Práctica 5: Reglas de Asociación

Introducción

"*Las reglas de asociación expresan patrones de comportamiento entre los datos en función de la aparición conjunta de valores de dos o más atributos*" [1].

Las reglas de asociación se trabajan sobre atributos cualitativos. Establecen relaciones entre conjuntos de elementos de la forma SI – ENTONCES:

Si **X** ENTONCES **Y** o bien, **X** => **Y**.

Donde X es el *antecedente* e Y el *consecuente*. Además, X e Y no tienen elementos comunes entre sí y ambos son subconjuntos de $\boldsymbol{I}$.

$\boldsymbol{I} = \{i_1, i_2, \ldots, i_n\}$ es un conjunto de n atributos binarios que se llaman ítems.

$\boldsymbol{D} = \{t_1, t_2, \ldots, t_n\}$ es un conjunto de transacciones almacenadas en una base de datos.

Algunas definiciones:

Itemset: colección de uno o más ítems del conjunto $\boldsymbol{I}$.

k-itemset: conjunto de k ítems.

Support count (conteo de soporte): cantidad de transacciones en $\boldsymbol{D}$ que contienen un itemset.

Itemset frecuente: itemset cuyo soporte es mayor o igual a un soporte determinado (mínimo soporte).

Las métricas relacionadas a una regla de asociación, para conocer su calidad, son el soporte o cobertura (support) y la confianza (confidence). El soporte (S) es la cantidad de transacciones que contienen a los subconjuntos X e Y sobre el total de transacciones (n). La confianza (C) es la cantidad de transacciones que contienen a los subconjuntos X e Y sobre las que contienen a X.

Las reglas de asociación se expresan de la siguiente forma:

$$\mathbf{X} \rightarrow \mathbf{Y} \ (S,C)$$

Aplicación

Las reglas de asociación se utilizan generalmente para tareas de tipo descriptiva, es decir, encontrar patrones en los datos de atributos aparentemente independientes.

Se utilizan en distintos tipos de "negocio", por ejemplo: en medicina para ayudar a diagnosticar pacientes; en ventas, para determinar que productos tienen mayor probabilidad de ser comprados juntos; o en la experiencia de usuario de páginas web, recopilando información para optimizar la interfaz de acuerdo a donde hacen click los usuarios en la página.

Algoritmos

Existen varios algoritmos para encontrar las reglas de asociación: Apriori, FP-Growth y Eclat, entre otros. Para esta clase se utilizará el algoritmo Apriori.

Apriori

Este algoritmo fue uno de los primeros desarrollados para la búsqueda de reglas de asociación. Tiene dos etapas:

1. Identificar los itemsets frecuentes.
2. Encontrar las reglas de esos itemsets frecuentes.

Ver ejemploApriori.pdf.

Implementación en R

Para implementar el algoritmo A priori para encontrar itemsets frecuentes y reglas de asociación se utilizará la función *apriori()* del paquete **arules** [2]. Los argumentos de la función son:

- Data: objeto de la clase *transactions*.
- Parameter: lista con los parámetros para el algoritmo (valor mínimo de soporte y confianza) y el objetivo que se busca, itemsets frecuentes o reglas de asociación con la opción **target** igual a "frequent itemsets" o “rules” respectivamente.

Para cargar las transacciones, se utiliza la función *read.transactions()*, que tiene los siguientes parámetros:

- File: archivo de datos.
- Format: formato de los datos (“single” o “basket”).
- Header: variable lógica, indica si se encuentran los nombres de las variables en la primera fila del archivo.
- Sep: caracter que indica cómo están separados los datos del archivo.
- Cols: para el formato “single”, es un vector de longitud dos para indicar el nombre o número de las columnas.
- Rm.duplicates: variable lógica para indicar si se quiere eliminar ítems duplicados.

El objeto de tipo transactions es una matriz booleana con las transacciones en las filas y los ítems en las columnas; con el valor 1 si el ítem está presente en la transacción y 0 si no está presente.

Formato de los datos

La forma en que se expresan los datos en una base de datos de transacciones puede ser de dos tipos: horizontal o *basket*, donde se representa un conjunto de artículos por fila; o vertical o *single*, donde se representa una columna con los IDs de las transacciones y otra columna con los artículos. En las siguientes tablas se pueden ver los dos formatos.

Formato *basket*	
TID	**Ítems**
1	{leche, cerveza}

2	{leche, pan, huevos}
3	{pan, servilletas}
4	{leche, pan, huevos, servilletas}

Formato *single*	
TID	**Ítem**
1	Leche
1	Cerveza
2	Leche
2	Pan
2	Huevos
3	Pan
3	Servilletas
4	Leche
4	Pan
4	Huevos
4	Servilletas

.

Actividades

1 – El archivo “orders.csv” contiene 12500 transacciones de un supermercado.

a) Explore los datos. ¿Qué puede decir de la forma en que están representados?
b) Encuentre los itemsets frecuentes para un soporte de 1% (el soporte mínimo tiene que ser bajo ya que el número de ítems y transacciones es muy alto) y muestre los 5 itemsets frecuentes de mayor soporte.
c) Obtenga las reglas de asociación para todas las transacciones del supermercado para una confianza mínima de 0,7.

2 – En el archivo titanic.raw.rdata se encuentra la información expandida del dataset de R Titanic, que nos muestra la clase, el sexo y el rango de edad de los pasajeros; y si sobrevivieron o no.

a) Explore los datos. ¿Qué características tiene la representación de los datos?
b) Encuentre las reglas de asociación para un soporte del 0,5% y una confianza del 80%, donde el consecuente sea si ese pasajero sobrevivió o no.

Referencias

1. Orallo, J. et all. “Introducción a la Minería de Datos” (2004). Capítulo 9.
2. Paquete arules. Disponible en: https://cran.r-project.org/web/packages/arules/arules.pdf