Minería de Datos

Técnicas de Minería de datos
Asociación - ejemplos

Agenda:

- **Aprendizaje de reglas de asociación**
- **Algoritmo Apriori**
- **Aplicaciones**

Aprendizaje de reglas de asociación

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*.**

Algoritmo Apriori

A) Generación de los conjuntos de items frecuentes

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

➡ ***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 infrecuentes

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

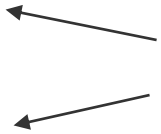
1-itemset	Soporte
{cerveza}	3
{pan}	4
{agua}	2
{servilleta}	4
{leche}	3
{huevos}	1

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 a principio Apriori

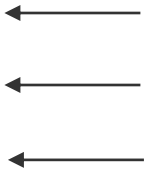
2-itemset	Soporte
{cerveza, pan}	2
{cerveza, servilleta}	3
{cerveza, leche}	2
{pan, servilleta}	3
{pan, leche}	3
{servilletas, leche}	3



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

B) Generación de las reglas de asociación

¿Cómo extraer reglas de asociación de manera eficiente a partir de los conjuntos de items frecuentes?

- **Las reglas que son generadas a partir de CI frecuentes satisfacen el umbral del soporte.**
- **Cada CI de k items puede generar (2^k - 2) reglas.**

Ejemplo: **del conjunto {cerveza, servilletas, leche}**

Se pueden derivar las siguientes reglas:

{cerveza} ⟶ {servilletas, leche}

{servilletas} ⟶ {cerveza, leche}

{leche} ⟶ {cerveza, servilletas}

{cerveza, servilletas} ⟶ {leche}

{cerveza, leche} ⟶ {servilletas}

{servilletas, leche} ⟶ {cerveza}

Algoritmo Apriori

- **Calcular la confianza de una regla de asociación no requiere búsquedas adicionales en el conjunto de datos**

Ejemplo:

Para la regla {cerveza, servilletas} ⟶ {leche}

Generada a partir del CI frecuente {leche, servilletas, cerveza}

La confianza será $\dfrac{\sigma(\{\text{cerveza, servilletas, leche}\})}{\sigma(\{\text{cerveza, servilletas}\})}$ ⟵

Es un CI frecuente debido al principio Apriori

El soporte de ambos ya fue calculado durante la generación de los CI frecuentes

Algoritmo Apriori

Para explorar el espacio de posibles reglas se puede utilizar el siguiente teorema:

Sea Y un CI frecuente, entonces:

Si una regla $X \longrightarrow Y - X$ no satisface el umbral de la confianza entonces cualquier regla $X' \longrightarrow Y - X'$ (donde $X' \subseteq X$), tampoco lo hará.

Ejemplo:

Y = {cerveza, servilletas, leche, pan}

Si la regla: {cerveza, servilletas, leche} $\longrightarrow$ {pan}

No satisface el umbral para la confianza, entonces tampoco lo hará la regla:

{cerveza, servilletas} $\longrightarrow$ {pan, leche}

Algoritmo Apriori

Demostración:

La confianza de las reglas es:

$$\frac{\sigma(\{X\} \cup \{Y - X\})}{\sigma(\{X\})} = \frac{\sigma(\{Y\})}{\sigma(\{X\})}$$

$$\frac{\sigma(\{X'\} \cup \{Y - X'\})}{\sigma(\{X\})} = \frac{\sigma(\{Y\})}{\sigma(\{X'\})}$$

Como $X' \subseteq X$ entonces
$\sigma(\{X'\}) \geq \sigma(\{X\})$

$$\frac{\sigma(\{Y\})}{\sigma(\{X\})} \geq \frac{\sigma(\{Y\})}{\sigma(\{X'\})}$$

La confianza de la regla $X' \longrightarrow Y - X'$ no puede ser mayor que la confianza de la regla $X \longrightarrow Y - X$ (donde $X' \subseteq X$).

Algoritmo Apriori

- **En el algoritmo Apriori se utiliza un enfoque por niveles para generar reglas de asociación, donde cada nivel corresponde al número de items en el consecuente de una regla.**
- **Inicialmente se generan todas las reglas de alta confianza con un item como consecuente. Estas reglas son entonces utilizadas para generar las reglas con dos items en el consecuente y así sucesivamente**

Ejemplo:

De las reglas

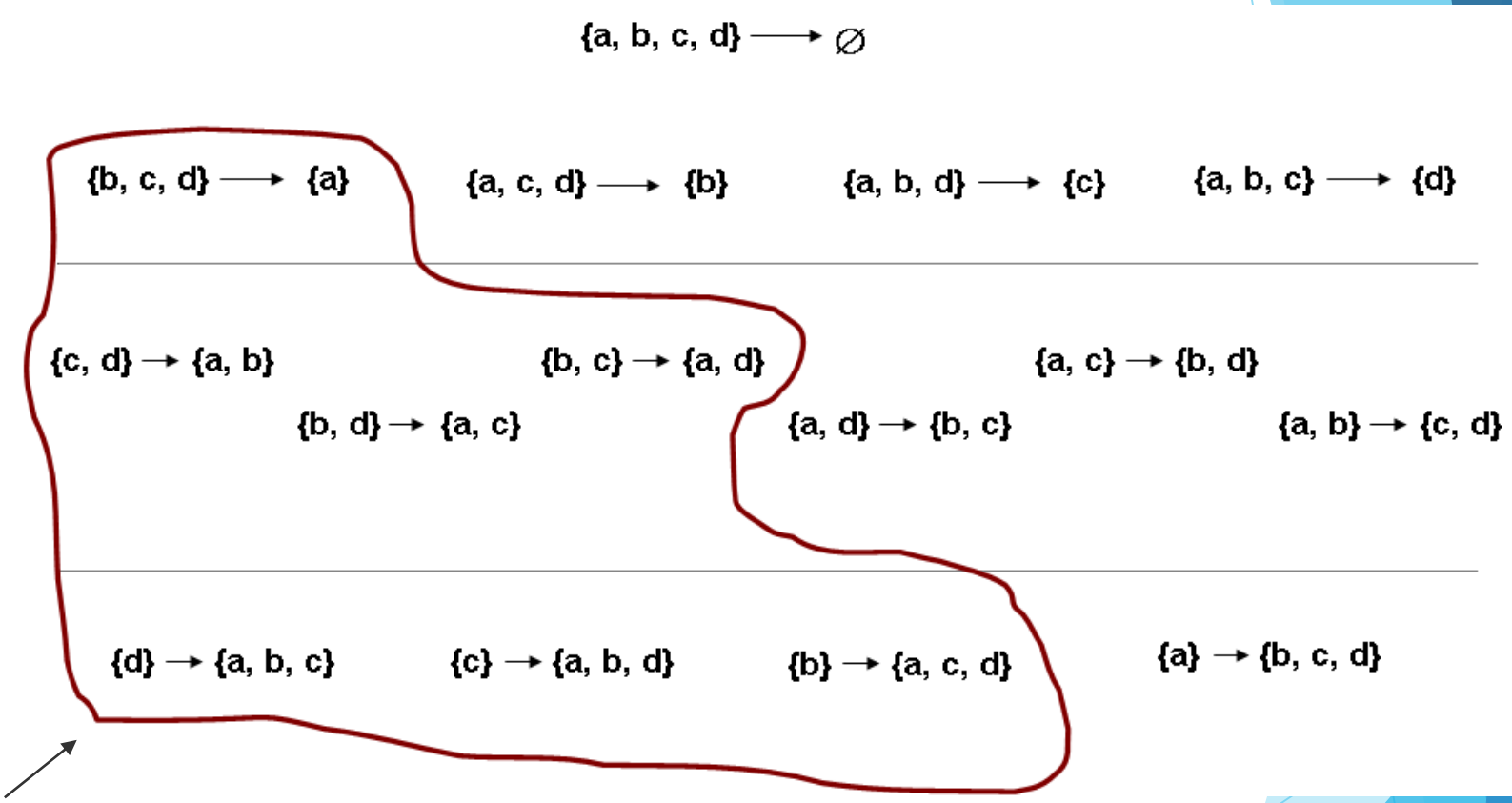
{a, c, d} ⟶ {b}

{a, b, d} ⟶ {c}

Se genera la regla ⟶ {a, d} ⟶ {b, c}

Entonces, el árbol de búsqueda para el Cl {a, b, c, d} será

Algoritmo Apriori



Puede ser eliminado

➡ Si la confianza de {b, c, d} ⟶ {a} no supera el umbral entonces todas las reglas que tengan al item "a" en el consecuente tampoco lo harán

Algoritmo Apriori

Ejemplo: **{cerveza, servilleta, leche} es CI frecuente**

¿Cuántas reglas se pueden generar si la confianza es del 70%?

1er. Nivel:

			Confianza:	
{cerveza, servilletas}	→	**{leche}**	**(2/3) = 0.66**	
{cerveza, leche}	→	**{servilletas}**	**(2/2) = 1.00**	←
{servilletas, leche}	→	**{cerveza}**	**(2/3) = 0.66**	

Se considera sólo esta regla en el siguiente nivel

2do. Nivel:

			Confianza:
{cerveza}	→	**{servilletas, leche}**	**(2/3) = 0.66**
{leche}	→	**{cerveza, servilletas}**	**(2/3) = 0.66**

➡ **La única regla sería** **{cerveza, leche}** → **{servilletas}**

Este procedimiento se repite para cada CI frecuente

Ejercicio

Dado el conjunto de datos de transacciones:

Transacción	Items comprados
1	{a, e}
2	{a, b, c, e}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, e}
6	{b, d, e}
7	{c, d}
8	{b, c, e}
9	{a, d, e}
10	{a, b, e}

Si el umbral del soporte es del 60% y el umbral de la confianza es del 80%
¿Cuáles son las reglas de asociación que se generarían con el algoritmo Apriori?

Aplicaciones

- Minería de uso de la Web

Es posible descubrir todas las asociaciones entre los accesos y usos de la Web por parte de los usuarios. Cada transacción consiste de un conjunto de URLs accedidas por un cliente en un servidor. Aplicando reglas de asociación se pueden encontrar reglas como:

El 40% de los clientes que accedieron la página con URL /entidad/productos/producto1.html también accedieron a /entidad/productos/producto2.html"

Esta información puede servir para mejorar el sitio Web

- Uso de Técnicas no Supervisadas en la Construcción de Modelos de Clasificación en Ingeniería del Software

(Moreno, M. y* otros. *Tendencias de la Minería de datos en España. pp:143-153. 2007. ISBN: 84-688-8442-1)

Se utiliza la técnica de reglas de asociación para construir un modelo de clasificación para realizar estimaciones en al área de Ingeniería de Software

Aplicaciones

- **Aplicación de Minería de Datos para la extracción de reglas en Objetos de Aprendizaje**

(*Zapata, A. y otros. *Recursos Digitales para el Aprendizaje. EditorialUADY_México_ISBN_9876077573173)

Se aplica la minería de datos para la extracción de reglas de asociación, utilizando el algoritmo Apriori, que permitan la caracterización de objetos de aprendizaje.

Aplicaciones

➡ **Estudio descriptivo del nivel de interacción de los estudiantes de postgrado en la plataforma Moodle de la Facultad de Ciencias**
(Luis Arredondo)

PROBLEMA:

☞ **¿Cómo los estudiantes utilizan estas herramientas?**

☞ **¿Este uso puede dar evidencias de un perfil de los estudiantes?**

Objetivo: ***Determinar diferentes comportamientos de los estudiantes de postgrado en cursos dictados a través de la plataforma Moodle***

Tarea de minería de datos: Asociación

Aplicaciones

Recolección de los datos:

- **Para cada usuario matriculado en un curso dado, se logró recopilar su registro de actividades en la plataforma Moodle.**

Preparación de los datos:

- **Se construyeron variables en función de los porcentajes de utilización en Foros, Tareas, Recursos, Chats, SCORMS, Wiki y Cuestionario.**
- **Se encontraron ausencias en algunas variables; sin embargo, tenían una interpretación asociada con la interacción del estudiante en la plataforma (no utilizó el recurso).**
- **Selección de variables: Algoritmo InfoGainAttributeEval con el método Ranker. Se seleccionaron las variables:**

%Foros

%Tareas

%Recursos

%Chats

Aplicaciones

- **Se efectuó una discretización de las variables a intervalos equidistantes.**

Intervalo	Etiqueta
[0, 0.2]	Muy bajo
(0.2, 0.4]	Bajo
(0.4, 0.6]	Medio
(0.6, 0.8]	Alto
(0.8, 1]	Muy alto

A cada intervalo se le asignó una etiqueta asociada al nivel de uso de la herramienta.

Aplicaciones

Minería de datos:

- **Tarea de minería de datos: Asociación**
- **Lenguaje de representación: no es un requerimiento**
- **Algoritmo: Apriori**
- **Medida de rendimiento: Soporte y confianza**
- **Confianza = 75%, Soporte = 50%**

Aplicaciones

Reglas encontradas:

R1: El 100% de los individuos en cuyos porcentajes tanto de entrega de tareas y de participación en chats fueron muy bajos también lo fue el porcentaje de participación en los foros.
R2: El 100% de los individuos con un porcentaje de tareas y de recursos muy bajos también tienen un porcentaje muy bajo de participación en los foros.
R3: El 100% de los estudiantes con un nivel bajo de entrega de tareas también tienen un porcentaje muy bajo de participación en los foros.
R4: Un 88% de los estudiantes cuya participación en los chats en muy baja también tienen un porcentaje muy bajo de participación en los foros.
R5: El 86% de los estudiantes cuyos niveles de participación es muy bajo tanto en foros como en chats también tienen un nivel muy bajo de entrega de tareas propuestas.
R6: El 75% de los individuos con un porcentaje muy bajo de participación en los chats también tienen un nivel muy bajo en la participación en los foros y en la entrega de tareas.
R7: El 75% de los individuos con un nivel muy bajo de participación en los chats también tienen un nivel muy bajo de entrega de tareas.