Joan Puigdomenech
Joel Romia

Análisis de Groceries Dataset

Utilizando herramientas como R y Apriori, un algoritmo de minería de datos utilizado para la generación de reglas de asociación en un conjunto de datos, podemos saber que ítems aparecen frecuentemente en las transacciones y utilizar esta información para incrementar los beneficios. El algoritmo utiliza un enfoque basado en la frecuencia para identificar los elementos más comunes en los datos, y luego utiliza estos elementos comunes para generar reglas de asociación. El enfoque de frecuencia se basa en la idea de que los elementos que aparecen juntos con más frecuencia en un conjunto de datos tienen más probabilidades de estar relacionados entre sí. Apriori es una herramienta útil para identificar patrones y relaciones en grandes conjuntos de datos transaccionales.

El conjunto de datos Groceries contiene 1 mes (30 días) de datos reales de transacciones en el punto de venta de un típico supermercado local. El conjunto de datos contiene 9835 transacciones y los artículos están agregados en 169 categorías.

Primeramente, cargamos el dataset en R y visualizamos los datos en crudo. Observamos que tenemos 9835 filas (transacciones) y 169 columnas (Items). Podemos observar que los items con mas frecuencia dentro del dataset son “Whole Milk”, “Other vegetables”, “rolls/buns”, “soda” y “yogurt”. Sin preprocesar los datos aplicamos el algoritmo A priori y jugamos con el umbral de soporte y la confianza para determinar 10 reglas, es decir, 10 estancias que describen la relación entre dos o mas items en el dataset. Es muy importante afinar en el umbral de soporte mínimo para identificar los elementos mas frecuentes, ya que se refiere a la proporción de transacciones en las que aparece un conjunto de elementos dado, también es muy importante la confianza de la regla, ya que se refiere a la proporción de transacciones que contienen tanto A como B, respecto del número de transacciones que contienen.

Dicho esto modificamos los parametros en Apriori de support y confiance en 0.009 y 0.55 para obtener 10 reglas, las cuales son las siguientes:

```
     lhs                                   rhs                support     confidence coverage   lift
[1]  {citrus fruit, root vegetables}    => {other vegetables} 0.010371124 0.5862069  0.01769192 3.029608
[2]  {tropical fruit, root vegetables}  => {other vegetables} 0.012302999 0.5845411  0.02104728 3.020999
[3]  {butter, yogurt}                   => {whole milk}       0.009354347 0.6388889  0.01464159 2.500387
```

Podemos observar ordenando por peso decreciente el lift, una medida de interés que mide la proporción entre la confianza de la regla y la proporción de veces que aparece B en el conjunto de transacciones, que obtenemos un antecedente como “citrus fruit, root vegetables” -> “other vegetables”, eso quiere decir que la gente que compra como antecedente fruta cítrica y verduras de raíz como la zanahoria, cebolla… suele comprar otro tipo de verduras, quizá no es una de las reglas más representativas pero ya nos indica una tendencia de compra. Una regla interesante podría ser la [3] “Butter, yogurt” -> “Whole milk”, es decir, gente que compra mantequilla y yogurt, suele comprar leche entera, con esta información se podría pensar un plan de venta, subir o bajar algún precio de los productos o pensar en packs de compra para conseguir un beneficio.

Visto que el ítem con más frecuencia del dataset es whole milk, hemos decidido preprocesar los datos eliminando este item para así poder estudiar otros productos y sus relaciones.

Eliminando este ítem observamos que los productos actuales con más frecuencia son “soda”, “rolls/buns”, “otrher vegetables”, “yogurt” y “bottled wáter”. Aplicando el algoritmo Apriori y ajustando las reglas con parámetros de support a 0.0022 y confidence a 0.60, obtenemos las siguientes 7 reglas:

```
    lhs                                     rhs                support     confidence coverage    lift
[1] {liquor, red/blush wine}             => {bottled beer}     0.002594919 0.9047619  0.002868069 11.209250
[2] {root vegetables, yogurt, rolls/buns} => {other vegetables} 0.002321770 0.6800000  0.003414368  4.266461
[3] {sausage, dessert}                   => {other vegetables} 0.002731494 0.6666667  0.004097241  4.182805
[4] {root vegetables, chocolate}         => {other vegetables} 0.002458345 0.6428571  0.003824092  4.033419
[5] {root vegetables, dessert}           => {other vegetables} 0.002868069 0.6363636  0.004506965  3.992677
[6] {tropical fruit, onions}             => {other vegetables} 0.002321770 0.6071429  0.003824092  3.809340
[7] {hard cheese, rolls/buns}            => {other vegetables} 0.002731494 0.6060606  0.004506965  3.802550
```

Podemos ver reglas de asociación interesantes, la primera que es la que más peso en lift tiene, nos dice que la gente que compra como antecedente licor y vino también suele comprar cerveza. Quizá una regla interesante es la [7] ya que normalmente en transacciones donde la gente compra queso fuerte y bollos, suelen comprar verduras, probablemente para hacer una receta de cocina en concreto como sausage rolls. Visto esto también podríamos pensar un plan de venta para maximizar los beneficios de la tienda.

Otra alternativa de preprocesado de datos ha sido eliminar tanto “Whole milk” como “Other Vegetables” ya que hemos visto que son los items más frecuentes tanto en LHS como RHS, es decir, como antecedentes y consecuentes. Con esto lo que se busca es obtener alguna regla interesante poco común con bastante peso en lift. Modificando los parámetros de support y confidence a valores como 0.001 y 0.57 obtenemos 10 reglas. El resultado es el siguiente:

```
     lhs                                                    rhs              support     confidence
[1]  {bottled beer, red/blush wine}                      => {liquor}         0.003086921 0.6333333
[2]  {liquor, red/blush wine}                            => {bottled beer}   0.003086921 0.9047619
[3]  {soda, liquor}                                      => {bottled beer}   0.001787165 0.6875000
[4]  {sausage, rolls/buns, canned beer}                  => {shopping bags}  0.001624695 0.6250000
[5]  {frankfurter, mustard}                              => {rolls/buns}     0.001624695 0.7692308
[6]  {rolls/buns, bottled water, fruit/vegetable juice}  => {soda}           0.001624695 0.7142857
[7]  {coffee, fruit/vegetable juice}                     => {soda}           0.002274574 0.6666667
[8]  {pork, canned beer}                                 => {soda}           0.001787165 0.6470588
[9]  {yogurt, rolls/buns, bottled water}                 => {soda}           0.002761982 0.6071429
[10] {pastry, salty snack}                               => {soda}           0.002274574 0.5833333
     coverage    lift      count
[1]  0.004874086 42.371377 19
[2]  0.003411860 10.983845 19
[3]  0.002599513  8.346277 11
[4]  0.002599513  6.690217 10
[5]  0.002112104  4.692384 10
[6]  0.002274574  3.870096 10
[7]  0.003411860  3.612089 14
[8]  0.002761982  3.505851 11
[9]  0.004549147  3.289581 17
[10] 0.003899269  3.160578 14
```

Podemos observar reglas de asociación interesantes como la [5] donde la gente que compra Frankfurt y mostaza compra también bollos, lo cual es obvio y de aquí se podría hacer un pack para aumentar los beneficios de compra o básicamente situarlos juntos en la propia tienda. También podemos observar que la “soda” está bastante presente en los RHS dado que es un ítem muy frecuente, podemos ver en la regla [7] la gente que compra café, fruta/ zumo de verdura también compra soda.

Otra alternativa de preprocesado para seguir estudiando los ítems y reglas ha sido focalizarnos en un RHS en concreto, hemos puesto el foco en “Newspapers”, se han modificado los parámetros support y confidence a 0.001 y 0.57 para poder encontrar una regla con este ítem como consecuente. El resultado es el siguiente:

```
    lhs                                  rhs           support     confidence coverage    lift
[1] {canned beer, liquor (appetizer)} => {newspapers} 0.001137287 0.5833333  0.001949634 8.800041
    count
```

Podemos observar que con un peso de 8.8 en Lift, la gente que compra cerveza de lata y licor de aperitivo también compra el periódico. Otra regla interesante para posicionar de manera inteligente los productos dentro de la propia tienda y así maximizar los beneficios de compra.

Por último, se ha decido agrupar algunos ítems como “Whole milk”, “UHT-milk” y “Condesed milk” para ver qué tipo de reglas obtenemos. Aplicando el algoritmo y ajustando los parámetros support y confidence a 0.11 y 0.60 obtenemos 10 reglas, entre las cuales están las siguientes:

```
     lhs                     rhs          support   confidence coverage  lift      count
[1]  {whipped/sour cream} => {whole milk} 0.1104145 0.9135447  0.1208638 1.0436875  317
[2]  {root vegetables}    => {whole milk} 0.1675374 0.9127135  0.1835597 1.0427379  481
[3]  {other vegetables}   => {whole milk} 0.2563567 0.8997555  0.2849181 1.0279340  736
[4]  {rolls/buns}         => {whole milk} 0.1940091 0.8912000  0.2176942 1.0181596  557
[5]  {pastry}             => {whole milk} 0.1138976 0.8910082  0.1278300 1.0179405  327
[6]  {tropical fruit}     => {whole milk} 0.1448972 0.8888889  0.1630094 1.0155193  416
```

Obtenemos reglas con un lift próximo a uno, lo cual nos indica que la ocurrencia de los elementos en la regla de asociación es independiente. La regla [1] nos sugiere que la gente que compra crema también compra leche entera. Por el valor bajo de Lift nos sugiere que la probabilidad de que ocurra el antecedente no está relacionada con la probabilidad de que ocurra el consecuente.

También se ha limitado el tamaño de LHS a dos, pero el resultado es parecido al anterior, un lift bajo. Por lo tanto, podemos concluir que en términos prácticos esta regla de asociación no tiene ningún valor predictivo y no es útil para la toma de decisiones en términos de recomendaciones de productos.

```
    lhs                                rhs          support    confidence coverage   lift
[1] {root vegetables, other vegetables} => {whole milk} 0.07941484 0.9382716  0.08463950 1.071937
[2] {other vegetables, yogurt}          => {whole milk} 0.07628004 0.9201681  0.08289794 1.051254
```