Minería de Datos

Técnicas de Minería de datos – Análisis de asociación

Agenda:

- **Análisis de asociación**
- **Conceptos básicos**
- **Aprendizaje de reglas de asociación**
- **Algoritmo Apriori**

Análisis de asociación

- Dada una colección de registros o instancias

Donde cada registro contiene un conjunto de atributos o items, no hay salida definida

- Encontrar combinaciones o asociaciones de items (atributos) que ocurren frecuentemente.

- Objetivo: descubrir patrones que describen características fuertemente asociadas en los datos.

Útil para descubrir relaciones interesantes ocultas en los datos
Estas relaciones expresan patrones de comportamiento en función de la aparición conjunta de valores de dos o más atributos

Análisis de asociación

El análisis de asociación es aplicable a muchos dominios, por ejemplo:

- Búsqueda de patrones en páginas Web:

Determinar cuáles son los itinerarios más seguidos por los visitantes de un sitio Web.

- Análisis de peticiones de servicios médicos:

Encontrar aquellas pruebas o exámenes médicos que frecuentemente se realizan juntas.

- Análisis de ventas de productos:

Determinar la frecuencia de clientes que al comprar un producto A, seis meses después compran un producto B.

- Ciencias de la tierra:

Determinar las conexiones más interesantes entre cielo, tierra y procesos atmosféricos.

- Análisis de la cesta de compra:

Encontrar los productos que se compran juntos más frecuentemente

Conceptos básicos

- ¿Cómo son las reglas de asociación?

Dado un conjunto de transacciones (instancias), las reglas predicen la ocurrencia de un item (o conjunto de items) basado en la ocurrencia de otros items.

Ejemplo: análisis de la cesta de compra

ID	Items
1	{Pan, leche}
2	{Pan, servilletas, cerveza, huevos}
3	{leche, servilletas, cerveza, agua}
4	{Pan, leche, servilletas, cerveza}
5	{Pan, leche, servilletas, agua}

Cada fila = transacción = conjunto de items comprados por un cliente dado

Analizar estos datos puede ser de gran utilidad, ya que a partir de ellos se puede aprender el comportamiento de compras de los clientes

Esta información puede ser utilizada para la toma de decisiones sobre promociones de mercadeo, gestión de inventarios y manejo de las relaciones con los clientes

Conceptos básicos

- Términos básicos:

Los datos de la cesta de compras pueden representarse como una tabla binaria:

	IDt	Pan	Leche	Servilleta	Cerveza	Huevos	Agua
t_1	1	1	1	0	0	0	0
t_2	2	1	0	1	1	1	0
t_3	3	0	1	1	1	0	1
t_4	4	1	1	1	1	0	0
t_5	5	1	1	1	0	0	1

Un item = variable binaria

Se tiene entonces

$$I = \{i_1, i_2, i_3, \ldots, i_d\} = \text{Conjunto de items}$$

$$T = \{t_1, t_2, t_3, \ldots., t_N\} = \text{Conjunto de transacciones}$$

Cada transacción t_i contiene un subconjunto de items de I

Conceptos básicos

- ***Conjunto de items:*** **Conjunto de uno o más items.**

  Ejemplo: {leche, pan, servilletas}

- ***Conjunto de k items:*** **Conjunto que contiene exactamente k items**

- ***Soporte de un conjunto de items:*** **Número de transacciones que contienen ese conjunto de items (= frecuencia de ocurrencia de un conjunto de items)**

  Si X es un conjunto de items entonces su soporte será

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

Ejemplo: $\sigma(\{\text{cerveza, servilletas, leche}\}) = | \{ t_3, t_4 \} | = 2$

- ***Conjunto de items frecuente*****: Conjunto de items cuyo soporte es igual o mayor que un umbral** SOP_{Min}

Conceptos básicos

- ¿Cómo evaluar una regla de asociación?

Para la evaluación de las reglas de asociación se utilizan las siguientes métricas:

Sea la regla de la forma $X \longrightarrow Y$, donde X e Y son conjuntos de items

☞ ***Soporte*: fracción de transacciones que contienen a X y a Y. Determina cuán frecuentemente una regla es aplicable a un conjunto de datos.**

$$s\,(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

☞ ***Confianza:*** **medida de cuán frecuentemente los items en Y aparecen en transacciones que contienen a X. Mide la confiabilidad de una regla.**

$$c\,(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Conceptos básicos

- **El soporte de una regla es una medida muy importante. Reglas con bajo soporte no son significativas.**
- **A más alta la confianza, será más probable que Y esté presente en transacciones que contienen a X.**
- **La confianza también proporciona un estimado de la probabilidad condicional de Y dado X.**

Ejemplo:

IDt	Pan	Leche	Servilleta	Cerveza	Huevos	Agua
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

Conceptos básicos

Si se tiene la regla R: {leche, servilletas} ⟶ {cerveza}

- Soporte de la regla:

$$s(R) = \frac{\sigma(\{\text{cerveza, servilletas, leche}\})}{5} = \frac{2}{5} = 0.40$$

- Confianza de la regla:

$$c(R) = \frac{\sigma(\{\text{cerveza, servilletas, leche}\})}{\sigma(\{\text{servilletas, leche}\})} = \frac{2}{3} = 0.67$$

Aprendizaje de reglas de asociación

- ¿Cómo extraer reglas de asociación a partir de datos?

La tarea de minería de reglas de asociación se basa en:

Dado un conjunto de transacciones T, encontrar todas las reglas que tengan:

$$\text{Soporte} \geq \text{Sop}_{\text{Min}}$$

$$\text{Confianza} \geq \text{Conf}_{\text{Min}}$$

Umbrales para el soporte y la confianza de las reglas. Definidas por el usuario.

Un posible enfoque:

- **Determinar todas las posibles reglas**
- **Calcular el soporte y la confianza para cada regla**
- **Eliminar las reglas que no satisfagan los umbrales para el soporte y la confianza**

➡ **¡Muy ineficiente!**

Aprendizaje de reglas de asociación

La estrategia más común adoptada por muchos algoritmos de minería de reglas de asociación

➡ **Descomponer el problema en dos subtareas**

1.- Generación de los conjuntos de items frecuentes:

- **Encontrar todos los conjuntos de items que satisfacen el umbral Sop_{Min}**

El soporte de una regla depende sólo del soporte de su correspondiente conjunto de items $(X \cup Y)$.

2.- Generación de las reglas:

- **Encontrar todas las reglas de alta confianza (que cumplan con el umbral para la confianza $Conf_{Min}$), a partir de los conjuntos de items frecuentes encontrados en el paso previo. Esta reglas se llaman *reglas fuertes*.**

Aprendizaje de reglas de asociación

Ejemplo:

IDt	Pan	Leche	Servilleta	Cerveza	Huevos	Agua
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

Soporte mínimo: 0.40 → **Soporte de los CI = σ = 2**

Confianza mínima: 0.60

➡ **Algunos conjuntos de items frecuente con soporte mayor o igual a 0.40:**

{pan, leche}

{pan, servilleta}

{pan, cerveza}

{leche, cerveza}

{cerveza, servilletas, leche}

{pan, servilletas, cerveza}

Aprendizaje de reglas de asociación

Del conjunto: **{cerveza, servilletas, leche}**

Se pueden derivar las siguientes reglas:

	Confianza
{cerveza} ⟶ {servilletas, leche}	(2/3) = 0.66
{servilletas} ⟶ {cerveza, leche}	(2/4) = 0.50
{leche} ⟶ {cerveza, servilletas}	(2/4) = 0.50
{cerveza, servilletas} ⟶ {leche}	(2/3) = 0.66
{cerveza, leche} ⟶ {servilletas}	(2/2) = 1.00
{servilletas, leche} ⟶ {cerveza}	(2/3) = 0.66

Las reglas seleccionadas son:

R1: Si {cerveza} entonces {servilletas, leche}
R2: Si {cerveza, servilletas} entonces {leche}
R3: Si {cerveza, leche} entonces {servilletas}
R4: Si {servilletas, leche} entonces {cerveza}

Aprendizaje de reglas de asociación

- ¿Cómo determinar un modelo de asociación?

Algunas técnicas:

- **Algoritmo Apriori**
- **Algoritmo Eclat**
- **Algoritmos genéticos**
- **Métodos para descubrir reglas de asociación secuenciales**
- **Otros**

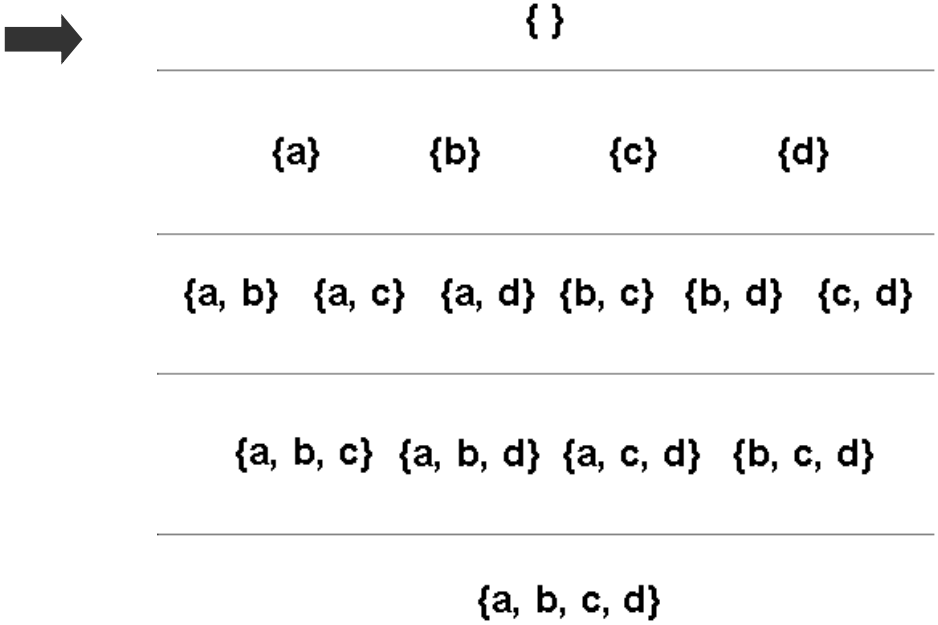
Algoritmo Apriori

Primer algoritmo de minería de reglas de asociación en utilizar la búsqueda basada en el soporte de los conjuntos de items.

A) Generación de los conjuntos de items frecuentes

Ejemplo: Sea I = {a, b, c, d}

Se deben tomar en cuenta todas las posibles combinaciones de items



¿Cómo reducir la complejidad de la búsqueda?

Reducir el número de posibles items candidatos

Algoritmo Apriori

Principio Apriori: **Utilizar el soporte para reducir el número de conjuntos de items (CI) explorados durante la generación de los conjuntos de items frecuentes**

Entonces,

a) Si un conjunto de items es frecuente entonces todos sus subconjuntos también serán frecuentes.

Ejemplo: si {b, c, d} es un CI frecuente (cumple con el umbral para el soporte)

entonces {b}, {c}, {d}, {b, c}, {b, d}, {c , d} también serán CI frecuentes

b) Si un conjunto de items es infrecuente entonces todos sus superconjuntos (aquellos que lo contienen) también serán infrecuentes.

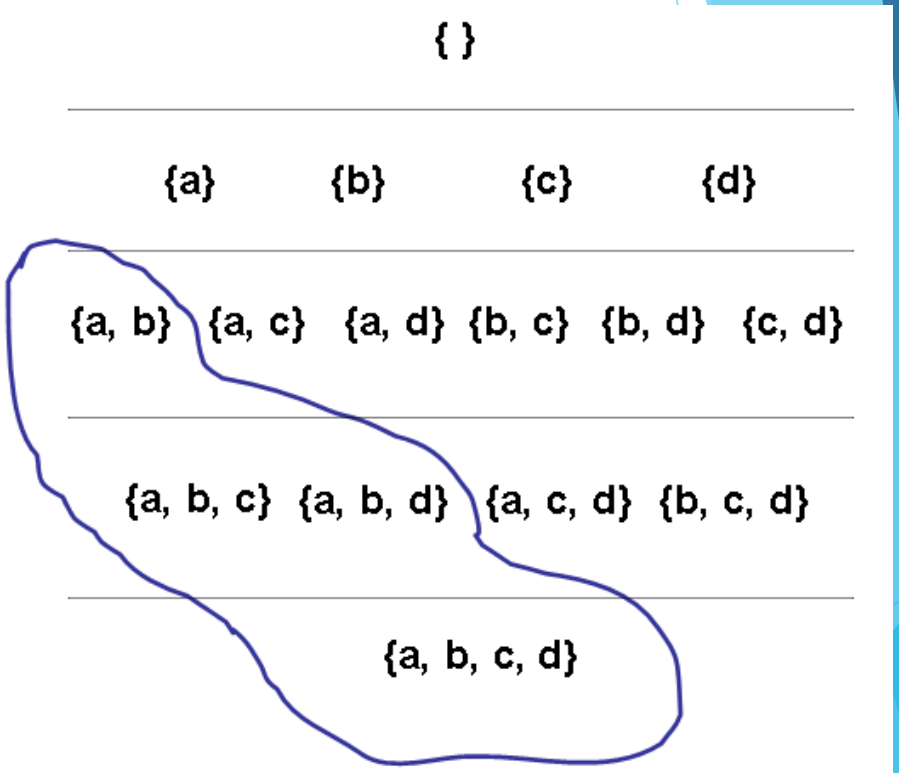
Ejemplo: si {a, b} es un CI infrecuente (no cumple con el umbral para el soporte)

entonces {a, b, c}, {a, b, d}, {a, b,c, d} también serán CI frecuentes

El principio b) permite mejorar la búsqueda en el espacio de CI candidatos

Algoritmo Apriori

Este principio se utiliza para evaluar cada posible nodo (CI) del árbol de búsqueda:



Si {a, b} es un CI infrecuente, no tiene sentido seguir la búsqueda por esta rama

Algoritmo Apriori

Ejemplo:

IDt	Pan	Leche	Servilleta	Cerveza	Huevos	Agua
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

1. Inicialmente cada item se considera como un CI de 1 item (1-itemset)

Sop_{Min} = 60% → **El soporte que deben cumplir los CI es:**

N x 0.60 = 5 x 0.60 = 3

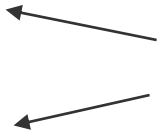
1-itemset	Soporte
{cerveza}	3
{pan}	4
{agua}	2
{servilleta}	4
{leche}	3
{huevos}	1

Se descartan, ya que no cumplen el soporte mínimo

Algoritmo Apriori

2. En la siguiente iteración, los CI con 2 items (2-itemset) se generan utilizando sólo los conjuntos de items frecuentes de 1 item, debido al principio Apriori

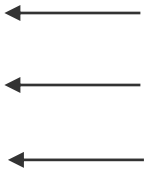
2-itemset	Soporte
{cerveza, pan}	2
{cerveza, servilleta}	3
{cerveza, leche}	2
{pan, servilleta}	3
{pan, leche}	3
{servilletas, leche}	3



Se descartan, ya que no cumplen el soporte mínimo

3. En la siguiente iteración, los CI con 3 items (3-itemset) se generan utilizando sólo los conjuntos de items frecuentes de 2 item.

3-itemset	Soporte
{cerveza, servilleta, pan}	2
{cerveza, servilleta, leche}	2
{pan, servilleta, leche}	2



Se descartan, ya que no cumplen el soporte mínimo