UNIVERSIDAD CATÓLICA BOLIVIANA “SAN PABLO”
FACULTAD DE INGENIERÍA
CARRERA INGENIERÍA DE SISTEMAS
MAESTRÍA EN CIENCIA DE DATOS

Reglas de Asociación

Introducción

- Las “Reglas de Asociación” corresponden a un concepto importante del aprendizaje automático que generalmente se utiliza en el análisis de “Cesta de la compra”.

- En una tienda la colocación de los productos puede tener un grado de impacto mayor en la medida de la aplicación de técnicas que ayuden a identificar la interrelación entre productos. Ej. la colocación de los productos como verduras, lácteos, carnes, etc., en los anaqueles de las tiendas.

- Invertir tiempo y recursos en el estudio de las ubicaciones de productos no solo reduce el tiempo de compra de un cliente, sino que también le recuerda qué artículos relevantes podría estar interesado en comprar.

- Las reglas de asociación ayudan a descubrir todas esas relaciones entre elementos en grandes bases de datos.

Ejemplo de transacciones

TID	Productos
1	LECHE, PAN, HUEVOS
2	PAN, AZUCAR
3	PAN, CEREAL
4	LECHE, PAN, AZUCAR
5	LECHE, CEREAL
6	PAN, CEREAL
7	LECHE, CEREAL
8	LECHE, PAN, CEREAL, HUEVOS
9	LECHE, PAN, CEREAL

Base de datos de transacciones: Ejemplo

ITEMS:
A =leche
B= pan
C= cereal
D= azúcar
E= huevos

TID	Productos
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

TID	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

Definiciones

- Item: par *atributo*=*valor* o simplemente *valor*
  - normalmente los atributos se convierten a binarios para cada valor, ej. **producto="A"** se escribe como atributo **"A"**
- Itemset *I*: subconjunto de items posibles
  - ej: *I* = {A,B,E} (el orden no importa)
- Transacción: (TID, conjunto de items)
  - TID es el ID de la transacción

Propiedad de los subconjuntos

- Idea fundamental en el análisis de asociaciones: “Todo subconjunto de un conjunto de ítems frecuente es frecuente.”

- Ej: supongamos que {A,B} es frecuente. Como cada ocurrencia de A,B incluye a A y a B, entonces tanto A como B también deben ser frecuentes

- Argumento similar para conjuntos de mayor tamaño

- Casi todos los algoritmos de reglas de asociación se basan en esta propiedad de los subconjuntos

Encontrar reglas de asociación fuertes

- El primer paso en la generación de reglas de asociación es obtener todos los conjuntos de elementos frecuentes en los que se pueden realizar particiones binarias para obtener el antecedente y el consecuente.

- Por ejemplo, si hay 6 elementos {Pan, mantequilla, huevo, leche, cuaderno, cepillo de dientes} en todas las transacciones combinadas, los conjuntos de elementos se verán como {Pan}, {Mantequilla}, {Pan, Cuaderno}, {Leche, cepillo de dientes} , {Leche, Huevo} etc.

- El tamaño de un conjunto de elementos puede variar de uno a la cantidad total de elementos que tenemos.

- Buscamos solo conjuntos de elementos frecuentes de estos.

Encontrar reglas de asociación fuertes

- Los conjuntos de elementos frecuentes son los que ocurren al menos un número mínimo de veces en las transacciones que están por encima de un umbral mínimo: minsupport

- Se tiene el algoritmo Apriori que ayuda a que esta búsqueda sea eficiente.

Algoritmo Apriori

- Generar todos los itemsets (support >= minsup) que tengan solo un item.

- Generar item sets que tengan dos items, como también todas las combinaciones de itemsets identificados. “Podar” los itemsets para los cuales el valor de support sea menor que minsup.

- Generar item sets que tengan tres items como también todas las combinaciones de itemsets de tamaño dos (que quedaron luego de haber “podado” en punto anterior), y llevar a cabo el mismo chequeo respecto al valor de support.

Algoritmo Apriori

- Se continua incrementando de a uno el tamaño de los itemsets y se verifica el minsup threshold en cada paso.

- Luego de generados los itemsets frecuentes, se identifican las reglas con estos.

Generación de Reglas de Asociación

- Las reglas son formadas por medio de la participación binaria de cada itemset, por ejemplo, Si {Pan, Huevo, Leche, Mantequilla} es un itemset, las reglas candidatas son:

(Huevo, Leche, Mantequilla -> Pan)

(Pan, Leche, Mantequilla -> Huevo)

(Pan, Huevo -> Leche, Mantequilla)

(Huevo, Leche -> Pan, Mantequilla)

(Mantequilla -> Pan, Huevo, Leche)

Support y Confidence

- Support: Es la proporción de transacciones t en el dataset que contienen el itemset A and C.

$$\text{support}(A \rightarrow C) = \text{support}(A \cup C), \quad \text{range: } [0, 1]$$

- Confidence: Es la proporción de transacciones que contienen el consecuente C que también contienen el antecedente A.

$$\text{confidence}(A \rightarrow C) = \frac{\text{support}(A \rightarrow C)}{\text{support}(A)}, \quad \text{range: } [0, 1]$$

- Cuando el antecedente y el consecuente ocurren juntos el valor de confidence es uno.

Lift

- Lift es comunmente utilizada para medir cuan mas a menudo el antecedente y el consecuente de una regla A->C ocurren juntos de lo que se pudiera esperar si fueran estadísticamente independientes.

$$\text{lift}(A \rightarrow C) = \frac{\text{confidence}(A \rightarrow C)}{\text{support}(C)}, \quad \text{range: } [0, \infty]$$

- Si A y C fueran independientes lift sería exáctamente uno.
- En casos donde A conduce a C, el valor de lift es mayor a uno.
- En casos donde lift sea menor a uno A y C negativamente correlacionados.

Conviction

- Es el ratio de que la regla realice una prediccion incorreta.

$$\text{conviction}(A \rightarrow C) = \frac{1 - \text{support}(C)}{1 - \text{confidence}(A \rightarrow C)}, \quad \text{range: } [0, \infty]$$

- Un valor alto de esta medida significa que el consecuente es altamente dependiente del antecedente.
- En el caso de un Confidence perfecto, el denominador se convierte en cero para lo cual esta medida se define como Infinito.
- Similar a Lift, si los items son independientes, Conviction es 1.

Leverage

- Leverage mide la diferencia entre la frecuencia observada de A y C que aparecen juntas y la frecuencia que se esperaría si A y C fueran independientes.

$$\text{leverage}(A \rightarrow C) = \text{supp}(A \rightarrow C) - \text{supp}(A) \times \text{supp}(C) \qquad \text{range: } [-1, 1]$$

- Un leverage de cero indica independencia.

Reglas de Asociación

- Regla de asociación *R* :

  Itemset1 => Itemset2

  - *Itemset 1, 2* son disjuntos; *Itemset2* es no vacío
  - Significado: si la transacción incluye *Itemset1* entonces también tiene *Itemset2*
- Ejemplos
  - A,B => E,C
  - A => B,C
  - => B, C

De los conjuntos de ítems frecuentes a las reglas de asociación

- *Dado un conjunto de items frecuente {A,B,E}, ¿cuáles son las reglas de asociación posibles?*
  - A => B, E
  - A, B => E
  - A, E => B
  - B => A, E
  - B, E => A
  - E => A, B
  - __ => A,B,E (regla vacío), o Verdadero => A,B,E

Reglas de Asociación-Ejemplo

Dado el conjunto frecuente {A,B,E}, ¿qué reglas de asociación tienen minsupport = 2 y minconf= 50% ?

A, B => E: confidence=2/4 = 50%

TID	List of items
1	**A, B, E**
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	**A, B, C, E**
9	A, B, C

Reglas de Asociación-Ejemplo

Dado el conjunto frecuente {A,B,E}, ¿qué reglas de asociación tienen minsup = 2 y minconf= 50% ?

A, B => E : conf=2/4 = 50%

A, E => B : conf=2/2 = 100%

TID	List of items
1	**A, B, E**
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	**A, B, C, E**
9	A, B, C

Reglas de Asociación-Ejemplo

Q: Dado el conjunto frecuente {A,B,E}, ¿qué reglas de asociación tienen minsupport = 2 y minconf= 50% ?

A, B => E : conf=2/4 = 50%

A, E => B : conf=2/2 = 100%

B, E => A : conf=2/2 = 100%

TID	List of items
1	**A, B, E**
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	**A, B, C, E**
9	A, B, C

Reglas de Asociación-Ejemplo

Q: Dado el conjunto frecuente {A,B,E}, ¿qué reglas de asociación tienen minsupport = 2 y minconf= 50% ?

A, B => E : conf=2/4 = 50%

A, E => B : conf=2/2 = 100%

B, E => A : conf=2/2 = 100%

E => A, B : conf=2/2 = 100%

No lo cumplen

A =>B, E : conf=2/6 =33%< 50%

B => A, E : conf=2/7 = 28% < 50%

_ => A,B,E : conf: 2/9 = 22% < 50%

TID	List of items
1	**A, B, E**
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	**A, B, C, E**
9	A, B, C

Evaluación de Support y Confidence

- De la lista de todas las posibles reglas candidatas, se identifican las reglas que tienen un nivel de confidence por encima del minconf.

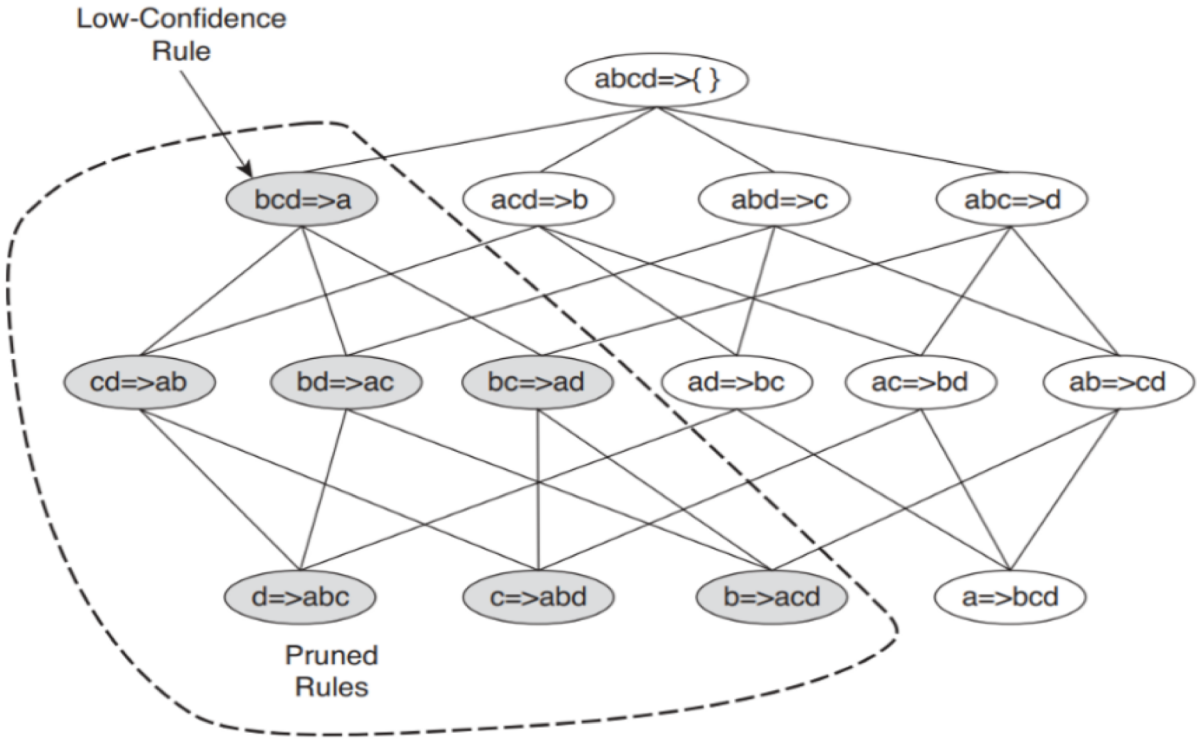
- Esto significa que la confidence de

  (A,B,C -> D) >= (B,C -> A,D) >= (C -> A,B,D)

- Como conocemos que el support de todas las reglas generadas del mismo itemset se mantienen el mismo, la diferencia ocurre solamente en el cálculo de confidence del denominador. Es así que a medida que los items en A disminuyan, el support{A} se incrementa, y el valor de confidence disminuye.

Evaluación de Support y Confidence

- Se inicia con un itemset frecuente {a, b, c, d}, y se va formando reglas con un consecuente solamente, luego se quita las reglas que no satisfacen la condición de minconf.



Ref: https://www-users.cs.umn.edu/~kumar001/dmbook/ch6.pdf

Evaluación de Support y Confidence

- Luego se continua formando reglas combinando los consecuentes de las que quedan. Se repite este proceso hasta que quede solamente un item en el antecedente. Este proceso se hace para todos los itemsets frecuentes.

- Con estos pasos identificamos un conjunto de Reglas de Asociación las cuales satisfacen ambas condiciones, un support y confidence mínimos.

- Luego, en este conjunto de reglas generado se buscan los valores más altos de lift para tomar las decisiones de negocio correspondientes.

Ejemplo con datos del tiempo: ¿Jugar?

Aspecto	Temperatura	Humedad	Viento	Jugar
soleado	calor	alta	falso	no
soleado	calor	alta	verdadero	no
nublado	calor	alta	falso	sí
lluvioso	templado	alta	falso	sí
lluvioso	fresco	normal	falso	sí
lluvioso	fresco	normal	verdadero	no
nublado	fresco	normal	verdadero	sí
soleado	templado	alta	falso	no
soleado	fresco	normal	falso	sí
lluvioso	templado	normal	falso	sí
soleado	templado	normal	verdadero	sí
nublado	templado	alta	verdadero	sí
nublado	calor	normal	falso	sí
lluvioso	templado	alta	verdadero	no

Ejemplo: generar reglas de un conjunto de ítems

- Conjunto frecuente de ítems de datos del tiempo con support 4

Humedad = Normal, Viento = Falso, Jugar = Sí

- Siete reglas potenciales

Si Humedad=Normal y Viento=Falso entonces Jugar=Sí	4/4
Si Humedad=Normal y Jugar=Sí entonces Viento=Falso	4/6
Si Viento=Falso y Jugar=Sí entonces Humedad=Normal	4/6
Si Humedad=Normal entonces Viento=Falso y Jugar=Sí	4/7
Si Viento=Falso entonces Humedad=Normal y Jugar=Sí	4/8
Si Jugar=Sí entonces Humedad=Normal y Viento=Falso	4/9
Si True entonces Humedad=Normal y Viento=Falso y Jugar=Sí	4/12

Reglas para los datos del tiempo

- Reglas con support > 1 y confianza=100%:

	Regla de asociación		Sup.	Conf.
1	Humedad=Normal Viento=Falso	⇒Jugar=Sí	4	100%
2	Temperatura=Fresco	⇒Humedad=Normal	4	100%
3	Aspecto=Nublado	⇒Jugar=Sí	4	100%
4	Temperatura=Frío Jugar=Sí	⇒Humedad=Normal	3	100%
…	…	…	…	…
58	Aspecto=Soleado Temperatura=Caliente	⇒Humedad=Alta	2	100%

- En total: 3 reglas con cobertura 4, 5 con cobertura 3, y 50 con cobertura 2

Ejemplo

Transaction id	Items
t1	{1, 2, 4, 5}
t2	{2, 3, 5}
t3	{1, 2, 4, 5}
t4	{1, 2, 3, 5}
t5	{1, 2, 3, 4, 5}
t6	{2, 3, 4}

Ref.: https://www.philippe-fournier-viger.com/spmf/AssociationRulesWithLift.php

Filtrado de reglas de asociación

- Problema: cualquier conjunto grande de datos puede producir gran cantidad de reglas de asociación, incluso con valores de min support y confidence razonables
- La confidence sola no es suficiente

Dificultades en las aplicaciones

- Encontrar reglas no significa saber cómo usarlas
- Wal-Mart sabe que todos los clientes que compran muñecas Barbie tienen una probabilidad del 60% de comprar alguno de tres tipos de caramelos
- ¿Qué hacer con esa información?
  - “No tengo ni idea” (gerente de ventas de Wal-Mart, Lee Scott)
  - www.kdnuggets.com/news/98/n01.html

Bibliografía

- Han J., Kamber M. Pei J.: “Data Mining Concepts and Techniques”. Ed. 3ra. Morgan Kaufmann

- Witten I., Frank E., Hail A.: “Data Mining”. Ed. 3ra., 2011. Morgan Kaufmann.

- Complete Guide to Association Rules 1 and 2.

  https://towardsdatascience.com/complete-guide-to-association-rules-2-2-c92072b56c84

- Alicia Pérez PhD - Diapositivas