Minería de patrones frecuentes y reglas de asociación

Máster Online en Ciencia de Datos

UCO ONLINE

Dr. José Raúl Romero

Profesor Titular de la Universidad de Córdoba y Doctor en Ingeniería Informática por la Universidad de Málaga. Sus líneas actuales de trabajo se centran en la democratización de la ciencia de datos (*Automated ML* y *Explainable Artificial Intelligence*), aprendizaje automático evolutivo y analítica de software (aplicación de aprendizaje y optimización a la mejora del proceso de desarrollo de software).

Miembro del Consejo de Administración de la *European Association for Data Science*, e investigador senior del Instituto de Investigación Andaluz de *Data Science and Computational Intelligence*.

Director del **Máster Online en Ciencia de Datos** de la Universidad de Córdoba.

Introducción a Frequent Pattern Mining

Taxonomía

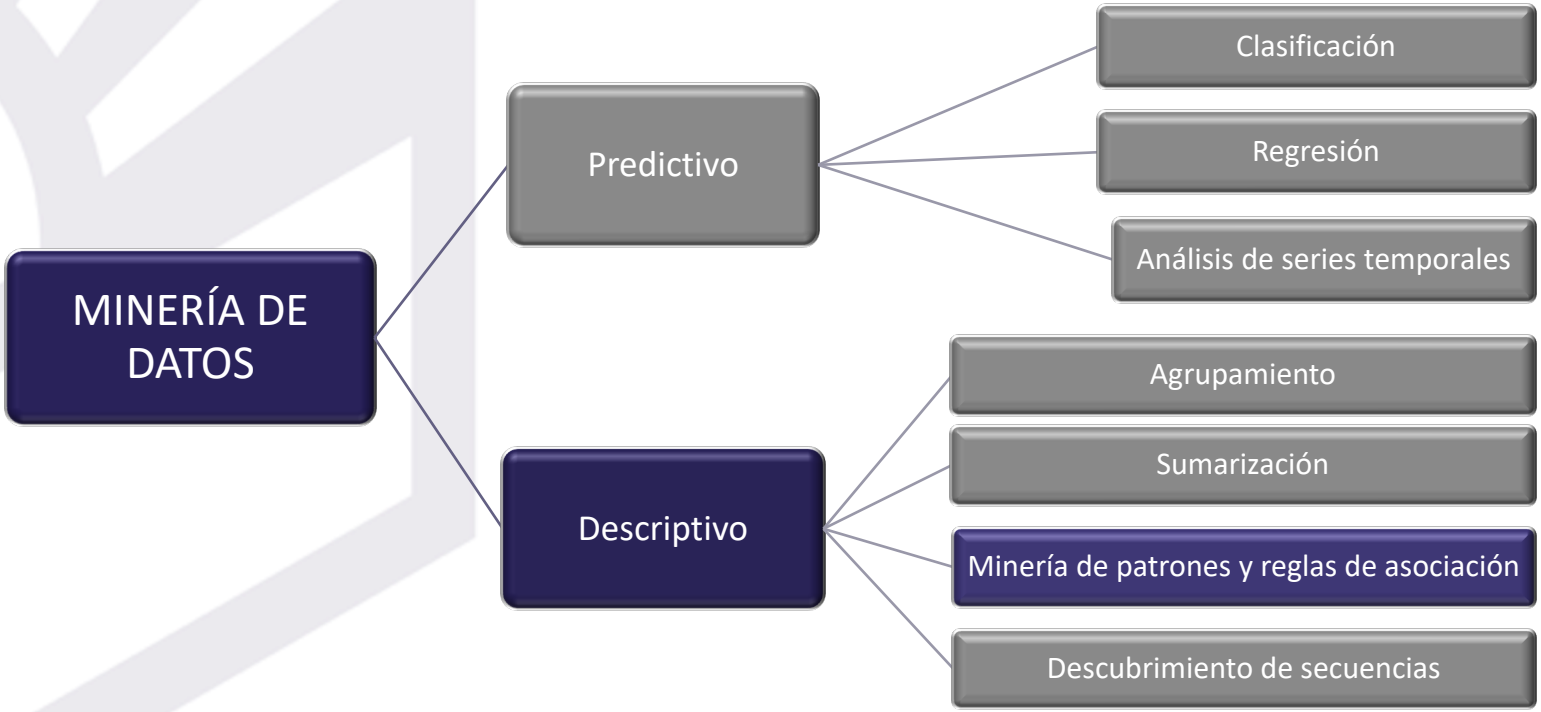
UCO ONLINE

Aprendizaje supervisado

- Requiere **conjunto de datos de entrenamiento**
- Al modelo se le dice a qué clase pertenece cada elemento del conjunto de entrenamiento
- Aprendizaje **basados en ejemplos**
- Ejemplo de técnica más característica: **Clasificación**

Aprendizaje no supervisado

- La **etiqueta Clase no es conocida** para el conjunto de entrenamiento
- El **número de clases puede ser desconocido**
- Aprendizaje **por observación**
- Ejemplos de técnicas relevantes: **Reglas de asociación**, ***Clustering***


MINERÍA DE DATOS
Predictivo
Clasificación
Regresión
Análisis de series temporales
Descriptivo
Agrupamiento
Sumarización
Minería de patrones y reglas de asociación
Descubrimiento de secuencias

- Permiten descubrir nuevos patrones y relaciones dentro de los datos
- Utilizados durante las fases de exploración de datos
- Algunas cuestiones típicamente contestadas por la minería de datos descriptiva son:
  - ¿Qué hay en los datos?
  - ¿Cómo parece la información?
  - ¿Hay patrones inusuales?
  - ¿Qué sugieren los datos de cara a la segmentación de clientes?
- Los usuarios pueden no tener idea del tipo de patrones que les pueden resultar de interés
- Los patrones pueden tener distintas granularidades:

Universidad – Facultad – Departamento – Área docente

Introducción a Frequent Pattern Mining

FPM y reglas de asociación

UCO ONLINE

Frequent itemset mining surge a principios de los 90 como un método de análisis de la cesta de la compra

El **objetivo** es encontrar algún tipo de uniformidad u homogeneidad en los hábitos de compra

Por ejemplo, encontrar conjuntos de productos que son frecuentemente comprados conjuntamente

TID	*Items*
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Coca-cola
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Coca-cola

<Pan, Leche> TIDs: 1, 4, 5

<Leche, Pañales> TIDs: 3, 4, 5

<Pan, Pañales, Cerveza> TIDs: 2, 4

Los *Frequent itemsets* pueden expresarse en modo de **reglas de asociación**:

SI un cliente compra Huevos **ENTONCES** es bastante probable que compre también Pan

Esta información es de suma relevancia pues permite:

- Organizar los productos
- Sugerir productos
- Detectar fraudes o comportamientos raros

TID	*Items*
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Coca-cola
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Coca-cola

La implicación implica co-ocurrencia,
no causalidad

El **orden de la regla** es muy importante, no siendo lo mismo:

SI un cliente compra Huevos **ENTONCES** es bastante probable que compre también Pan

SI un cliente compra Pan **ENTONCES** es bastante probable que compre también Huevos

TID	*Items*
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Coca-cola
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Coca-cola

SI Huevos **ENTONCES** Pan:

siempre que alguien compra huevos también compra pan

SI Pan **ENTONCES** Huevos:

comprar pan no es sinónimo de comprar huevos

Por tanto, el problema de **minería de reglas de asociación** consiste en:

Dados:

(1) base de datos de transacciones

(2) cada transacción es una lista de items (*p.ej. compras de un cliente en una determinada visita*)

Encontrar:

Todas las reglas que correlacionan la presencia de uno de esos conjuntos de elementos con la de otros conjuntos de items

* P.ej. El 98% de quien compra neumáticos, también adquiere el servicio de instalación de los mismos

Introducción a *Frequent Pattern Mining*

Conceptos básicos

UCO ONLINE

- **Patrones frecuentes:**
  - Patrones (conjuntos de elementos, secuencias, etc.) que ocurren frecuentemente en la base de datos

- **Minería de patrones frecuentes**: cumple con el objetivo de encontrar regularidades en los datos
  - ¿Qué combinación de productos se compra normalmente?
    - ¿Pañales y cerveza?
  - ¿Cuáles son las compras posteriores a adquirir un coche?
  - ¿Es posible establecer perfiles de nuestros clientes?

- **Transacciones,** $D = \{t_1, t_2, .., t_n\}$

  sea el conjunto de transacciones aquel en el que una transacción **t** es un *itemset* de la base de datos **D**

- ***Itemset* o conjunto de ítems**
  - Una colección de uno o más ítems

    <Leche, Pan, Pañales>
  - *k-itemset,* $I=\{I_1, I_2, \ldots, I_k\}$

    Un *itemset* que contiene *k* items

- **Soporte absoluto (σ)**
  - Frecuencia absoluta de aparición de un *itemset* en **D**

    σ(<Leche, Pan, Pañales>) = 2

- **Soporte relativo (s)**
  - Frecuencia relativa de transacciones en **D** que contienen un determinado *itemset*

    s(<Leche, Pan, Pañales>) = 2/5 = 0.4

TID	*Items*
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Coca-cola
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Coca-cola

- ***Itemset* frecuente**

  Un *itemset* cuyo soporte sea mayor o igual que un umbral ***minsup***

Espacio de búsqueda

Dados ***k*** elementos o ítems: $\mathbf{2^k-1}$

Para el ejemplo dado, existen: $2^6-1 = 63$ itemsets

<Pan>, <Leche>, <Pañales>, …. <Pan, Leche>, <Pan, Pañales>….., <Pan, Leche, Pañales, Cerveza, Huevos, Coca-cola>

$C_{6,1} = 6$ <Pan>, <Leche>….

$C_{6,2} = 15$ <Pan, Leche>, <Pan, Pañales>….

$C_{6,3} = 20$ <Pan, Leche, Pañales>…..

$C_{6,4} = 15$ <Pan, Leche, Pañales, Cerveza>….

$C_{6,5} = 6$ <Pan, Leche, Pañales, Cerveza, Huevos>……

$C_{6,6} = 1$ <Pan, Leche, Pañales, Cerveza, Huevos, Coca-cola>

TID	*Items*
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Coca-cola
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Coca-cola

Espacio de búsqueda prohibitivo para grandes conjuntos de datos (número de ítems):

Con sólo 50 ítems diferentes, tendríamos un total de 1.1259×10^{15} *itemsets* diferentes

¡¡Importante!! Se debe diferenciar claramente entre ítems y transacciones

Espacio de búsqueda

Reducir el enorme tamaño del espacio de búsqueda es un problema no trivial y supone un reto para FPM

- Se hace necesario para aplicaciones del mundo real
- Se emplea la propiedad anti monótona para reducir el espacio de búsqueda de forma efectiva

TID	Items
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Coca-cola
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Coca-cola

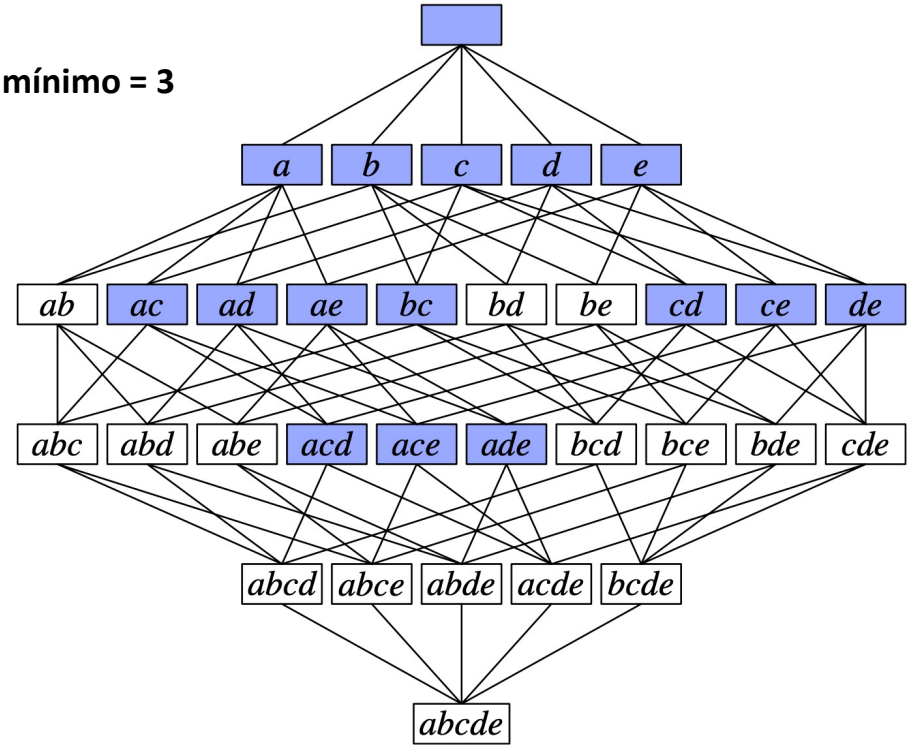
- **Propiedad anti monótona**
  - Ningún superconjunto de un conjunto infrecuente puede ser frecuente
    - σ(<Huevos>) = 1
    - σ(<Pan>) = 4
    - σ(<Huevos, Pan>) = 1
  - Todos los subconjuntos de un conjunto frecuente son frecuentes
    - σ(<Pan, Leche>) = 3
    - σ(<Pan>) = 4
    - σ(<Leche>) = 4

Si $X \subset Y$, entonces $sup(X) \geq sup(Y)$.
Por tanto: si $sup(X) \not\geq minSup$, entonces $\forall Y \supset X$, $sup(Y) \not\geq minSup$.

TIDs: Itemsets

1: $\{a, d, e\}$
2: $\{b, c, d\}$
3: $\{a, c, e\}$
4: $\{a, c, d, e\}$
5: $\{a, e\}$
6: $\{a, c, d\}$
7: $\{b, c\}$
8: $\{a, c, d, e\}$
9: $\{b, c, e\}$
10: $\{a, d, e\}$

Soporte mínimo = 3



¿Cómo encontrar los patrones de interés?

- Encontrar **todos** los patrones de interés: **Compleción**
  - ¿Puede un sistema de minería de datos encontrar **todos** los patrones de interés?
  - ¿Necesitamos encontrar **todos** los patrones de interés?
  - Búsqueda heurística Vs búsqueda exhaustiva

- Búscar **sólo** patrones de interés: **Problema de optimización**
  - ¿Puede un sistema de minería de datos encontrar **sólo** los patrones de interés?
  - **Propuestas**:
    a) Encontrar todos primero, filtrar los no interesantes después
    b) Generar sólo los patrones de interés (optimización de consultas)

Introducción a Frequent Pattern Mining

Medidas de interés

UCO ONLINE

La minería de datos puede generar miles de patrones

- No todos los patrones encontrados son interesantes

Medidas de interés:

Un patrón es **interesante** si resulta…

- **Fácilmente comprensible** por los humanos
- **Válido** para nuevos datos o datos de test con cierto grado de certidumbre
- Potencialmente **útil** y **novedoso**
- O **valida alguna hipótesis** que el usuario pretende confirmar

Medidas de interés objetivas

- Basadas en la estadística o en la estructura de los patrones
- Las **medidas más representativas**:
  - **Soporte**. Proporción de transacciones en el conjunto de datos que contienen el *itemset*.
    - **$X \Rightarrow Y$, $P(X \cup Y)$**: probabilidad de que una transacción contenga **X** e **Y**
  - **Confianza**. Grado de certeza de una asociación detectada
    - **P(Y I X)**: Probabilidad condicionada a que una transacción que contenga **X**, también contenga **Y**
- El usuario establece y controla los umbrales mínimos para estos valores (*thresholds*)
- Las reglas que no satisfagan un *threshold* mínimo de confianza no son consideradas interesantes

- Ejemplo de base de datos con 4 items y 5 transacciones

transaction ID	*milk*	*bread*	*butter*	*beer*
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Fuente: Wikipedia

  - El *itemset* {milk,bread,butter} tiene un **soporte** de 1 / 5 = 0.2
  - La regla {milk, bread} → {butter} tiene una **confianza** de 0.2 / 0.4 = 0.5

- **Medidas subjetivas:**
  - Basadas en las creencias y necesidades del usuario, como la **novedad**, **aplicabilidad** (*¿puede el usuario sacar provecho de la regla?*), **inesperabilidad** (*¿era desconocida o contraria al conocimiento actual?*), etc.

¡Gracias!

UCO ONLINE