Trabajo Práctico 1

Nicolás Muschitiello
Roberto Velez
Jairo Jiménez

June 22, 2015

1 Introducción

2 Resultados técnicos esperados

2.1 Preprocesamiento

En este apartado, describimos cómo preprocesamos las bases previo a introducirlas en el algoritmo de generación de reglas de asociación.

Comenzamos aplanando la base, usando MS Access.

Removimos de la base aplanada todos aquellos renglones que no tenían transacciones asociadas (entre ellos, 1 Categoría y 16 SubCategorías).

Eliminamos las transacciones con cantidades negativas, ya que representan devoluciones. En el anexo se enumeran los Venta_ID eliminados, que totalizan XXX.

Detectamos la existencia productos con precio 0. Se decidió no hacer nada al respecto, aunque distorsionan las mediciones del gasto de los clientes y la ponderación de reglas por su impacto en las ventas.

Separamos el campo DescAdic, utilizando (descripción del algoritmo usado por Robert) Determinamos que casi toda la información contenida en este campo se encuentra codificada dentro de Prod_ID, por lo que decidimos no trabajar a la hora de generar reglas con DescAdic.

Como medida de volumen, decidimos trabajar con el campo Cantidad UM1.

Clasificamos los clientes siguiendo tres ejes: la cantidad de compras realizadas, el gasto incurrido en las mismas, y la categoría de los productos que compran. Para las dos primeras clasificaciones, utilizamos tres categorías para

cada una, a saber:

Cantidad de transacciones. Muy Frecuente: ¿ = 14 transacciones (una compra por mes en promedio o más); Frecuente entre 7 y 13 (una compra cada dos meses o más); Poco Frecuente ¡ = 7 (menos de una compra cada dos meses).

Gasto. Gasto Alto ¿ = $ 500.000. Gasto Medio entre $ 100.000 y $ 499.999. Gasto Bajo ¡ = $ 100.000.

Adicionalmente, determinamos su rubro (o rubros), a partir de los productos que compran, agrupados por categoría de acuerdo al siguiente esquema:

Los clientes que compraron un 50% o más de productos (COUNT DISTINCT Prod_ID & Venta_ID) de una misma categoría, se clasifican con esa categoría.

A los clientes que no entran en la clasificación anterior, si entre las dos categorías mayoritarias suman más del 80%, se utilizan ambas para su clasificación. Si no, se los clasifica como POLIRUBRO.

2.2 Software utilizado

En la elaboración del presente trabajo, se usaron gran variedad de herramientas de análisis de datos.El preprocesamiento de los datos se hizo conjuntamente con las herramientas Microsoft Excel, Microsoft Access y QlikView.
Para la la generación y evaluación de las reglas presentadas en este trabajo se exporaron las herramientas *Weka* [Witten and Frank, 2005] y *R* [R Core Team, 2013]. Dada la flexibilidad de programación del software estadístico *R*, se optó por este último. En dicho software, se implementaron los códigos necesarios para la generación de los conjuntos de datos en el formato requerido por el algoritmo, el cálculo de las medidas de interés adicionales y la poda de las reglas.

2.3 Justificación de la elección de los parámetros del algoritmo

La determinación del minsup se realizó a partir del análisis exploratorio de los datos (Mostrar una tabla de frecuencia o gráfico que vincule la cantidad de ítems por nivel de SUP).

La de la confianza, a partir de iteraciones sucesivas del algoritmo y observar cuántas reglas interesantes generaba.

Adicionalmente, al no contar con conocimiento experto sobre la base con la que trabajamos, decidimos emplear dos medidas más para seleccionar reglas interesantes: Kulc e IR. Las consideramos apropiadas dado que cumplen con la null-invariance property, que consiste en no verse afectadas ante la presencia de

transacciones nulas. Una transacción es nula cuando no contiene ninguno de los ítems del ítemsets en análisis.

2.4 Criterio para la selección de los resultados no técnicos

Para el análisis de las reglas interesantes, se decidió utilizar medidas adicionales a las medidas clásicas [Tan et al., 2005]. Con la ayuda de las medidas adicionales, se determinaron las reglas más interesantes para kas categorías, subcategorías, etc.

2.4.1 Medidas adicionales

Las medidas que fueron elegidas, tienen la propiedad de ser *null-invariantes*, es decir que no se ven afectadas por el efecto de la falta de la categoría en el conjunto de datos. Las medidas utilizadas son la medida coseno, la medida de Kulczynsky y la razón de desbalanceo [Hall et al., 2009]. Estas medidas se presentan a continuación

Medida coseno:

$$cosine(A, B) = \sqrt{P(A|B)P(B|A)}$$

Medida de Kulczynsky:

$$Kulc(A, B) = \frac{1}{2}\left(P(A|B)P(B|A)\right)$$

Razón de desbalanceo:

$$IR = \frac{|sup(A) - sup(B)|}{sup(A) + sup(B) - sup(A \cup B)}$$

3 Resultados no técnicos esperados

3.1 Características más habituales de las ventas de la empresa

3.2 Reglas generadas a partir de las variables demográficas

3.3 Reglas a nivel de monto y cantidad de ventas de la empresa

3.4 Reglas generadas de un año a otro

4 Bonus

References

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

[R Core Team, 2013] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Tan et al., 2005] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.