LA CONTRIBUCIÓN DE LAS REGLAS DE ASOCIACIÓN A LA MINERÍA DE DATOS

MARLY ESTER DE MOYA AMARIS

Ingeniera de Sistemas, candidata a Magister en Ingeniería de Sistemas de la Universidad Nacional de Colombia. Administradora del Sistema de Información Geográfica del Instituto Geográfico Agustín Codazzi.
me_demoya@igac.gov.co.

JORGE ENRIQUE RODRÍGUEZ RODRÍGUEZ

Ingeniero de Sistemas, Especialista en Diseño y Construcción de Soluciones Telemática, Especialista en Ingeniería del Software, candidato a Magíster en Ingeniería de Sistemas de la Universidad Nacional de Colombia. Profesor de tiempo completo adscrito a la Facultad Tecnológica de la Universidad Distrital Francisco José de Caldas (F.J.C.)
jrodri@udistrital.edu.co.

Clasificación del artículo: Reflexión
Fecha de recepción: agosto 27 de 2003
Fecha de aceptación: diciembre 04 de 2003

Palabras clave: Minería de datos, reglas de asociación, algoritmos a priori, bases de datos, bodegas de datos

Key words: Data mining, association rules, a priori algorithms, data bases, data warehouses

Resumen

El artículo hace parte del proyecto de investigación «Desarrollo de Herramientas para la Minería de Datos - UDMiner»; en este documento los autores se centran en el estudio de una de las técnicas utilizadas para aplicar minería de datos en grandes volúmenes de información: las *reglas de asociación*. Se inicia con una breve introducción a la minería de datos y las reglas de asociación, luego se aborda un ejemplo para explicar cómo se generan estas reglas, y se profundiza sobre reglas de asociación simples y reglas de asociación multinivel. Por último, se presentan algunas conclusiones, producto del avance de la investigación y de la reflexión acerca de sus potencialidades en el campo.

Abstract

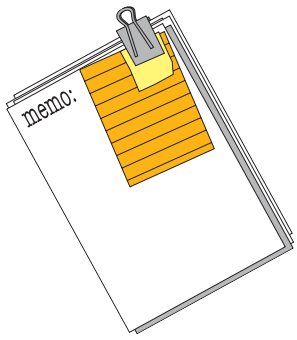
TThe article makes part of the investigation project: "Development of tools for the data mining – UDMiner"; in this document, the authors concentrate on the study of one of the techniques used to apply the data mining to high information volumes: *the association rules*. It starts with a short introduction to the data mining and the association rules, then an example is presented to explain how these rules are

generated and deepens on simple association rules and multilevel association rules. Lastly, some conclusions are presented as a consequence of the research advance and the reflection about its potentialities in the field.

1. Introducción

La minería de datos (Han, 2000), entendida como la extracción de patrones y reglas significativas de una gran cantidad de información, es útil en cualquier campo en el cual exista una importante cantidad de datos y algo valioso que aprender. Podría ser sorpresivo saber, por ejemplo, que las organizaciones de inteligencia militar emplean este tipo de técnicas para procesar grandes cantidades de imágenes satelitales con el propósito de clasificar objetivos visibles en tierra como tanques o tractores.

En el contexto de negocios aplica la misma regla: su utilidad en cualquier campo donde existan grandes cantidades de datos y un conocimiento potencial (Imielinski, 1996). En este tipo de actividades hay una definición explícita *«algo se considera valioso si el resultado de su conocimiento aporta más dinero que lo que costó su descubrimiento»*; de manera más estricta puede decirse que *«el conocimiento de algo es valioso si el retorno de la inversión requerida para su aprendizaje es mayor que el retorno de utilizar estos fondos en otro tipo de inversión»*. Así, mientras que en el contexto académico se considera «conocimiento» la posesión de los valores intrínsecos de algo, independiente de cualquier aplicación práctica, en el campo de los negocios el conocimiento puede ser valioso de dos formas: si aumenta la utilidad por medio de la disminución de los costos o del incremento de los ingresos. El conocimiento también puede tener valor cuando aumenta el precio en bodega a partir de la promesa de un incremento futuro de la utilidad, por la vía de algunos de los dos mecanismos anteriores.



Pueden identificarse cuatro campos fundamentales de trabajo en los que se está comenzando a emplear la minería de datos en los negocios:

- Como herramienta de investigación
- En el proceso de mejoramiento del negocio
- En mercadotecnia
- En la administración de las relaciones con los clientes.

Al respecto surge una pregunta importante: ¿qué papel juegan las reglas de asociación en el contexto de minería de datos? Las reglas de asociación se derivan de un tipo de análisis que extrae información por coincidencias (Agrawal, 1993); el análisis permite descubrir correlaciones o coocurrencias en los sucesos de la base de datos por analizar y se formaliza en la obtención de reglas de tipo sí ... entonces ..., las cuales se convierten en un importante punto de apoyo para el descubrimiento de conocimiento a partir de la información analizada (Piatetsky-Shapiro, 1996).

Un ejemplo común del uso de las reglas de asociación es el análisis de la canasta familiar adquirida por los compradores, el cual es utilizado para encontrar asociaciones entre los diferentes productos comprados por los clientes; el descubrimiento de tales asociaciones puede ayudar a desarrollar estrategias de mercadeo orientadas al aumento de las ventas.

Interrogantes adicionales surgen con respecto a la obtención y aplicación de este tipo de reglas: ¿cómo pueden encontrarse reglas de asociación

de grandes cantidades de datos, cuando estos son de transacciones o relaciones diferentes? ¿qué reglas de asociación son las más interesantes? ¿cómo puede ayudarse o guiarse el proceso de minería de datos para descubrir asociaciones interesantes? ¿qué herramientas y/o lenguajes de desarrollo son útiles en la definición de lenguajes de consultas de reglas de asociación para la minería de datos? En este trabajo se presenta un estudio detallado de los conceptos involucrados en reglas de asociación, dando respuestas a los anteriores interrogantes.

2. ¿Qué es una regla de asociación?

La Gerencia de Mercadeo de DISTCOL[1], desea dar respuesta al siguiente interrogante: ¿cuáles son los productos que más compran sus clientes? Para responder a esta pregunta puede hacerse un estudio de mercado acerca de los productos que los clientes compran al detal, empleando sus resultados para adoptar un plan de mercadeo o plantear una estrategia, por ejemplo, diseñar el catálogo de electrodomésticos que la compañía ofrece.

En forma premeditada, los productos que frecuentemente compran los clientes pueden ser ubicados en sitios algo distantes, mientras que aquellos productos que se venden con poca frecuencia se localizarán en sitios más próximos a la vista de los clientes. Si los clientes compran computadores y a la vez se inclinan por la compra de software financiero, entonces el sitio donde se ubique el hardware puede ayudar a incrementar las ventas del software: los clientes que escojan un computador observarán en la vitrina un software contable que despierte también su interés.

La información referente a los productos que compran los clientes, puede ser representada a partir de una regla de asociación, como sigue:

Si compra computador *entonces* compra software contable
[Soporte = 2%, Confianza = 60%] (Regla 1)

El soporte y la confianza son dos criterios de medida interesantes que reflejan, respectivamente, la utilidad y certeza de la regla. Un soporte del 2% indica que este porcentaje de todas las transacciones bajo análisis muestran que el computador y el software contable son comprados conjuntamente. Una confianza del 60% muestra que este porcentaje de los clientes que compran computadores adquieren también software contable. Las reglas de asociación son apropiadas si satisfacen el valor del mínimo soporte (*min_sop*) y de la mínima confianza (min_conf).

El soporte y la confianza para la regla *si* A *entonces* B está dada por:

$$\text{soporte}\ (A \Rightarrow B) = P(A \cup B) \qquad \text{(Ecuación 1)}$$

$$\text{confianza}(A \Rightarrow B) = P(B \mid A) = \text{soporte}\ (A \cup B\ ) / \text{soporte}(A) \qquad \text{(Ecuación 2)}$$

El interés debe centrarse en el descubrimiento de reglas que tienen mucho soporte; por lo tanto, independientemente de donde surjan, se buscan pares atributo-valor que cubran gran cantidad de instancias. Ellos se conocen como *ítemsets*, y cada par atributo-valor como *ítem*. Si un *ítemset* satisface el *min_sop*, entonces se le llama *ítemset* frecuente.

La respuesta a la pregunta: ¿cómo se hallan reglas de asociación para la minería de datos?, se encuentra en la ejecución de dos pasos:

[1] *DISTCOL de Colombia es una empresa cuya misión es la comercialización de electrodomésticos; se encuentra ubicada en la ciudad de Tunja (Colombia).*

- Hallar todos los *ítemsets* frecuentes (Agrawal, 2000)
- Generar los valores de las reglas de asociación a partir de los *ítemsets* frecuentes.

Las reglas de asociación pueden ser clasificadas en varios grupos, de acuerdo con los siguientes criterios:

- *Con base en los tipos de valores que manejan las reglas:* si una regla se refiere a asociaciones entre la presencia o ausencia de un ítem, se trata de una regla de asociación booleana. La Regla 1, presentada anteriormente, es de este tipo.

Si una regla describe asociaciones entre ítems cuantitativos o atributos, entonces se dice que es una regla de asociación cuantitativa. En ella los *ítems* o atributos son divididos en intervalos. El siguiente es un ejemplo de una regla de asociación cuantitativa, donde X es una variable que representa a un cliente (Pedro, Jorge, Diana, ...):

edad(X, «20...30») ∧ ingreso(X, «$500.000 ..... $1.000.000»⇒ (compra(X, computadores)[2]
(Regla 2)

Nótese que los atributos cuantitativos edad e ingreso tienen valores discretos.

- *Con base en las dimensiones de datos que involucra una regla:* si en los ítems o atributos de una regla de asociación se referencia solo en una dimensión, se trata de una regla de asociación de dimensión simple. Al reescribir la Regla 1 se tiene:

compra(X, «computador»⇒ compra(X, «software contable») (Regla 3)

La Regla 1 es de dimensión simple porque referencia solo una dimensión: la compra. Si se referencian dos o más dimensiones se dice que es una regla de asociación multidimensional, por ejemplo la Regla 3.

- *Con base en los niveles de abstracción que involucra la regla:* algunos métodos para encontrar reglas de asociación pueden descubrir reglas con diferentes niveles de abstracción. Por ejemplo, supóngase que en un conjunto de reglas de asociación se tiene:

edad(X, «20...30») ⇒ compra(X, «computadores portátil») (Regla 4)
edad(X, «20...30») ⇒ compra(X, «computadores») (Regla 5)

. . . . . .

[2] La representación de la regla se cambia por simplicidad en la notación.

En las reglas 4 y 5 los *ítems* son referenciados a diferentes niveles de abstracción («computador» es un nivel más alto de abstracción que «computador portátil»).

- *Con base en varias extensiones para la asociación minera:* la asociación minera puede extenderse al análisis de correlación, donde ésta puede ser identificada por la presencia o ausencia de un ítem.

3. Reglas de asociación simples en bases de datos transaccionales

A continuación se profundizará en el estudio de reglas de asociación booleanas de una sola dimensión, describiendo algunos métodos a través de un ejemplo:

3.1 El algoritmo a priori (Agrawal, 1994)

Busca *ítemsets* frecuentes usando generación de candidatos. Su nombre se debe a que usa conocimiento a priori para la generación de *ítemsets* frecuentes. Este algoritmo se resume en dos pasos:

i) Generación de todos los *ítemsets* que contienen un solo elemento, utilización de estos para generar *ítemsets* que contengan dos elementos, y así sucesivamente. Se toman todos los posibles pares de *ítems* que cumplen con las medidas mínimas de soporte inicialmente preestablecidas; esto permite ir eliminando posibles combinaciones: aquellas que no cumplan con los requerimientos de soporte no entrarán en el análisis.

ii) Generación de las reglas revisando que cumplan con el criterio mínimo de confianza. Es interesante observar que si una conjunción de consecuentes de una regla cumple con los niveles mínimos de soporte y confianza, sus subconjuntos (consecuentes) también los cumplen; en el caso contrario, si algún ítem no los cumple no tiene caso considerar sus superconjuntos.

Así se obtiene un método para construir reglas con un solo consecuente, a partir de ellas construir reglas de dos consecuentes y así sucesivamente; todo se realiza mediante una pasada por la base de datos para cada conjunto de *ítems* de diferente tamaño. El esfuerzo computacional depende principalmente de la cobertura mínima requerida, y se lleva prácticamente todo en el primer paso. El proceso de iteración del primer paso se llama *level-wise* y va considerando los superconjuntos nivel por nivel. De esta manera se tiene una propiedad anti-monótona: si un conjunto de ítems no pasa la prueba de soporte ninguno de sus subconjuntos la pasa; esto se aprovecha en la construcción de candidatos, para no considerar todas las opciones.

En la Tabla 1 se presenta un ejemplo de datos de compra de productos.

Transacción	Listado de productos adquiridos
T1	Computador, impresora
T2	Impresora, DVD
T3	Impresora, cámara de vídeo
T4	Computador, impresora, DVD
T5	Computador, cámara de vídeo
T6	Impresora, cámara de vídeo
T7	Computador, cámara de vídeo
T8	Computador, impresora, cámara de vídeo, scanner
T9	Computador, impresora, cámara de vídeo
T10	Impresora, scanner
T11	Computador, DVD
T12	Computador, impresora, DVD

Tabla 1. Un ejemplo de datos de compras de productos

Cada transacción (T1, T2,...) representa una compra realizada por diferentes clientes de DISTCOL; al frente aparecen los productos comprados en cada transacción; por simplicidad, a cada transacción se asigna un identificador. A continuación se explica el algoritmo a priori con base en el contenido de la Tabla 1, que registra 12 transacciones de venta por parte de la empresa.

En la primera iteración del algoritmo, cada *ítem* es un miembro del conjunto de candidatos 1-*ítemsets,* C1. El algoritmo explora en orden to-

das las transacciones para contar el número de ocurrencias de cada *ítem*.

C_1	
Ítemset	**Contador de soporte (cont_sop)**
{computador}	8
{impresora}	9
{cámara de vídeo}	6
{DVD}	4
{scanner}	2

Tabla 2. Contador de soporte para cada producto

Supóngase que el contador de soporte requerido son dos elementos (min_sop = 2/12 = 16%). El conjunto de los 1-*ítemsets* frecuentes, L1, puede entonces ser determinado considerando los 1-*ítemsets* candidatos que satisfacen este mínimo soporte (min_sop).

L_1	
Ítemset	**Contador de soporte (cont_sop)**
{computador}	8
{impresora}	9
{cámara de vídeo}	6
{DVD}	4
{scanner}	2

Tabla 3. Comparación de los candidatos del contador de soporte con el min_sop

Para descubrir el conjunto de los 2-ítemsets frecuentes L2, el algoritmo a priori usa L1 * L2 para generar un conjunto de candidatos de 2-ítemsets, C2.

C_2
Ítemset
{computador, impresora}
{computador, cámara de video}
{computador, DVD}
{computador, scanner}
{impresora, cámara de video}
{impresora, DVD}
{impresora, scanner}
{cámara de video, DVD}
{cámara de video, scanner}
{DVD, scanner}

Tabla 4. Generación de C2 candidatos desde L1

Las transacciones son exploradas y el contador de soporte de cada *ítemset* candidato en C2 es acumulado, como se muestra en la Tabla 5.

C_2	
Ítemset	**cont_sop**
{computador, impresora}	5
{computador, cámara de vídeo}	4
{computador, DVD}	3
{computador, scanner}	1
{impresora, cámara de vídeo}	4
{impresora, DVD}	3
{impresora, scanner}	2
{cámara de vídeo, DVD}	0
{cámara de vídeo, scanner}	1
{DVD, scanner}	0

Tabla 5. Contador de soporte para 2-ítemsets

El conjunto de los 2-*ítemsets* frecuentes, L2, es entonces determinado por aquellos que cumplan con el mínimo soporte.

L_2	
Ítemset	**cont_sop**
{computador, impresora}	5
{computador, cámara de vídeo}	4
{computador, DVD}	3
{impresora, cámara de vídeo}	4
{impresora, DVD}	3
{impresora, scanner}	2

Tabla 6. Candidatos que cumplen con el min_sop

La generación del conjunto de candidatos de 3-*ítemsets,* C3, se muestra en la Tabla 7.

C_3
Ítemset
{computador, impresora, cámara de vídeo}
{computador, impresora, DVD}

Tabla 7. Generar C3 candidatos desde L2

Las transacciones son exploradas y el contador de soporte de cada *ítemset* candidato en C3 es acumulado, como se muestra en la Tabla 8; posteriormente, en la Tabla 9 se muestran los ítemsets que cumplen con el min_sop.

C_3	
Ítemset	**cont_sop**
{computador, impresora, cámara de vídeo}	2
{computador, impresora, DVD}	2

Tabla 8. Contador de soporte para 3-ítemsets

L_3	
Ítemset	**cont_sop**
{computador, impresora, cámara de vídeo}	2
{computador, impresora, DVD}	2

Tabla 9. Candidatos que cumple con el min_sop

El algoritmo usa L_3*L_3 para generar un conjunto de candidatos de 4-*ítemsets,* C4; sin embargo, no existen candidatos de 4-ítemsets que cumplan con el min_sop, por lo cual el algoritmo a priori termina. A continuación se presenta el algoritmo a priori en seudo-código. (Fuente tomada y adaptada de Han, 2000).

Algoritmo: **A priori.** *Encuentra los ítemsets frecuentes.*
Entradas: Transacciones de una base de datos D; min_sop
Salidas: L, ítemsets frecuentes de la base de datos D.
Método:
(1) L1 = encontrar_1-ítemsets_frecuente(D);
(2) Para (k=2; L_{k-1}≠∅; k++)
(3) C_k = generar_apriori(L_{k-1}, min_sop);
(4) Para cada transacción t ∈ D // examinar D para el contador
(5) C_t = subconjuntos(C_k) // obtener los subconjuntos t que son candidatos
(6) Para cada candidato c ∈ C_t
(7) c.contador++
(8) fin-para
(9) L_k = {c ε C_k | c.contador≥min_sop}
(10) fin-para
(11) retornar L = $U_k L_k$;

función generar_apriori(L_{k-1} : (k-1)-ítemsets frecuentes; min_sop : mínimo soporte)
(1) Para cada ítemset $l_1 \in L_{k-1}$
(2) Para cada ítemset $l_2 \in L_{k-1}$
(3) Si ($l_1[1] = l_2[1] \wedge (l_1[2] = l_2[2]) \wedge ... \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1])$)
Entonces
(4) c = l_1 X l_2; //generar candidatos
(5) Si subconjunto_infrecuente(c, L_{k-1}) Entonces
(6) eliminar c;
(7) Si No
(8) adicionar c a C_k
(9) fin-para
(10) retornar C_k;

función subconjunto_infrecuente(c : k-ítemset candidato; L_{k-1} : (k-1)-ítemset frecuente)
(1) Para cada (k-1)-subconjunto s de c
(2) Si (s ∉ L_{k-1}) Entonces
(3) retornar Verdadero;
(4) retornar Falso;

3.2 Generación de reglas de asociación desde ítemsets frecuentes

Una vez se tienen los conjuntos de ítems, generar las reglas es relativamente sencillo. Para cada conjunto l de ítems se generan todos sus subconjuntos. Para cada subconjunto s (l, genera una regla: s ⇒ (l - s) si:

$$\frac{soporte(l)}{soporte\ (s)} \geq nivel\ \_\ confianza$$

3.3 Algunas mejoras

Se han realizado algunas mejoras al algoritmo básico de reglas de asociación (a priori) para hacerlo más eficiente:

i) Usar tablas *hash* para reducir el tamaño de los candidatos de los ítemsets (Park, 1995)
ii) Eliminar transacciones (elementos en la base de datos) que no contribuyan en superconjuntos a considerar (Agrawal, 1994)
iii) Dividir las transacciones en particiones disjuntas, evaluar *ítemsets* locales y luego, con base en sus resultados, estimar los globales (Savasere, 1995)
iv) Hacer aproximaciones con muestreos en la lista de productos para no tener que leer todos los datos
v) Evitar generar candidatos usando estructuras de datos alternativas, como por ejemplo, los *FP-trees (Frequent Pattern tree)*.

3.4 Algunas extensiones

Dentro de ellas pueden citarse:

i) Encontrar reglas de asociación a diferentes niveles de abstracción. Normalmente se empieza con las clases superiores y los resultados sirven para filtrar clases inferiores. Por ejemplo, considerar reglas de asociación sobre computadoras e impresoras, y luego sobre DVD y cámaras de video, de una parte, y sobre impresoras láser y de punto, de otra. Pueden realizarse las siguientes acciones: a) considerar un criterio de soporte uniforme reduciéndolo para las subclases; b) considerar todas las subclases independientemente de este criterio tomando en cuenta el de una de las superclases de un ítem o k superclases de *k ítems;* c) considerar *ítems* cuyo nivel de soporte de sus padres no cumpla con el criterio de soporte, pero que sea mayor que cierto umbral. En este tipo de proceso es común generar reglas redundantes o que no dicen nada nuevo (v.gr., la regla más general ya decía lo mismo), por lo que es necesario incorporar mecanismos de filtrado.
ii) Encontrar reglas de asociación combinando información de múltiples tablas o reglas de asociación multidimensionales; estas últimas pueden encontrarse a partir de cubos de datos.
iii) Las reglas de asociación, al igual que los árboles de decisión y las reglas de clasificación, funcionan en su forma original, con atributos discretos. Al igual que en las otras técnicas, se han propuesto mecanismos para manejar atributos continuos. Los enfoques más comunes son: antes de explorar, discretizar en rangos usando jerarquías predefinidas, y durante el proceso discretizar dinámicamente tratando de maximizar algún criterio de confianza o reducción de longitud de reglas.

La ACRS (*Association Rule Clustering System)* (Lent, 1997), por ejemplo, mapea atributos cuantitativos a una rejilla y luego utiliza técnicas de agrupación. Primero asigna datos a *contenedores* delimitados por rangos, que después pueden cambiar. Los esquemas más comunes de contenedores son: del mismo tamaño, con el mismo número de elementos, y con elementos uniformemente distribuidos. Posteriormente, utilizando los contenedores, se encuentran reglas de asociación; una vez se tienen estas se agrupan, si forman rectángulos más grandes dentro de la rejilla. Luego se discretizan utilizando información semántica y se forman grupos con elementos cercanos (posiblemente haciendo agrupación sobre los atributos), para luego encontrar las reglas de asociación con esos grupos basados en distancias o similaridades.

4. Reglas de asociación multinivel en bases de datos transaccionales

En esta sección se estudiarán métodos para realizar minería de datos con reglas de asociación multinivel (Han, 1995), es decir, reglas que involucran *ítems* con diferentes niveles de abstracción. También se analizarán métodos para verificar multiniveles redundantes.

4.1 Reglas de asociación multinivel

Para muchas aplicaciones es difícil encontrar asociaciones fuertes entre los ítems de datos a un nivel primitivo de abstracción, debido a su dispersión en espacios multidimensionales; asimismo, asociaciones fuertes descubiertas en altos niveles conceptuales pueden representar conocimiento de sentido común. Sin embargo, lo que puede representar sentido común para un usuario puede no serlo para otro, por lo cual los sistemas de minería de datos deberían tener capacidades para realizar su función a diferentes niveles de abstracción. El concepto de jerarquía define una secuencia de diagramas, iniciando en un conjunto de conceptos de bajo nivel hasta llegar a conceptos de alto nivel; en otras palabras, se inicia con conceptos específicos en el nivel más bajo, y se va subiendo en jerarquía a conceptos más generales. La Figura 1 representa un ejemplo.

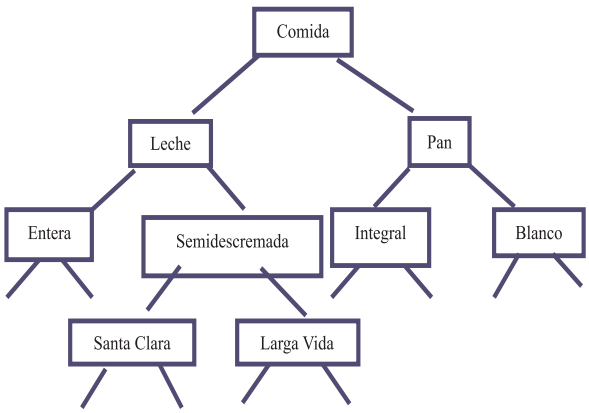


Figura 1. El concepto de jerarquía a través de un ejemplo

El nivel más alto de la jerarquía en la Figura 1 corresponde a «comida», de la cual se desprenden dos ramas: «leche» y «pan»; si se desciende en jerarquía a partir de «leche» se encuentran dos clases: «entera» y «semidescremada», y así sucesivamente. Así puede apreciarse la denominada conceptualización jerárquica. En minería de datos, las reglas de asociación generadas con conceptos jerárquicos son llamadas de niveles múltiples o reglas de asociación multinivel, dado que consideran más de un nivel conceptual.

4.2 Minería de datos con reglas de asociación multinivel

¿Cómo pueden generarse eficientemente reglas de asociación multinivel usando conceptos jerárquicos?. En términos generales, siempre se utiliza una estrategia top-down (arriba-abajo); los conteos son realizados iniciando en el nivel conceptual 1 y luego se va disminuyendo en nivel, hasta que ya no se encuentren ítemsets frecuentes. Para cada nivel puede utilizarse cualquier algoritmo de descubrimiento de ítemsets frecuentes, como el algoritmo a priori o alguna de sus variaciones.

Para la generación de reglas de asociación multinivel existen básicamente dos métodos que se describen a continuación.

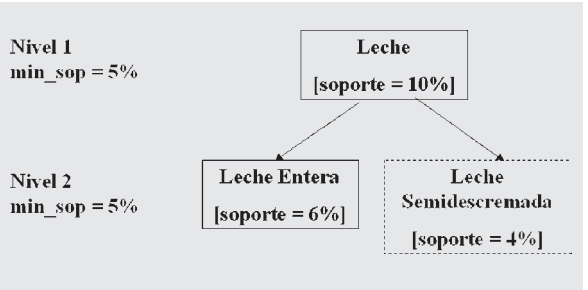


Figura 2. Minería multinivel con un soporte uniforme

i) Utilizando el mismo soporte mínimo para todos los niveles (soporte uniforme). Este soporte (ver Figura 2) es empleado cuando se realiza minería en cada nivel de abstracción, simplificando de esta manera el procedimien-

to de búsqueda. Sin embargo, el método tiene algunas dificultades: es improbable que los ítems en un nivel de abstracción inferior ocurran tan frecuentemente como aquellos que están en un nivel de abstracción superior. Si el mínimo soporte establecido es muy alto podrían perderse muchas asociaciones significativas en niveles inferiores de abstracción; si es muy bajo podrían generarse muchas asociaciones poco interesantes en niveles de abstracción más altos. Este problema motiva la utilización del siguiente método.

ii) Utilizando soportes mínimos más reducidos en niveles de abstracción inferiores. Cada nivel de abstracción tiene su propio soporte mínimo (ver Figura 3), por lo cual el nivel más bajo tendrá el soporte más pequeño de toda la jerarquía. Existen muchas estrategias para implementar este método, entre ellas:

• Nivel por nivel independiente: se trata de una búsqueda total en amplitud en la cual el conocimiento previo no es utilizado, es decir, cada nodo es examinado sin importar si su padre es un ítemset frecuente o no.

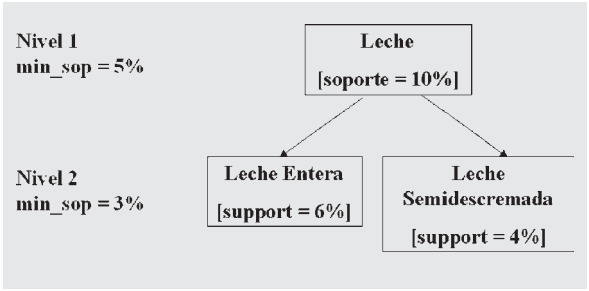


Figura 3. Minería multinivel con un soporte reducido

• Filtrado de niveles por un único ítem (Figura 4): un ítem es examinado sí y solo sí su padre es un ítem frecuente. En otras palabras, si un nodo es frecuente su hijos serán examinados; de otro modo sus descendientes serán eliminados de la búsqueda.

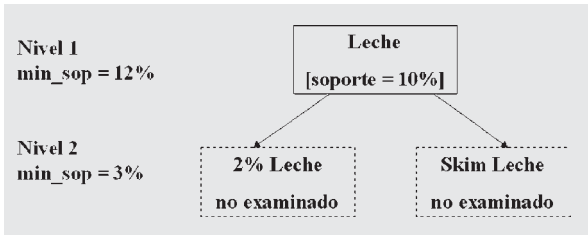


Figura 4. Minería multinivel con soporte reducido usando niveles por un ítem simple

• Filtrado de niveles por un k-*ítemset* (Figura 5): un k-*ítemset* es examinado sí y solo sí su correspondiente k-*ítemset* padre es frecuente.

La estrategia nivel por nivel independiente puede llevar a examinar numerosos *ítems* poco frecuentes en niveles inferiores, encontrando asociaciones entre *ítems* de muy poca importancia. La estrategia de filtrado de niveles por un k-*ítemset* permite examinar únicamente los hijos de k-*ítemsets* frecuentes; esta restricción es muy fuerte en conjuntos de datos donde no existen muchos k-*ítemsets*, e implica que patrones importantes puedan ser eliminados del análisis usando esta técnica. Por su parte, la estrategia de filtrado de niveles por un único *ítem* representa un compromiso entre los dos extremos (Figura 6); sin embargo, este último método puede perder asociaciones entre *ítems* de niveles por la reducción del soporte mínimo, cuando sus padres o ancestros no satisfagan el soporte mínimo.

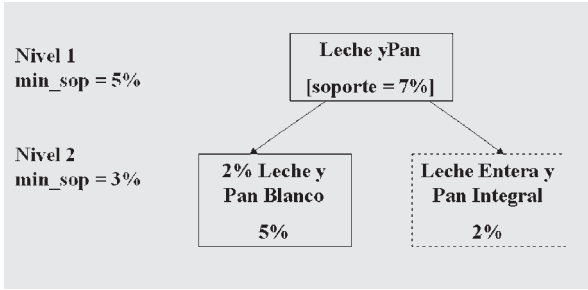


Figura 5. Minería multinivel con soporte reducido, usando niveles por un K-ítemset, con k=2

Una versión modificada de esta última estrategia, conocida como filtrado controlado de niveles por un único *ítem*, permite establecer un umbral de paso de niveles que es utilizado para bajar de un nivel a otro. En otras palabras, el método alternativo permite que los hijos de *ítems* que no cumplen con el soporte mínimo sean examinados, si estos cumplen con el umbral de paso de niveles.

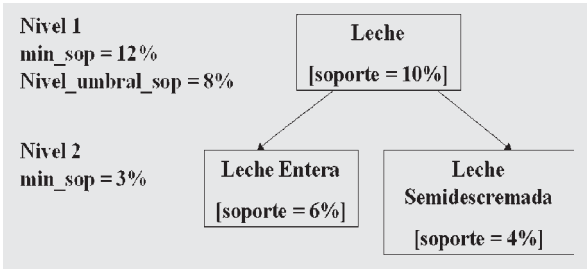


Figura 6. Minería multinivel con control de K-ítemsets de un solo ítem

4.3 Verificación de reglas de asociación redundantes

En la minería de datos el concepto de jerarquías es útil si permite descubrir conocimiento a diferentes niveles de abstracción, como las reglas de asociación multinivel; sin embargo, cuando estas últimas se generan, algunas de ellas pueden ser redundantes debido a que sus ancestros ya han representado las relaciones entre los *ítems.*

5. Generación de reglas de asociación multidimensionales en bases de datos relacionales y en bodegas de datos

A continuación se estudiarán los métodos para generar reglas de asociación multidimensionales (Kamber, 1997), es decir, aquellas que involucran más de una dimensión o predicado. Estos métodos pueden ser organizados de acuerdo con su tratamiento de atributos cuantitativos.

5.1 Reglas de asociación multidimensionales

Hasta el momento se han estudiado reglas de asociación que implican un único predicado, en este caso «compras». Por ejemplo, si tenemos la regla de asociación Pan Blanco -> Leche Entera, también podemos escribirla como:

compra(X, «Pan Blanco «) -> compra(X, «Leche Entera») (Regla 6)

En la regla anterior, X es una variable que representa clientes. Empleando la terminología de bases de datos multidimensionales, cada predicado distinto en una regla es una dimensión, luego la Regla 6 es unidimensional o intradimensional.

Supóngase, sin embargo, que en lugar de utilizar bases de datos transaccionales se necesitan bases de datos relacionales o bodegas de datos, en las cuales los datos almacenados son multidimensionales. Por ejemplo, una base de datos relacional puede almacenar otros atributos asociados con los ítems, tales como cantidad, precio y marca. Además puede almacenar información adicional de los clientes, como su edad, dirección, enfermedades, etc. Considerando cada dimensión de la base de datos como un predicado, es posible generar reglas de asociación conteniendo múltiples predicados, tales como:

edad(X, «60...70»)^enfermedad(X, «diabetes») -> compra(X, «Pan Integral») (Regla 7)

Las reglas de asociación que involucran dos o más predicados son conocidas como *multidimensionales;* entre ellas, aquellas que no tienen predicados repetidos son llamadas reglas

de asociación *interdimensionales*, y las que los tienen son conocidas como *híbrido-dimensionales.*

En una base de datos un atributo puede ser categórico o cuantitativo. Los primeros tienen un número finito de posibles valores, sin tener necesariamente valores ordenados entre cantidades, por ejemplo ocupación, marca y color; ellos también son conocidos como nominales, debido a que sus valores son «nombres de cosas». Los atributos cuantitativos son numéricos y tienen un orden implícito de valores, por ejemplo, edad y precio.

Las técnicas para generar reglas de asociación multidimensionales pueden ser categorizadas de acuerdo con tres acercamientos básicos:

i) *Atributos cuantitativos discretizados usando conceptos jerárquicos predefinidos*. La discretización se realiza antes de comenzar el proceso de minería; por ejemplo, un concepto jerárquico para ingresos puede ser usado para reemplazar los valores numéricos originales de estos atributos por rangos, tales como «0...20k», «21k...30k», «31k...40k», y así sucesivamente; la discretización es entonces estática y predeterminada. La discretización numérica de atributos con sus rangos de valores puede ser tratada como atributos categóricos, considerando cada rango como una categoría.
ii) *Atributos cuatitativos discretizados en «conjuntos» basados en la distribución de los datos.* Este proceso de discretización es dinámico y establecido para satisfacer ciertos criterios de la minería, tales como la maximización de la confidencia de las reglas. Las reglas de asociación generadas con esta técnica son también llamadas *cuantitativas,* debido a que esta estrategia trata los valores numéricos de los atributos como cantidades, en lugar de emplear rangos o categorías predefinidas.
iii) *Atributos cuantitativos discretizados basados en la distancia entre puntos de los datos.* Se trata de una discretización dinámica que considera la distancia entre puntos de datos, por lo cual sus reglas de asociación se denominan *basadas en distancia.*

5.2 Reglas de asociación multidimensionales usando discretización estática de los atributos cuantitativos

En este caso los atributos cuantitativos son discretizados antes de comenzar el proceso de minería, utilizando conceptos jerárquicos predefinidos, en los cuales los valores numéricos son reemplazados por rangos; si se desea, los atributos categóricos también pueden ser generalizados a niveles conceptuales más altos. Si los datos resultantes son almacenados en una tabla relacional, entonces el algoritmo a priori requiere una modificación sencilla para encontrar todos los conjuntos de predicados frecuentes, en lugar de encontrar los *ítemsets* frecuentes; encontrar todos los conjuntos de k-predicados frecuentes, requerirá de k o k+1 búsquedas sobre la tabla. También pueden emplearse otras estrategias, como el *hashing* o el *particionamiento.*

Alternativamente, los datos transformados de tareas relevantes pueden ser almacenados en un cubo de datos; estos cubos se ajustan perfectamente al proceso de generación de reglas de asociación multidimensionales, debido a que ellos son multidimensionales por definición. La Figura 7 muestra un cubo de datos para las dimensiones edad, ingresos y compras.

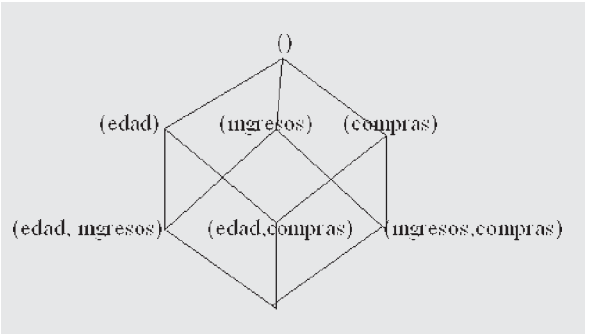


Figura 7. Cubo de Datos

Las celdas de un cubo n-dimensional son usadas para almacenar el conteo de soporte de los correspondientes conjuntos de n-predicados. Debido al creciente uso de las bodegas de datos y la tecnología OLAP, es posible que un cubo de datos que contenga las dimensiones de interés para el usuario, pueda realmente existir. Una estrategia similar a la empleada en el algoritmo *a priori* puede ser usada, basada en el hecho a priori que cada subconjunto de un conjunto de predicados frecuentes también debe ser frecuente; esta propiedad puede ser usada para reducir el número de conjuntos de predicados candidatos.

5.3 Reglas de asociación cuantitativas

En ellas los atributos numéricos son discretizados dinámicamente durante el proceso de minería, para satisfacer algunos criterios como la maximización de la confidencia. En esta sección se centrará el estudio en cómo generar reglas de asociación cuantitativas, teniendo dos atributos cuantitativos al lado izquierdo de cada regla y un atributo categórico en el lado derecho de ella:

$$Aquan_1 \ \& \ Aquan_2 \rightarrow Acat \quad \text{(Regla 8)}$$

Estas reglas de asociación son conocidas como *bidimensionales,* debido a que contienen dos dimensiones cuantitativas. Un ejemplo podría ser la asociación existente entre la edad, los ingresos de un consumidor y el tipo de televisor que este cliente compra:

edad(X, «30...29») & ingresos(X, «42k... 48k») -> compra(X, «TV alta resolución» (Regla 9)

A continuación se describirá un sistema llamado ARCS *(Association Rule Clustering System)*, el cual permite la generación de reglas bidimensionales. El sistema mapea parejas de atributos cuantitativos en una grilla 2D para registros que satisfagan la condición del atributo categórico; en esta grilla se buscan grupos de puntos, a fin de generar las reglas de asociación. Los pasos a seguir en este sistema son:

? Binning: realización de un proceso de particionamiento o de creación de intervalos, cada uno de los cuales se considera un *bin.* En el sistema ARCS todos los intervalos son de un mismo tamaño; el usuario introduce el que considere adecuado para cada atributo cuantitativo; cada celda del arreglo, contiene la correspondiente distribución de conteo para cada posible clase del atributo categórico de la regla ubicada al lado derecho. El mismo arreglo 2D puede ser usado para generar reglas para cualquier valor del atributo categórico, basado en los dos mismos atributos cuantitativos.

- Búsqueda de los conjuntos de *ítemsets* frecuentes: una vez que el arreglo 2D contiene la distribución de conteo para cada categoría, éste puede ser escaneado con el fin de encontrar los conjuntos de predicados frecuentes, que satisfacen la confidencia y el soporte mínimo.
- Agrupación de las reglas de asociación (ver Figura 8): las reglas de asociación fuertes, obtenidas en el paso anterior son ubicadas en la grilla 2D:

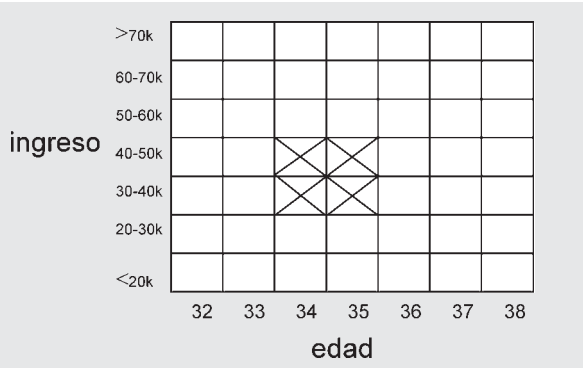


Figura 8. Una matriz 2D, representativa de los clientes que compran TV de alta resolución

Las reglas de asociación encontradas en el ejemplo anterior son:

edad(X,»34"), ?ingreso(X,»30K - 40K») —> compra(X,»high resolution TV»)
edad(X,»35"), ? ingreso(X,»30K - 40K») —> compra(X,»high resolution TV»)
edad(X,»34"), ? ingreso(X,»40K - 50K») —> compra(X,»high resolution TV»)
edad(X,»35"), ? ingreso(X,»40K - 50K») —> compra(X,»high resolution TV»)

Estas reglas pueden ser agrupadas en una sola, de tal forma que:

edad(X,»34-35"), ?ingreso(X,»30K - 50K») ->?compra(X,»high resolution TV»)

ARCS utiliza un algoritmo de agrupamiento para este propósito.

5.4 Generación de reglas de asociación basados en la distancia

En la sección anterior se describieron las reglas de asociación cuantitativas; en ellas los atributos cuantitativos son discretizados inicialmente por métodos de agrupamiento, y los intervalos resultantes son combinados. Sin embargo, esta técnica puede no capturar la semántica de los intervalos, debido a que no considera la distancia entre ellos. Considérese por ejemplo la siguiente tabla:

Transacción	Listado de productos adquiridos
T1	Computador, impresora
T2	Impresora, DVD
T3	Impresora, cámara de vídeo
T4	Computador, impresora, DVD
T5	Computador, cámara de vídeo
T6	Impresora, cámara de vídeo
T7	Computador, cámara de vídeo
T8	Computador, impresora, cámara de vídeo, scanner
T9	Computador, impresora, cámara de vídeo
T10	Impresora, scanner
T11	Computador, DVD
T12	Computador, impresora, DVD

Tabla 10. Una partición del atributo «precio» de acuerdo con los métodos binning

En la Tabla 10 se muestra el atributo «precio» particionado de acuerdo con las técnicas de *binning equidistancia* y *equicantidad*, versus el particionamiento basado en distancia. Este último es más intuitivo, debido a que agrupa los valores que están más cercanos entre sí en el mismo intervalo; por lo tanto este tipo de discretización es más significativa que las otras dos.

Un algoritmo de dos fases puede ser usado para generar reglas de asociación basadas en distancia. La primera fase emplea agrupamiento *(clustering)* para encontrar los intervalos o clusters; la segunda fase obtiene reglas de asociación basadas en distancia, mediante la búsqueda de grupos de cluster que ocurran juntos, frecuentemente.

En la primera fase se hace necesario definir un diámetro de medida, a fin de determinar los registros más cercanos. Sea S[X] un conjunto de N registros, $t_1, t_2, ...t_N$, proyectadas sobre el conjunto de atributos X; el diámetro de S[X] es el promedio de la distancia entre las registros proyectadas en X; la distancia puede ser la *euclidiana* o la de *Manhatan.*

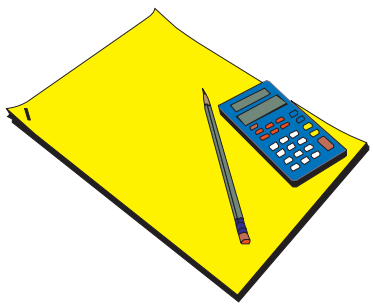
En la segunda fase, los grupos o *clusters* son combinados para forma reglas de asociación basadas en distancia. Considérese una regla de asociación basada en distancia de la forma C_X->C_Y; supóngase que X es el atributo edad y Y el atributo ingresos. Se quiere tener certeza de que la implicación entre el *cluster* C_X (edad) C_Y (ingresos) es fuerte; esto significa que cuando los registros de edad en C_X son proyectadas en el

atributo ingresos, sus valores correspondientes de ingresos estén dentro del cluster CY o cerca de él. Un grupo CX proyectado dentro del conjunto de atributos Y es denotado por $C_X[Y]$. En consecuencia, la distancia entre $C_X[Y]$ y $C_Y[Y]$ debe ser pequeña. La distancia mide el grado de asociación entre C_X y C_Y; entre más pequeña sea esta distancia más fuerte es su grado de asociación.

Los grupos o *clusters* pueden ser combinados para encontrar reglas de asociación de la forma: C_{X1} C_{X2}.... C_{XX} -> C_{Y1} C_{Y2}.... C_{YY}, donde Xi y Yj son pares de atributos de conjuntos disjuntos que cumplen las siguientes tres condiciones: a) cada grupo o *cluster* en la regla antecedente está fuertemente asociado con cada grupo o cluster en la regla consecuente; b) los grupos en el antecedente deben ocurrir colectivamente; c) los grupos en el consecuente deben ocurrir colectivamente.

6. Conclusiones

- La minería de datos ha surgido del potencial del análisis de grandes volúmenes de información, con el fin de obtener resúmenes y conocimiento que apoye la toma de decisiones y pueda construir una experiencia a partir de los millones de transacciones detalladas que registra una empresa en sus sistemas informáticos.
- La implementación de reglas de asociación es una interesante opción en el proceso de selección de la técnica a utilizar para realizar minería de datos; su aplicación es fundamental para el descubrimiento de relaciones de asociación en grandes cantidades de datos, siendo útil en la selección de una estrategia de mercadeo, procesos de toma de decisiones, manejo de negocios, etc.
- Actualmente existen numerosos algoritmos eficientes para encontrar reglas de asociación, pero la mayoría de ellos trabajan con bases de datos transaccionales o requieren que los dominios de los atributos de la base de datos sean discretos; no obstante, en el mundo real existen numerosas bases de datos en las cuales la información es numérica. La mayor parte de las herramientas que trabajan sobre dominios continuos simplemente se limitan a discretizarlos mediante alguna estrategia concreta que les permita tratarlos posteriormente como si fueran discretos. Sin embargo, los métodos de discretización desvirtúan, en cierta medida, el resultado final de las reglas obtenidas, porque durante el proceso de división de los dominios numéricos no tienen en cuenta algunas de las medidas indicadoras del interés de las reglas de asociación, entre otras el soporte y la confianza.
- La investigación en reglas de asociación hasta ahora está en sus inicios; existen muchos campos especializados que prometen grandes beneficios, tales como análisis de datos espaciales, datos multimediales y series de tiempo, entre otros. T&C



REFERENCIAS BIBLIOGRÁFICAS

[1] AGRAWAL, R. AGGARWAL, C. and PRASAD, V. V. V. (2000). A tree projection algorithm for generation of frequent itemsets. In Journal of Parallel and Distributed Computing (Special Issue on High Performance Data Mining)

[2] AGRAWAL, R. and SRIKANT, R. (1994). Fast algorithms for mining association rules. VLDB'94 487-499, Santiago, Chile

[3] AGRAWAL, R. IMIELINSKI, T. and SWAMI, A. (1993). Mining association rules between sets of items in large databases. SIGMOD'93, 207-216, Washington, D.C.

[4] HAN, J. and FU, Y. (1995). Discovery of multiple-level association rules from large databases. VLDB'95, 420-431, Zurich, Switzerland

[5] HAN, J. and KAMBER, M. (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann

[6] IMIELINSKI, T. and MANNILA, H. (1996). A database perspective on knowledge discovery. Communications of ACM, 39:58-64.

[7] KAMBER, M. HAN, J. and CHIANG J. Y. (1997). Metarule-guided mining of multi-dimensional association rules using data cubes. KDD'97, 207-210, Newport Beach, California.

[8] LENT, B. SWAMI, A. and WIDOM, J. (1997). Clustering association rules. ICDE'97, 220-231, Birmingham, England.

[9] PARK, J.S. CHEN, M. S. and Yu, P.S. (1995). An effective hash-based algorithm for mining association rules. SIGMOD'95, 175-186, San Jose, CA.

[10] PIATETSKY-SHAPIRO, G. FAYYAD, U. and SMITH, P. (1996). From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press.

[11] SAVASERE, A. OMIECINSKI, E. and NAVATHE, S. (1995). An efficient algorithm for mining association rules in large databases. VLDB'95, 432-443, Zurich, Switzerland