Práctica 6: Reglas de Asociación – Parte 2

Algoritmo Eclat

Este algoritmo para encontrar patrones frecuentes fue propuesto en el año 2000 por Zaki. Se diferencia de Apriori en la forma en que analiza las transacciones, ya que lo hace de forma vertical. Es decir, en cada fila contiene un ítem y las transacciones en las que aparece ese ítem. El conjunto de transacciones en las que aparece el ítem se denomina *tidset*.

Este enfoque es el que lo hace más rápido que el algoritmo Apriori.

Eclat es un algoritmo para obtener itemsets frecuentes, no para generar reglas de asociación directamente. Para ellos, una vez que se han obtenido los itemsets frecuentes, se utiliza la segunda parte del algoritmo Apriori para generar las reglas.

Otras métricas

Además de las métricas ya vistas: el soporte y la confianza, se puede utilizar el estadístico *lift*. Este compara la frecuencia observada de una regla con la frecuencia esperada por el azar (no existe la regla).

El lift de la regla X => Y es igual a:

$$Lift(\mathbf{X} => \mathbf{Y}) = c(\mathbf{X} => \mathbf{Y}) / s(\mathbf{Y})$$

Es la proporción de transacciones en las que la presencia del antecedente da lugar a la presencia del consecuente.

Desde el punto de vista de la probabilidad:

$$Lift(X => Y) = P(X|Y) / P(Y)$$

Interpretación

Lift(X => Y) > 1	La probabilidad de que ocurra el consecuente, dado que ocurrió el antecedente, aumenta.
Lift(X => Y) = 1	No hay incidencia entre las probabilidades de ocurrencia.
Lift(X => Y) < 1	La probabilidad de que ocurra el consecuente, dado que ocurrió el antecedente, disminuye.

Implementación en R

Para implementar el algoritmo Eclat para encontrar itemsets frecuentes y luego, reglas de asociación se utilizará la función *eclat()* 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, número mínimo o máximo de ítems por itemset).

Para cargar las transacciones, se utiliza la función *read.transactions()*, de la misma forma que en la guía anterior.

Una vez que se generaron los itemsets frecuentes con el algoritmo Eclat, las reglas de asociación se obtienen utilizando otra función del paquete **arules**: *ruleInduction()*. Que tiene los siguientes parámetros:

- x: intemsets frecuentes de donde se pretende obtener las reglas.
- transactions: si los itemsets encontrados tienen un tamaño mínimo, se debe utilizar el objeto con las transacciones. Si no, es NULL por defecto.
- confidence: mínimo valor de confianza.

La función aplica el algoritmo Apriori para encontrar las reglas.

Otras funciones de arules

Con la función *subset()* se pueden filtrar tanto itemsets, reglas o transacciones. Se puede utilizar un criterio que sea conveniente para el problema y elegir los elementos de interés. Los parámetros que se utilizarán son:

- x: objeto a filtrar.
- subset: expresión lógica para seleccionar los elementos.

Para construir la expresión lógica se pueden utilizar, además de los operadores lógicos tradicionales, los siguientes operadores:

Operador	Significado
%in%	Cualquiera de los siguientes elementos.
%ain%	Contiene todos los siguientes elementos.
%oin%	Contiene solo los siguientes elementos.
%pin%	Contiene parcialmente los siguientes elementos.

La función *itemFrequency()* devuelve el soporte de los elementos individuales de un conjunto de ítems.

La función *interestMeasure()* devuelve medidas de interés sobre un conjunto de itemsets o reglas. Los parámetros necesarios son:

- x: el conjunto de interés.
- measure: nombre o vector de nombres con las medidas de interés (“coverage”, “fishersExactTest”, etc. [2]).

Actividades

1 – El archivo “groceries.csv” contiene las compras de los clientes de un supermercado.

a) Explore los datos. ¿En qué forma están representados los datos? ¿Cuántas transacciones y cuántos ítems contienen los datos?
b) Con el resultado de la función *itemFrequency()* [2] puede encontrar el soporte de cada elemento en el conjunto de datos. Utilice un criterio estadístico adecuado para definir el soporte mínimo y encuentre los 5 itemsets frecuentes de mayor soporte.
c) Obtenga las reglas de asociación para todas las transacciones del supermercado para una confianza mínima de 0,5 y de 0,25. ¿Cuántas reglas encuentra en cada caso? ¿Qué puede decir de las reglas obtenidas para el caso de menor confianza?
d) Para las reglas obtenidas con una confianza de 0,25, obtenga:
    - las reglas que tienen un lift mayor a 2;
    - las reglas que contienen el elemento “other vegetables” en el consecuente con un lift mayor a 2;
    - las reglas que contienen solamente “other vegetables" y "yogurt” en el antecedente y un lift mayor a 2; y
    - las reglas que contienen “other vegetables" o "yogurt” en el consecuente y un lift mayor a 2.

2 – Utilice el archivo “orders.csv” de la actividad 1 de la guía anterior y obtenga:

a) Los itemsets frecuentes con el algoritmo Eclat, para un soporte mínimo de 0.1% y una longitud mínima de 3 ítems.
b) Encuentre las reglas de asociación para una confianza mínima del 70%, con los itemsets frecuentes obtenidos anteriormente y ordenelas por *lift*.
c) Compare las reglas obtenidas con el algoritmo Eclat, con las obtenidas para los mismos valores de soporte y confianza con el algoritmo Apriori.
d) Compare las métricas comentadas en la teoría (coverage, test exacto de Fisher y lift) para los dos conjuntos de reglas encontrados en el punto anterior.

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