

Minería de Datos: Aprendizaje no Supervisado

Reglas de asociación: Trabajo en laboratorio

Descripción

El objetivo de esta práctica es realizar un análisis sobre una base de datos (BD) seleccionada por el estudiante utilizando técnicas de extracción de reglas de asociación (RAs). La base de datos seleccionada podrá ser una que esté integrada en alguno de los paquetes de R o que el estudiante haya obtenido desde otra fuente (Internet, etc.). **Destacar que la BD AdultUCI y Zoo no pueden ser utilizadas para este trabajo.**

En este análisis se deberá hacer uso de las herramientas que se han aplicado sobre la BD AdultUCI en el guion de prácticas “Laboratorio-ReglasAsociacion1.pdf” para guiar la búsqueda de información útil hacia los detalles del problema que consideréis que pueden ser interesantes (algo similar a lo que hicisteis sobre la BD Zoo). Destacar que NO es aplicarlas en el mismo orden que en el guion del trabajo en laboratorio, sino hacer uso de ellas como creáis adecuado para guiar la búsqueda hacia la información que consideréis interesante para vuestro problema. Además, este análisis deberá también incluir un análisis de ítems negados del problema y un análisis por grupos de las reglas.

Los ítems negativos se incluirán solo en 1 o 2 variables de problema para evitar una explosión de combinaciones al aplicar el método Apriori. Para añadir los ítems negativos de una variable remplazaremos esta variable por tantas variables nuevas como posibles valores podía tomar la variable original. En las variables nuevas los ejemplos tomaran como valor True si el ejemplo contiene el valor correspondiente y False en caso contrario. Por ejemplo, supongamos que tenemos la variable color que puede tomar los valores rojo, amarillo y verde. En este caso la variable color se remplazará por las variables colorRojo, colorAmarillo y colorVerde. Cada ejemplo tomará el valor True o False para cada una de estas variables en función del valor original del ejemplo. Por ejemplo, si un ejemplo de la BD tiene color rojo, las nuevas variables tomaran para ese ejemplo los valores colorRojo = True, colorAmarillo=False, colorVerde=False.

Considerar que al introducir los ítems negativos habrá que realizar un post procesamiento de las reglas más fino para limpiar reglas que sean iguales o que contengan ítems redundantes. Por ejemplo, la regla “Si el colorRojo = True y ColorVerde=False Entonces Edad= Joven” es igual que la regla “Si el colorRojo = True Entonces Edad= Joven”.

Recordar que el objetivo del análisis no es obtener un conjunto de reglas que tengan valores altos para las medidas de interés. El objetivo es obtener reglas que proporcionen información útil y que no nos indiquen cosas que son obvias o irrelevantes. Por ejemplo, si analizamos la BD Zoo y obtenemos esta regla “Si tiene plumas -> Vuela” con una confianza 0.95, en principio no nos dice algo útil, pero si nos fijamos en su medida podemos ver que hay algunas especies que

aunque tengan plumas no vuelan y nos centraríamos en analizar esa información que es de mayor utilidad.

De este análisis al menos se deberá extraer 5-10 reglas que sean interesantes y que, según vuestro juicio como expertos de la BD, son también útiles.

Adicionalmente, como tarea opcional, el estudiante que lo desee puede utilizar también el paquete de R RKEEL (disponible en el cran) que actualmente ha sido extendido para poder ejecutar en R cualquier método de reglas de asociación que hay incluido en la herramienta software KEEL. Entre ellos, se os recomienda probar el método MOPNAR, el cual permite generar RA cuantitativas con ítems positivos y negativos a partir de BD en formato data.frame y sin necesidad de dividir en intervalos las variables numéricas (aprende el intervalo para cada ítems de una variable numérica integrado en una variable). Una vez obtenidas las reglas, se tendrían que analizar para ver si se ha generado alguna regla que aporte información útil de la BD seleccionada.

Entregar el fichero.r con el script de R generado y un fichero PDF con el análisis realizado. Se valorará positivamente que el pdf no se extienda más de 15-20 páginas.