Práctica 8: Análisis de secuencias

Introducción

Dentro de las tareas de minería de datos descriptivas se encuentran las reglas de asociación y las reglas de asociación secuenciales. Las secuenciales son un caso especial de las reglas de asociación, donde se intenta obtener patrones que se basan en secuencias temporales: las relaciones entre los datos dependen del tiempo [1].

En el ejemplo clásico de reglas de asociación, donde se tienen transacciones de un supermercado, para el caso de análisis secuencial cada transacción incluirá un atributo que indica el tiempo en el que ocurrió. Aquí importa el orden de ocurrencia.

Entonces, cada transacción tendrá los datos del "cliente", el "tiempo" de ocurrencia de la transacción y los "ítems" de la misma.

Algoritmo SPADE

SPADE (Sequential Pattern Discovery using Equivalent Class) es un algoritmo propuesto por Zaki en 2001 para encontrar secuencias frecuentes, basado en técnicas de búsqueda en enrejado y uniones simples. Básicamente divide el problema en "subproblemas" optimizando el manejo de memoria.

En SPADE, la base de datos secuencial se transforma a un formato de "lista de identificación vertical" o *ID-list*, donde se asocia cada ID con los ítems correspondientes y en un instante de tiempo [2].

El primer paso de SPADE es calcular las frecuencias de las 1-secuencias (con un solo elemento). Luego se cuentan las secuencias de 2 elementos, transformando la representación vertical en una representación horizontal y contando el número de secuencias para cada par de elementos utilizando una matriz bidimensional.

Las n-secuencias siguientes pueden formarse uniendo las (n-1)-secuencias mediante sus listas de identificación. El tamaño de las listas de identificación es el número de secuencias en las que aparece un elemento. Si este número es mayor que el soporte mínimo, la secuencia es frecuente. El algoritmo se detiene cuando ya no se encuentran secuencias frecuentes. El algoritmo puede utilizar un método de búsqueda en profundidad o en amplitud para encontrar nuevas secuencias.

Implementación en R

Para implementar el algoritmo SPADE y encontrar secuencias frecuentes, se utilizará la función *cspade()* del paquete **arulesSequences** [3]. 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 máximo de ítems por itemset de una secuencia, número máximo de itemsets de una secuencia, etc.).

La información temporal se toma de los componentes sequenceID (identificador de secuencia o cliente) y eventID (identificador de evento) de *transactionInfo()* [4]. Tenga en cuenta que los identificadores enteros deben ser positivos y que las transacciones deben estar ordenadas por sequenceID y eventID.

Devuelve un objeto de la clase *sequences*. Es una colección de secuencias y sus medidas asociadas (soporte).

Otra función del paquete **arulesSequences** que se va a utilizar cuando se tienen datos en forma de lista vertical es *read_baskets().* Para abrir los datos en formato basket y crear un objeto de tipo *transactions*. Sus parámetros son:

- con: nombre del archivo.
- sep: expresión regular que indica que campos están separados en el archivo.
- info: un vector de caracteres con los nombres del “header” de las columnas.

Actividades

1 – El dataset “zaki” del paquete arulesSequences, es un conjunto de transacciones generadas a partir de un conjunto de datos de eventos.

Los datos originales se pueden observar en la siguiente tabla:

ID secuencia	Tiempo	N° items	Items
1	10	2	C, D
1	15	3	A, B, C
1	20	3	A, B, F
1	25	4	A, C, D, F
2	15	3	A, B, F
2	20	1	E
3	10	3	A, B, F
4	10	3	D, G, H
4	20	2	B, F
4	25	3	A, G, H

a) Explore los datos. ¿En qué forma están representados los datos? ¿Cuántas transacciones y cuántos ítems contienen los datos?

b) Encuentre las secuencias frecuentes para los siguientes valores de soporte mínimo de 0,1, 0,25, 0,4 y 0,7.

c) Para las secuencias encontradas con un soporte mínimo de 0,4, indique:

- ¿Cuáles son los ítems individuales más frecuentes?
- ¿Cuáles son los itemsets más frecuentes en los eventos?
- El conjunto de secuencias encontrado ordenado por el valor de soporte.

2 – En el archivo “sequences.txt” se recoge información pública de una página web, que muestra los favoritos de todos los usuarios. Cada favorito consta del usuario, la URL y las etiquetas que el usuario eligió para describirla.

La línea de tiempo pública muestra los favoritos del sistema en una secuencia de tiempo, de modo que podemos obtener secuencias de favoritos para usuarios específicos. En este caso, cada favorito es un evento, y cada etiqueta es un elemento.

a) Cargue el archivo de forma adecuada para obtener las transacciones en formato basket requeridas por la función *cspade()*. ¿Cuántas transacciones y cuántos ítems contiene? ¿Cuáles son los ítems más frecuentes?

b) Extraiga patrones temporales con el objetivo de generar reglas que predigan etiquetas útiles para un usuario específico (para un soporte mínimo de 0,2%).

Referencias

1. Orallo, J. et all. “Introducción a la Minería de Datos” (2004). Capítulo 9.

2. Zhao Q. y Bhowmick S., “Sequential Pattern Mining: A Survey” (2003). Disponible en: https://cs.nju.edu.cn/zhouzh/zhouzh.files/course/dm/reading/reading04/zhao_techrep03.pdf

3. Paquete arulesSequences. Disponible en: https://cran.r-project.org/web/packages/arulesSequences/arulesSequences.pdf

4. Paquete arules. Disponible en: https://cran.r-project.org/web/packages/arules/arules.pdf