Modelos de Asociación

Prof. Dr. Jorge Zavaleta
zavaleta.jorge@gmail.com

Modelos de Asociación (MAs)

- **Introducción**:
  - Las empresas acumulan enormes cantidades de datos de sus operaciones diarias.
  - El comercio minorista se interesa en analizar los datos para aprender el comportamiento de compras de los clientes.
  - La minería de datos busca extraer conocimiento de grandes conjuntos de datos.
  - Una de las tareas más importantes de este proceso es el descubrimiento de **reglas de asociación**, que identifican relaciones frecuentes entre elementos de un conjunto de datos.

MAs - Definición

- Los modelos de asociación identifican **conjuntos de elementos** (**itemsets**) que frecuentemente coexisten en las transacciones.
- Las relaciones descubiertas se pueden representar en forma de **itemsets** presentes en muchas transacciones, que se conocen como **itemsets frecuentes** o **reglas de asociación**, que representan relaciones entre dos itemsets.
- Las **reglas de asociación** se representan en la forma **X -> Y**, donde **X** e **Y** son **itemsets** disjuntos (**X∩Y = 0**).

Transaciones de cesta de mercado

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

@Tan,2019

{pañales} → {cerveza}

MAs – Representación binaria

- Los datos de la cesta de la compra se pueden representar en formato binario, donde cada fila corresponde a una transacción y cada columna corresponde a un artículo.
- Un artículo puede tratarse como una variable binaria cuyo valor es **uno** si el artículo está presente en una transacción y **cero** en caso contrario.

Representación de la cesta básica binaria

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

@Tan,2019

MAs – Itemset

- Sea $I = \{i_1, i_2, \ldots, i_d\}$ el conjunto de todos los elementos en una **cesta de compras** y sea
- T= $\{t_1, t_2, \ldots, t_N\}$ el conjunto de todas las **transacciones**.
- Cada transacción t_i contiene un subconjunto de elementos elegidos de **I**
- Una colección de cero o más elementos se denomina **itemset** (conjunto de elementos).

- Si un itemset contiene **k** elementos, se denomina conjunto de **k** elementos.
- Ejemplo: *{cerveza, pañales, leche}* es un ejemplo de un 3-itemset.
- Se dice que una transacción t_j contiene un itemset X si X es un subconjunto de t_j

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0

MAs – Conteo de soporte

- Una propiedad importante de un itemset es su **conteo de soporte**, que *se refiere a la cantidad de transacciones que contienen un itemset en particular*.
- El **conteo de soporte** $\boldsymbol{\sigma(X)}$, para un itemset X, se puede expresar de la siguiente manera:

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}|,$$

Representación de la cesta básica binaria

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

@Tan,2019

- En la tabla, el **conteo de soporte** para **{Cerveza, Pañales, Leche}** es igual a **dos** porque sólo hay dos **transacciones** que contienen los 3 elementos: $\boldsymbol{\sigma(X)}$= 2

MAs – fracción de transacciones

- La **fracción de transacciones** es una propiedad de **soporte** en las que ocurre un itemset:

$$\mathbf{s}(\boldsymbol{X}) = \boldsymbol{\sigma}(\boldsymbol{X})/\boldsymbol{N}$$

- Cantidad de veces em que el conjunto X aparece.
- Un itemset **X** se llama **frecuente** si **s(X)** es mayor que algún umbral definido por el usuario, ***minsup***.

MAs – regla de asociación

- Una **regla de asociación** es una expresión de implicación de la forma **X** → **Y** , donde X e Y son itemsets disjuntos, es decir, **X** ∩ **Y = ∅**.
- La fuerza de una regla de asociación se puede medir en términos de su **soporte** y **confianza**.
- El **soporte** determina la **frecuencia** con la que se aplica una regla a un itemset determinado.

- La **confianza** determina la **frecuencia** con la que los elementos de **Y** aparecen en transacciones que contienen **X.**
- $\boldsymbol{\sigma(X \cup Y)}$, cantidad de veces em que los conjuntos X y Y aparecen en unión:
- **Soporte**: $\boldsymbol{s(X \cup Y) = \frac{\sigma(X \cup Y)}{N}}$
- **Soporte**: $\boldsymbol{s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}}$
- **Confianza**: $\boldsymbol{c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}}$
- **Elevación: lift**$\boldsymbol{(X \rightarrow Y) = \frac{s(X \cup Y)}{s(X) * s(Y)}}$

MAs – Ejemplo

- Sea o conjunto de transacciones.

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

- Considere la regla:

  {Milk, Diapers, Beer} →{Beer}

- **X={Milk, Diapers, Beer}**
- $\boldsymbol{\sigma(X)} = 2$, N = 5 (TIDs)
- $\mathbf{s}(\boldsymbol{X}) = \frac{\sigma(X)}{N} = \frac{2}{5} = 0{,}4$
- $\mathbf{c}(\boldsymbol{X} \rightarrow \boldsymbol{Y}) = \frac{\boldsymbol{\sigma(X)}}{\boldsymbol{\sigma(\{milk,Diaper\})}} = \frac{\mathbf{2}}{\mathbf{3}} = \mathbf{0,67}$

MAs – Ejemplo … cont.

- **Generación de conjuntos de elementos frecuente por fuerza bruta**
  - Enumerar todos los conjuntos de elementos posible y calcular su soporte
  - La primera tabla representa una base de datos de compras efectuadas.

TID	Itens
1	{Pão, Leite}
2	{Pão, Fraldas, Cerveja, Ovos}
3	{Leite, Fraldas, Cerveja, Cola}
4	{Pão, Leite, Fraldas, Cerveja}
5	{Pão, Leite, Fraldas, Cola}

Candidatos	Suporte
{Pão}	4/5 = 0.9
{Leite}	4/5 = 0.9
{Fraldas}	4/5 = 0.9
{Cerveja}	3/5 = 0.8
{Ovos}	1/5 = 0.2
{Cola}	2/5 = 0.4
{Pão,Leite}	3/5 = 0.8
{Pão,Fraldas}	3/5 = 0.9
{Pão, Cerveja}	2/5 = 0.4
…	…

@UFOP

MAs – Tipos de reglas de asociación

- **Reglas generales de asociación**: identificar relaciones entre conjuntos de elementos sin considerar restricciones.
- **Reglas de asociación con soporte mínimo**: Identificar relaciones entre conjuntos de elementos con un nivel mínimo de soporte.
- **Reglas de asociación con confianza mínima**: Identificar relaciones entre conjuntos de elementos con un nivel mínimo de confianza.

MAs – Medidas de evaluación

- **Soporte**: Frecuencia con la que se cumple la regla en los datos.
  - Medida que indica la proporción de X en Y.
- **Confianza**: Probabilidad de que Y ocurra dado X.
  - Medida que expresa la proporción de "Si X fuera comprado, cual es la posibilidad de Y ser comprado"
  - Si valor es alto, aparece en todas las transacciones.
- **Elevación (lift)**: Medida de la fuerza de la asociación entre X e Y.
  - Si **lift(x → y)>1**, entonces Y es probable de ser comprado, cuando X fuera comprado
  - Si **Si lift(x → y)≤1**, entonces NO es probable que Y sea comprado, caso X sea comprado.

MAs - Algoritmos

- Los algoritmos de asociación más utilizados son:
  - **A priori**: busca conjuntos frecuentes de elementos mediante un proceso iterativo.
  - **FP-Growth**: crea un árbol de prefijos para identificar conjuntos de elementos frecuentes.
  - **Eclat**: utiliza una estructura de datos hash para identificar conjuntos frecuentes de elementos.

MAs – Técnicas de mejora

- **Reducción de dimensionalidad**: Selección de atributos relevantes para el descubrimiento de reglas.
- **Generación de reglas de asociación negativa**: Identifica relaciones entre conjuntos de elementos que no coexisten con frecuencia.
- **Descubrimiento de reglas de asociación multidimensional**: identifica relaciones entre conjuntos de elementos en múltiples dimensiones.

MAs - Aplicaciones

- Los modelos de asociación se aplican en varias áreas, tales como:
  - **Marketing**: Análisis de la cesta de la compra para identificar productos que se compran juntos con frecuencia.
  - **Finanzas**: Detección de fraude y análisis de riesgos.
  - **Salud**: Diagnóstico de enfermedades y análisis de datos médicos.
  - **Bioinformática**: Análisis de genes y proteínas.

MAs – Herramientas para la minería asociativa

- **Mlxtend**: biblioteca Python
- **Pycaret**: biblioteca Python
- **RapidMiner**: herramienta de minería de datos de código abierto.
- **Weka**: herramienta de código abierto para minería de datos.
- **SAS Enterprise Miner**: Herramienta comercial para minería de datos.

MAs - Ventajas

- Descubrir relaciones frecuentes entre elementos.
- Identificación de patrones de compra.
- Predecir el comportamiento futuro.
- Generación de reglas de negocio.

MAs - Desventajas

- Alta dimensionalidad de los datos.
- Interpretación de las reglas de asociación.
- Sensibilidad al ruido en los datos.

MAs – Tendencias y avances

- El área de los modelos de asociación está en constante evolución, con nuevas investigaciones y avances en:
  - **Aprendizaje automático**: integración de técnicas de aprendizaje automático para mejorar el rendimiento del modelo.
  - **Descubrimiento de reglas en tiempo real**: desarrollo de algoritmos para procesar datos en tiempo real.
  - **Interpretabilidad**: Desarrollo de métodos para explicar los resultados del modelo de una manera más clara y transparente.

MAs - Conclusiones

- Los modelos de asociación son poderosas herramientas para el análisis de datos, con diversas aplicaciones en diferentes áreas.
- A pesar de sus ventajas, es importante ser consciente de sus desventajas y consideraciones éticas.
- A medida que avance la investigación y se desarrollen nuevas técnicas, los modelos de asociación seguirán siendo herramientas importantes para descubrir conocimiento en grandes conjuntos de datos.

Bibliografía

- Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.
- Aggarwal, C. C. (2015)., 29(5), 1023-1078. Cluster analysis: A survey of recent developments. The Journal of Knowledge Discovery and Data Mining
- Tan, P. N., Kumar, V., & Srivastava, J. (2006). Introduction to data mining. Pearson.