puede ocurrir que aparezcan muchas reglas **redundantes** debido a la relación de antecesor existente. Por ejemplo:

  Laptop computer → InkJet Printer

  HP Laptop Computer → InkJet Printer

Contenidos

- Problemas de Interpretabilidad
- Evaluación: Medidas de Interés
- Interpretaciones
- Reglas de Asociación Difusas
- Aspectos algorítmicos. El caso de las reglas de asociación jerárquicas
- Evaluación de reglas por grupos

Análisis de reglas por grupos

- El análisis de las reglas de asociación suele realizarse de forma individual, estudiando su novedad y potencial utilidad en base a los itemsets que la componen, las medidas objetivas y subjetivas realizadas sobre ellas, y el conocimiento previo del experto.
- Sin embargo, el análisis de conjuntos de reglas definidos según ciertos criterios puede proporcionar más información, con ciertas ventajas:
  - Permite descartar reglas que por si solas parecen relevantes, pero realmente no lo son.
  - Puede permitir discernir la semántica de una regla (ej: causalidad).
  - Si no hay excesivo solapamiento entre grupos de reglas y/o los grupos contienen bastantes reglas, el número de grupos puede ser menor que el de reglas.
  - Permiten definer otros tipos de patrones de conocimiento útiles como grupos de reglas que satisfacen ciertas condiciones.

Análisis de reglas por grupos

- Ejemplos: estudiar simultáneamente
  - $A \rightarrow C$ con $A \rightarrow \neg C$
  - $A \rightarrow C$ ó $\neg C \rightarrow \neg A$ (reglas muy fuertes)
  - Añadir $\neg A \rightarrow C$, $A \rightarrow \neg C$, $\neg A \rightarrow \neg B$, etc.
- Combinar lo anterior con el descarte de causalidad
  - Caso de que se confirmen $A \rightarrow C$, $A \rightarrow D$, $C \rightarrow D$ (se descarta o es dudosa la causalidad de C a D).
- Estudio de la evolución de conjuntos de reglas obtenidos a lo largo del tiempo
- Muchas más alternativas. Veamos tres ejemplos:

Análisis de reglas por grupos

- Excepciones: grupo de reglas de la forma
  - $A \rightarrow C$ es fuerte
  - $A, B \rightarrow \neg C$ tiene alto cumplimiento (aunque bajo soporte)
- B "caracteriza" excepciones a una regla general. Ejemplo: la ``aplicación de antibiótico" (A) se asocia con la ``mejoría del paciente" (C), salvo en aquellos casos en que aparecen estafilococos (B).
- Anomalías: grupo de reglas de la forma
  - $A \rightarrow C$ es fuerte
  - $A, \neg C \rightarrow B$ tiene alto cumplimiento (aunque bajo soporte)
  - $A, C \rightarrow \neg B$ tiene alto cumplimiento (aunque bajo soporte)
- B caracteriza alternativas a la asociación normal para A. Ejemplo: Si un paciente tiene el síntoma A, habitualmente tiene la enfermedad C ; cuando no es así, tiene la enfermedad B.

Análisis de reglas por grupos

- Clasificadores formados por conjuntos de reglas: se trata de seleccionar conjuntos de reglas obtenidas y estructurarlas de forma que permitan construir un clasificador.
- Muchos enfoques en la literatura.

```
P30 = A : TYPE = N (473|62)
P30 = C : TYPE = N (441|24)
P30 = T : TYPE = N (447|57)
clsc
  P28 = A and P32 = T : TYPE = EI (235|33)
  P28 = G and P32 = T : TYPE = EI (130|20)
  P28 = C and P32 = A : TYPE = IE (160|31)
  P28 = C and P32 = C : TYPE = IE (167|35)
  P28 = C and P32 = G : TYPE = IE (179|36)
  else
    P28 = A : TYPE = N (106|14)
    P28 = G : TYPE = N (94|4)
    clsc
      P29 = C and P31 = G : TYPE = EI (40|5)
      P29 = A and P31 = A : TYPE = IE (86|4)
      P29 = A and P31 = C : TYPE = IE (61|4)
      P29 = A and P31 = T : TYPE = IE (39|1)
      else
        P25 = A and P35 = G : TYPE = EI (54|5)
        P25 = G and P35 = G : TYPE = EI (63|7)
        clsc
          P23 = G and P35 = G : TYPE = EI (40|8)
          P23 = T and P35 = C : TYPE = IE (37|7)
          clsc
            P21 = G and P34 = A : TYPE = EI (41|5)
            else
              P28 = T and P29 = A : TYPE = IE (66|8)
              else
                P31 = G and P33 = A : TYPE = EI (62|9)
                clsc
                  P28 = T : TYPE = N (49|6)
                  else
                    P24 = C and P29 = A : TYPE = IE (39|8)
                    clsc
                      TYPE = IE (66|39)
```