Instituto Politécnico Nacional
Escuela Superior de Cómputo
Secretaría Académica
Departamento de Ingeniería en Sistemas Computacionales

Minería de datos (*Data Mining*)
Reglas de Asociación

1

Profesora: *Dra. Fabiola Ocampo Botello*

2

Larose & Larose (2015) establecen que El análisis de afinidad se refiere al estudio o características que "van juntas". Los métodos para el análisis de afinidad son conocidos como "análisis de la canasta de mercado", la cual busca descubrir asociaciones entre los atributos, es decir, busca descubrir reglas para cuantificar la relación entre dos o más atributos. Las reglas de asociación son de la forma:

Si el antecedente entonces el consecuente.

Considerando una medida de soporte (*support*) y una medida de confianza (*confidence*).

3

Concepto	Significado
Algoritmo a priori *A priori algorithm*	En minería de datos, un algoritmo a priori para generar reglas de asociación considera la relación entre los elementos para generar tales reglas y con ello reducir el tamaño del espacio de análisis en un tamaño más manejable.
I	Conjunto de todos los ítems que se analizan.
Soporte o apoyo *Support*	Suponga que tiene la regla que relaciona dos ítems (A, B), A → B y n el número total de transacciones analizadas. Es el cociente resultante de la división de B/n. Ejemplo: Existen 1000 transacciones, 200 de las notas compraron pañales y de éstos 200, 50 de ellas también compraron cerveza. Support = 50/1000 = 5%
Confianza *Confidence*	Es el cociente del resultado de dividir B/A. En este ejemplo Confidence = 50/200 = 25%
Conjunto de ítems *Itemset*	Conjunto de ítems del conjunto I en una transacción.
Conjunto de ítems k k-itemset	Es el conjunto que contiene k elementos.

Elaboración propia con base en lo expuesto por Larose & Larose (2015)

4

Frecuencia del conjunto de ítems *Itemset frequency*	Es el número de transacciones que contienen un *itemset* en particular.
Conjunto de elementos frecuentes *Frequent itemset*	Es un itemset que se presenta al menos un cierto número de veces y que tienen un itemset frequency > fi (φ) Ejemplo: Suponga que fi (φ) = 4, entonces el itemset que ocurre más de cuatro veces se denomina frecuente (*frequent*).
Fk	Es el conjunto de k-itemset frecuentes.
Reglas estrictas *Strong rules*	Son aquellas que cumplen o superan ciertos criterios mínimos de soporte o apoyo y confianza

Elaboración propia con base en lo expuesto por Larose & Larose (2015)

5

Figura tomada de Tan, Steinbach & Kumar (2014).

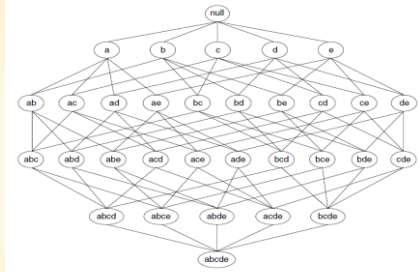

Figure 6.1. An itemset lattice.

6

Existen dos términos en las reglas de asociación (Han, Kamber & Pei, 2012):

- Soporte s, donde s es el porcentaje de transacciones en D que contienen AUB, la unión de los conjuntos de A y B, o ambos A y B.
- Confianza c, en el conjunto de transacciones D, donde c es el porcentaje de transacciones en D que contienen A y también a B.

Soporte $(A \Rightarrow B) = P(A \cap B)$

Confianza $(A \Rightarrow B) = P(B \mid A)$

7

Ejemplo de soporte:

El soporte (*support*) es la proporción de las transacciones que contienen al antecedente y al consecuente

Soporte $(A \Rightarrow B) = P(A \cap B)$

No. Transacción	p1	p2	p3	p4
1	1	1	0	0
2	1	1	1	0
3	1	1	0	1
4	1	0	0	0
5	1	0	0	1

t = {p1, p2} soporte(t) = 3/5
t = {p2, p3} soporte(t) = 1/5
t = {p3, p4} soporte(t) = 0/5 = 0

8

Ejemplo de confianza:

Confianza (c) en el conjunto de transacciones D, donde c es el porcentaje de transacciones en D que contienen A y también a B.

$$\text{Confianza } (A \Rightarrow B) = P(B \mid A) = \frac{support\_count(A \cap B)}{support\_count(A)}$$

t = {p1, p2} soporte(t) = 3/5 = 0.6 Confianza(t) = $\frac{3/5}{5/5} = \frac{0.6}{1} = 0.6$

t = {p2, p3} soporte(t) = 1/5 = 0.2 Confianza(t) = $\frac{1/5}{3/5} = \frac{0.2}{0.6} = 0.33$

t = {p3, p4} soporte(t) = 0/5 = 0

9

Las reglas que satisfacen un umbral de soporte mínimo (min_sup) y un umbral de confianza mínima (min_conf) son llamadas fuertes (Han, Kamber & Pei, 2012).

k-itemset es un itemset que contiene k ítems.

Algoritmo A priori, es un algoritmo básico para encontrar conjuntos de elementos (itemset) frecuentes (Han, Kamber & Pei, 2012).

Condición inicial:
Se establece un umbral fi (φ), para que el algoritmo a priori identifique subconjuntos de elementos que contienen φ elementos (Larose & Larose, 2015).

Propiedad a priori:
Si un conjunto de elementos Z no es frecuente, entonces para cualquier elemento A, Z U A no será frecuente (Larose & Larose, 2015).

10

Para crear el conjunto de elementos L_k se utiliza L_{k-1}.
Se utilizan dos pasos: join y poda (Han, Kamber & Pei, 2012:249):

1. El paso de join. Para encontrar Lk, un conjunto candidato de k-itemsets se genera a partir de aplicar la operación de join L_{k-1} consigo misma.
2. El paso de poda. Ck es un superconjunto de Lk, sus miembros podrían o no ser frecuentes, todos los k-itemsets frecuentes se incluyen en Ck.

11

Las reglas de asociación se generan en dos procesos (Han, Kamber & Pei, 2012:247):

Paso No. 1. Se encuentran todos los conjuntos de elementos frecuentes, cada uno de estos conjuntos de elementos ocurrirá al menos con la misma frecuencia que un umbral mínimo de soporte predeterminado, min_sup.

Paso No. 2. Se Generan reglas de asociación fuertes, a partir de conjuntos de elementos frecuentes: por definición, estas reglas deben satisfacer un soporte mínimo (min_sup) y una confianza mínima (min_conf).

12

EJEMPLO

Considere el siguiente conjunto de elementos:

I = {alcohol, algodón, cubreboca, Gel antibacterial, guantes}

Y el siguiente conjunto de transacciones:

No. Transacción	Ítems comprados
1	{algodón, cubreboca, Gel antibacterial}
2	{cubreboca, guantes}
3	{alcohol, cubreboca}
4	{cubreboca, Gel antibacterial, guantes}
5	{alcohol, Gel antibacterial}
6	{alcohol, cubreboca}
7	{alcohol, Gel antibacterial}
8	{alcohol, cubreboca, , Gel antibacterial}
9	{alcohol, cubreboca, Gel antibacterial}

13

Existen dos métodos para representar los datos de la canasta de mercado:

El formato de datos transaccionales. El cual consta de dos campos: el número de transacción y el valor.

No. Transacción	Ítem comprado
1	Algodón
1	cubreboca
1	Gel antibacterial
2	cubreboca
2	guantes
3	alcohol
3	cubreboca
:	:

14

Formato de datos tabular. Cada registro representa una transacción con 1 ó 0 dependiendo si está o no el ítem

Transacción	alcohol	algodón	cubreboca	Gel antibacterial	guantes
1	0	1	1	1	0
2	0	0	1	0	1
3	1	0	1	0	0
4	0	0	1	1	1
5	1	0	0	1	0
6	1	0	1	0	0
7	1	0	0	1	0
8	1	1	1	1	0
9	1	0	1	1	0
Suma	6	2	7	6	2

15

PASO NO. 1. GENERAR LOS CONJUNTOS DE ELEMENTOS FRECUENTES

PRIMERA ITERACIÓN (L1)

Se refiere (1-itemsets) a los conjuntos frecuentes de 1 ítem, esto es, los productos individuales.

C1 =

Itemset	Soporte
Gel antibacterial	6
Cubreboca	7
alcohol	6
Guantes	2
algodón	2

Se analiza si las ocurrencias cumplen con min_sup = 2. Todos cumplen. Por lo tanto

L1 =

Itemset	Soporte
Gel antibacterial	6
Cubreboca	7
alcohol	6
Guantes	2
algodón	2

16

SEGUNDA ITERACIÓN (L2)

Se generan los itemset candidatos para el conjunto C2, lo cual se realiza a partir de L1.

C2 =

Itemset	Soporte
{cubreboca, Gel antibacterial}	4
{alcohol, Gel antibacterial}	4
{Gel antibacterial, guantes}	1
{algodón, Gel antibacterial}	2
{alcohol, cubreboca}	4
{cubreboca, guantes}	2
{algodón, cubreboca}	2
{alcohol, guantes}	0
{alcohol, algodón}	1
{algodón, guantes}	0

Se analiza si las ocurrencias cumplen con min_sup = 2.

L2 =

Itemset	Soporte
{cubreboca, Gel antibacterial}	4
{alcohol, Gel antibacterial}	4
{algodón, Gel antibacterial}	2
{alcohol, cubreboca}	4
{cubreboca, guantes}	2
{algodón, cubreboca}	2

17

TERCERA ITERACIÓN (L3)

Paso número 1. Join

C3 = L2 ⋈ L2

Itemset	Soporte
{alcohol, cubreboca}	4
{alcohol, Gel antibacterial}	4
{algodón, cubreboca}	2
{algodón, Gel antibacterial}	2
{cubreboca, Gel antibacterial}	4
{cubreboca, guantes}	2

⋈

Itemset	Soporte
{alcohol, cubreboca}	4
{alcohol, Gel antibacterial}	4
{algodón, cubreboca}	2
{algodón, Gel antibacterial}	2
{cubreboca, Gel antibacterial}	4
{cubreboca, guantes}	2

Itemset	Soporte
{alcohol, cubreboca, Gel antibacterial}	4
{algodón, cubreboca, Gel antibacterial}	2
{cubreboca, Gel antibacterial, guantes}	4

18

Paso número 2. Poda.

Subconjunto	Frecuencia	Decisión
{Gel antibacterial, cubreboca, alcohol}		No se poda
{Gel antibacterial, cubreboca}	4	
{Gel antibacterial, alcohol}	4	
{cubreboca, alcohol}	4	
{Gel antibacterial, cubreboca, algodón}		No se poda
{Gel antibacterial, cubreboca}	4	
{Gel antibacterial, algodón}	2	
{cubreboca, algodón}	2	
{cubreboca, gel antibacterial, guantes}		Se poda
{cubreboca, gel antibacterial}	4	
{cubreboca, guantes}	2	
{gel antibacterial, guantes}	1	

19

Se analiza si las ocurrencias cumplen con min_sup = 2.

L3 =

Itemset	Soporte
{alcohol, cubreboca, Gel antibacterial}	2
{algodón, cubreboca, Gel antibacterial}	3

⋈

20

PASO NO. 2. GENERAR LAS REGLAS DE ASOCIACIÓN FUERTES

Las reglas de asociación fuertes deben satisfacer dos condiciones: soporte mínimo y confianza mínima (Han, Kamber & Pei, 2012:254):

Considerando las siguientes fórmulas:

$$\text{Confianza } (A \Rightarrow B) = P(B \mid A) = \frac{support\_count(A \cap B)}{support\_count(A)}$$

Esto significa lo siguiente:

Support_count (A ∩ B) es el número de transacciones que contiene los itemsets A∩B y support_count(A) es el número de transacciones que contienen el itemset A.

21

Las reglas se generan como sigue (Han, Kamber & Pei, 2012):

- Para cada conjunto de elementos frecuentes l, genere todos los subconjuntos no vacíos de l.
- Para cada conjunto no vacío s de l, se aplica la regla: s ⇒ (l – s) si se cumple:

$\frac{support\_count(l)}{support\_count(s)}$ >= min_cof, donde min_cof es el umbral mínimo de confianza.

Recordando que el **soporte** es la proporción de las transacciones en las cuales los elementos {x, y} y {z} ocurren.

Y la confianza significa {x, y} aparecen n1 veces del total de las transacciones, n2 de las cuales también contienen el elemento z, lo cual es una confianza de n2/n1.

22

Los conjuntos de elementos frecuentes finales son:
{alcohol, cubreboca, Gel antibacterial}
{algodón, cubreboca, Gel antibacterial}

Se considera una confianza del 50%
Las reglas generadas son:

Regla	Soporte	Confianza	Decisión
{cubreboca, Gel antibacterial} ⇒ algodón	2/9 = 0.22	2/4 = 50%	
{algodón, Gel antibacterial} ⇒ cubreboca	2/9 = 0.22	2/2 = 100%	Se queda
{algodón, cubreboca} ⇒ Gel antibacterial	2/9 = 0.22	2/2 = 100%	Se queda
Gel antibacterial ⇒ {algodón, cubreboca}	2/9 = 0.22	2/6 = 33%	
cubreboca ⇒ {algodón, Gel antibacterial}	2/9 = 0.22	2/7 = 29%	
algodón ⇒ {cubreboca, Gel antibacterial}	2/9 = 0.22	2/2 = 100%	Se queda
cubreboca, Gel antibacterial⇒ alcohol	2/9 = 0.22	2/4 = 50%	
alcohol, Gel antibacterial ⇒ cubreboca	2/9 = 0.22	2/4 = 50%	
alcohol, cubreboca ⇒ Gel antibacterial	2/9 = 0.22	2/4 = 50%	
Gel antibacterial ⇒ alcohol, cubreboca	2/9 = 0.22	2/6 = 0.33	
cubreboca ⇒ alcohol, Gel antibacterial	2/9 = 0.22	2/7 = 0.28	
alcohol ⇒ cubreboca, Gel antibacterial	2/9 = 0.22	2/6 = 0.33	

23

MÉTODOS DE EVALUACIÓN DE PATRONES

Las reglas estrictas no son necesariamente interesantes

Si una regla es interesante o no, puede evaluarse subjetiva u objetivamente. En última instancia, solo el usuario puede juzgar si una regla determinada es interesante, y este juicio, al ser subjetivo, puede diferir de un usuario a otro. Sin embargo, las medidas objetivas de interés, basadas en las estadísticas "detrás" de los datos, pueden utilizarse como un paso hacia el objetivo de eliminar las reglas poco interesantes que de otro modo se presentarían al usuario. (Han, Kamber & Pei, 2012:265).

24

No todas las reglas de asociación son igualmente útiles. Existe una medida que puede cuantificar la utilidad de una regla de asociación llamada carga (lift) que se define de la siguiente manera (Larose & Larose, 2015):

$$\text{Lift} = \frac{\text{Rule confidence}}{\text{Prior proportion of the consequent}}$$ Figura tomada de Larose & Larose (2015).

$$lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}.$$ Figura tomada de Han, Kamber & Pei (2012:265).

25

Para comprender este significado, Larose & Larose (2015) presentan un ejemplo muy ilustrativo, suponga que existen 1000 ventas registradas, 200 de las cuales compraron pañales y de estas 200 ventas, 50 compraron también cerveza.

La proporción a priori de quienes compraron cerveza es 50/1000 = 5% y la confianza es 50/200 = 0.25.

Por lo tanto, la carga (*lift*) de la regla de asociación: "Si compra pañales, entonces compra cerveza" es: Lift = 0.25/0.05 = 5

Esto puede interpretarse como "Los clientes que compran pañales tienen cinco veces más probabilidades de comprar cerveza que los clientes de todo el conjunto de datos".

26

¿Son las reglas de asociación un tipo de aprendizaje supervisado o no supervisado?

La mayoría de los métodos de minería de datos son del tipo aprendizaje supervisado, debido a que

1. Una variable objetivo está pre especificada, y
2. El algoritmo recibe una cantidad suficientes de ejemplos donde se puede descubrir una posible asociación entre la variable objetivo y las variables predictoras.

Por otro lado, en el aprendizaje no supervisado, no se identifica explícitamente ninguna variable objetivo. Más bien, el algoritmo de minería de datos busca patrones y estructura entre todas las variables. El agrupamiento (*cluster*) es quizás el método de minería de datos sin supervisión más común.

Sin embargo, la minería de reglas de asociación se puede aplicar de manera supervisada o no supervisada.

27

Por ejemplo, en el análisis de la canasta de la compra, simplemente se puede estar interesado en "qué artículos se compran juntos", en cuyo caso no se identificaría una variable objetivo (Larose & Larose, 2015).

Sin embargo, algunos conjuntos de datos están estructurados de manera natural para que una variable particular cumpla el papel de un consecuente, y no un antecedente (por ejemplo el juego de golf) (Larose & Larose, 2015).

Por ejemplo, suponga que los encuestadores políticos han recopilado datos demográficos en sus encuestas de salida, junto con la preferencia de voto del sujeto. En este caso, las reglas de asociación podrían extraerse de este conjunto de datos, donde la información demográfica podría representar posibles antecedentes, y la preferencia de voto podría representar el único consecuente de interés (Larose & Larose, 2015).

28

De esta manera, las reglas de asociación podrían usarse para ayudar a clasificar las preferencias de voto de los ciudadanos con ciertas características demográficas, en un proceso de aprendizaje supervisado (Larose & Larose, 2015).

Por lo tanto, la respuesta a la pregunta es que las reglas de asociación, aunque generalmente se usan para el aprendizaje no supervisado, también se pueden aplicar para el aprendizaje supervisado para una tarea de clasificación (Larose & Larose, 2015).

29

Patrones globales versus patrones locales

Los analistas de datos deben considerar la diferencia entre modelos y patrones (Larose & Larose, 2015).

Un modelo es una descripción global o explicación de un conjunto de datos, teniendo una perspectiva de alto nivel. Los modelos pueden ser descriptivos o inferenciales.
Los modelos descriptivos buscan resumir todo el conjunto de datos de manera sucinta.
Los modelos inferenciales tienen como objetivo proporcionar un mecanismo que permita al analista generalizar de las muestras a las poblaciones. De cualquier manera, la perspectiva es global y abarca todo el conjunto de datos.
Los patrones son esencialmente características locales de los datos. De hecho, los patrones reconocibles pueden ser válidos sólo para unas pocas variables o una fracción de los registros en los datos.

30

Las reglas de asociación son particularmente adecuadas para descubrir patrones locales en los datos, debido a que al aplicar la cláusula *if* en una regla de asociación, se está dividiendo los datos para que, por lo general, a la mayoría de los registros no se apliquen (Larose & Larose, 2015).

La aplicación de la cláusula *if* "profundiza" más intensamente en un conjunto de datos, con el objetivo de descubrir un patrón local oculto que puede o no ser relevante para la mayor parte de los datos (Larose & Larose, 2015).

31

Referencias bibliográficas

Expósito, Expósito, López, Melián y Moreno. ("s/f"). Minería de patrones de asociación. Material educativo del Departamento de Ingeniería Informática y de Sistemas.
Han, Jiawei; Kamber, Micheline & Pei, Jian. (2012). Data Mining: concepts and techniques. Third edition. Morgan Kaufman Series.
Larose, T. Daniel & Larose, D. Chantal. (2015). *Data Mining and Predictive Analytics*. Second Edition. Wiley.