Unidad 2: Minería de elementos frecuentes

Modelo mercado canasta

Blanca Vázquez y Gibran Fuentes-Pineda

29 de agosto de 2022

Los principales bloques en minería de datos

- Clasificación
- Clustering
- Detección de valores atípicos
- Minería de patrones

¿Por qué son especiales estos bloques?

Naturaleza de los datos

Consideremos una matriz *D* de tamaño *nxd*

	d_1	d_2	d_3	...	d_n
n_1	0	0	1	0	1
n_2	0	1	0	0	0
n_3	0	0	0	1	1
.....	...	...	...	...	...
n_n	0	0	1	0	0

Relaciones entre los elementos de una matriz

- Relaciones entre columnas
  - Clasificación
- Relaciones entre filas
  - Clustering
  - Detección de valores atípicos
  - Minería de patrones

Minería de patrones

De manera general, la minería de patrones se define en el contexto de las matrices binarias dispersas.

	Pan	Mantequilla	Leche	Huevos	Yogurt
Cliente_1	1	1	1	0	0
Cliente_2	0	0	1	1	1
Cliente_3	0	1	0	1	1
Cliente_4	1	0	1	0	0

Uno de los estudios más comunes es la minería de elementos frecuentes.

Minería de elementos frecuentes

- La minería de elementos frecuentes es el proceso de **descubrimiento de tendencias o patrones** a partir de grandes conjuntos de datos con el objetivo de guiar futuras decisiones.
- Surgió en el contexto de los datos de los supermercados

¡Vámonos de compras!

Figura 1: Imagen tomada de Puig [1]

¿Qué productos compran los clientes?, ¿Cuáles productos compran juntos?

¡Vámonos de compras!

Figura 2: Imagen tomada de Debashis Borgohain

¡Vámonos de compras!

La minería de elementos frecuentes ha sido ampliamente usada en los supermercados*.

Modelo mercado - canasta

Describe una relación de muchos a muchos entre dos clases:

- **Elementos** (*items*): son cada uno de los eventos o elementos en una transacción
- **Transacciones** (*baskets*): es una colección de items (*itemset*)

Objetivo: identificar el conjunto de elementos que son adquiridos en conjunto.

$$\{\text{fideos, queso rallado}\} \Rightarrow \{\text{salsa}\}$$

Modelo mercado - canasta

Se asume lo siguiente:

- El número de elementos en una transacción es más pequeño que el número total de elementos.
- El número de transacciones puede ser tan grande, que podemos llegar a saturar el espacio de almacenamiento.

Elementos frecuentes

Son aquellos conjuntos de elementos que pueden aparecer en muchas transacciones, y estos tendrán un umbral llamado soporte.

Definición: sea I un conjunto de elementos, el soporte de I es la fracción de transacciones para lo cual I es un subconjunto.

Ejemplo

Base de datos de transacciones de una tiendita

T	Elementos
1	{pan, mantequilla, leche}
2	{huevos, leche, yogurt}
3	{pan, queso, huevos, leche}
4	{huevos, leche, galletas, pan}
5	{queso, galletas, yogurt}

Cálculo de soporte

- Recordemos, cada transacción es un conjunto de elementos comprados al mismo tiempo.
- El total de transacciones en la bd = 5
- El soporte de un elemento I se denota por $sup(I)$
- El $sup(\{pan\}) = ?$
- El $sup(\{mantequilla\}) = ?$
- El $sup(\{yogurt\}) = ?$

Cálculo de soporte

- Cuando se calcula el soporte de un elemento, se dice que es un **conjunto único**.
- También podemos calcular el soporte para dos o más elementos: $sup(\{galletas, leche\}) = 1/5 = 0.2$
- Decimos que *I* es frecuente, si su soporte es igual o mayor al umbral definido para *sup*

Identificación de elementos frecuentes

Supongamos que el soporte mínimo $minsup = 0.5$, ¿qué pares de elementos, cumplen este umbral?

T	Elementos
1	{pan, mantequilla, leche}
2	{huevos, leche, yogurt}
3	{pan, queso, huevos, leche}
4	{huevos, leche, galletas, pan}
5	{queso, galletas, yogurt}

Reglas de asociación

Las reglas de asociación son los elementos más importantes en el modelo mercado canasta.

$$\{X\} \Rightarrow \{Y\}$$

* dónde X y Y son elementos individuales o conjuntos de elementos,
* *X* se conoce como **antecedente** y *Y* como **consecuente**

$$\{\text{fideos, queso rallado}\} \Rightarrow \{\text{salsa}\}$$

Reglas de asociación

Ejemplos de asociaciones comunes:

- {pañales} $\Rightarrow$ {cerveza}
- {leche} $\Rightarrow$ {pan}
- {pan para hot dogs} $\Rightarrow$ {mostaza}

Confianza

Sea X y Y dos conjuntos de elementos, la confianza $conf(X \cup Y)$ de la regla $X \cup Y$ es la probabilidad condicional de $X \cup Y$ que ocurre en una transacción dado que la transacción contiene X.

$$conf(X \Rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)}$$

Confianza

T	Elementos
1	{pan, mantequilla, leche}
2	{huevos, leche, yogurt}
3	{pan, queso, huevos, leche}
4	{huevos, leche, yogurt}
5	{queso, leche, yogurt}

$$conf(\{huevos, leche\} \Rightarrow \{yogurt\}) = \frac{sup(\{huevos, leche, yogurt\})}{sup(\{huevos, leche\})}$$

$$conf(\{huevos, leche\} \Rightarrow \{yogurt\}) = \frac{sup(2/5)}{sup(3/5)}$$

$$conf(\{huevos, leche\} \Rightarrow \{yogurt\}) = 2/3$$

Confianza y soporte

Las reglas de asociación, se definen por su soporte y su nivel de confianza.

- Semejante al soporte, en donde se define un soporte mínimo (*minsup*), también se define un nivel de confianza mínimo (*minconf*)
- El *minconf* puede ser usado para generar reglas de asociación relevantes
- La confianza (*conf*) es la fuerza de la asociación entre elementos

Cálculo de la confianza

T	Elementos
1	{pan, gelatina, mantequilla-maní}
2	{pan, mantequilla-maní}
3	{pan, leche, mantequilla-maní}
4	{cerveza, pan}
5	{cerveza, leche}

Cálculo de la confianza

Calcula la confianza de las siguientes reglas de asociación:

$conf(\{pan\} \Rightarrow \{mantequilla - mani\})$
$conf(\{cerveza\} \Rightarrow \{pan\})$
$conf(\{mantequilla - mani\} \Rightarrow \{gelatina\})$
$conf(\{gelatina\} \Rightarrow \{leche\})$

Cálculo de la confianza

Calcula la confianza de las siguientes reglas de asociación:

$conf(\{pan\} \Rightarrow \{mantequilla - mani\}) = (3/5)/(4/5) = 0.75$
$conf(\{cerveza\} \Rightarrow \{pan\}) = (1/5)/(2/5) = 0.5$
$conf(\{mantequilla - mani\} \Rightarrow \{gelatina\}) = (1/5)/(3/5) = 0.3$
$conf(\{gelatina\} \Rightarrow \{leche\}) = (0)/(2/5) = 0$

Importante en las reglas de asociación

- Reglas con bajo soporte: pueden haber aparecido por casualidad
- Reglas con baja confianza: es probable que no existe relación entre el antecedente y el consecuente
- $\{pepsi\} \Rightarrow \{coca - cola\}$

Tipos de reglas de asociación

- Binarias vs cuantitativas
- Unidimensionales vs multidimensionales
- De un nivel vs multinivel

Reglas binarias y cuantitativas

Se basan en los tipos de datos:

- $compra\{laptop\} \Rightarrow compra\{impresora\}$
- $edad\{> 30\} \wedge sueldo\{> 30,000\} \Rightarrow compra\{SmartTV\}$

Reglas unidimensionales y multidimensionales

Se basan en las dimensiones de los datos involucrados en la regla

- $compra\{laptop\} \Rightarrow compra\{impresora\}$
- $edad\{> 30\} \wedge sueldo\{> 30,000\} \Rightarrow compra\{SmartTV\}$
- $compra\{traje, camisa\} \Rightarrow compra\{zapatos, corbata\}$

Reglas de un nivel vs multinivel

Se basan en el nivel de abstracción involucrado:

Reglas de un nivel vs multinivel

Se basan en el nivel de abstracción involucrado:

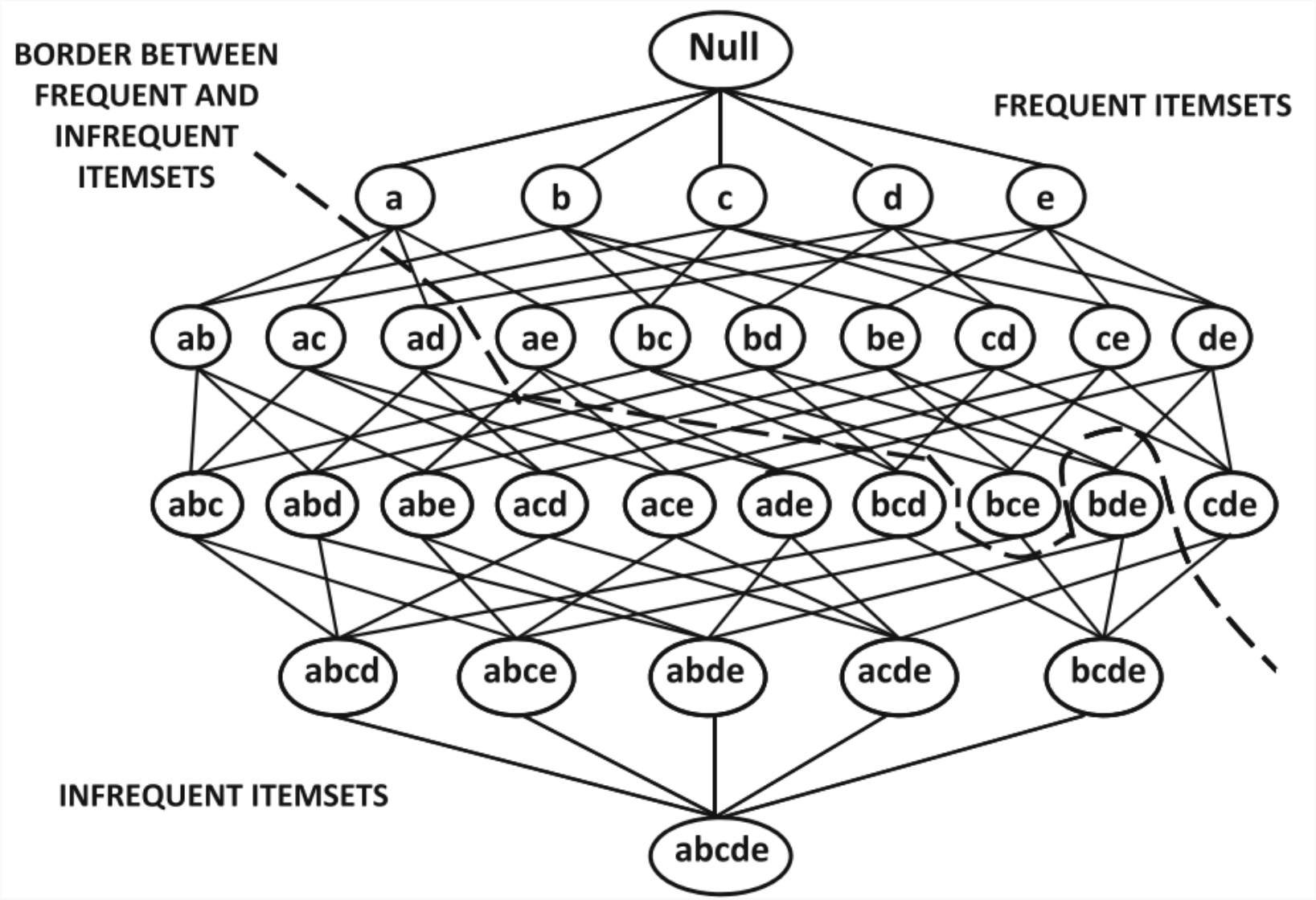

Imagen tomada de Aggarwal, 2015.

Ejemplos de reglas de asociación

- Casas inteligentes:
  $\{temp\_baja\} \Rightarrow \{encender\_calefaccion\}$
  $\{despertador\_activo\} \Rightarrow \{encender\_cafetera\}$
- Sistemas de recomendación:
  $\{El\_senor\_anillos\} \Rightarrow \{El\_hobbit\}$
- Dispositivos móviles:
  $\{mensaje\_noimportate, reunion\} \Rightarrow \{no\_notificar\}$

Aplicaciones de la minería de elementos frecuentes

- Identificación de patrones de compra (asociaciones)
- Minería de texto
- Detección de eventos espacio-temporales
- Detección de errores de programación
- Biomarcardores
- Identificación de plagio
- Identificación de patrones en bibliotecas, librerías.

Referencias

1. Introduction to machine learning, Albert Orriols-Puig. URL: *https://www.slideshare.net/aorriols/lecture13-association-rules*
2. Inteligencia Artificial, Ariel Monteserin.