Trabajo Práctico 1

Nicolás Muschitiello
Roberto Vélez
Jairo Jiménez

June 26, 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 que totalizan 363
- 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 el siguiente procedimiento:
  1. Realización de un query para sacar la categoría y subcategoría por producto con las descripciones generales y adicionales, estas últimas son las que vamos a separar en una tabla nueva que contendrá un registro por cada atributo con su valor asociado con la subcategoría

```
SELECT TP_Categoria.Cat_ID, TP_Categoria.
   Cat_Desc, TP_Sub_Categoria.SubCat_ID,
   TP_Sub_Categoria.SubCat_Desc,
   TP_Productos.DescGen, TP_Productos.
   DescAdic
```

```
FROM TP_Sub_Categoria INNER JOIN (
   TP_Categoria INNER JOIN TP_Productos ON
   TP_Categoria.Cat_ID = TP_Productos.Cat_ID
   ) ON TP_Sub_Categoria.SubCat_ID =
   TP_Productos.SubCat_ID;
```

2. Realización de un script para separar la columna DescAdic, este script contiene la separación de la subcategoría primeramente para después poder separar los atributos que están identificados con la notación:

   Nombre del Atributo: Valor del Atributo

3. Realización de un script (Ver Anexos) para generar un csv con los nuevos registros para la tabla TP_MarcaAtributos en R.

4. Creación de la tabla en Access con las siguientes columnas:
   - marca_ID
   - marca_Nombre
   - marca_AtributoNombre
   - marca_AtributoValor
   - SubCat_ID

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: $\geq$ 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 $\geq$ \$ 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.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 otras medidas, las cuales se presentan en el apartado 2.4.1

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 éstas, se determinaron las reglas más interesantes para las categorías, subcategorías, etc.

2.4.1 Medidas adicionales

Las medidas que fueron elegidas, tienen la propiedad de ser *null-invariantes*, es decir, 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:

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

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

En general las relaciones existentes entre las variables demográficas y los productos, descripción general de los productos, subcategorías y categorías es muy poca. A nivel de producto, las reglas encontradas suelen no ser muy interesantes pues éstas solamente relacionan botellas con botellas o termos con termos, sin embargo, éstas tienen dos particularidades: todas tienen una correlación positiva entre el antecedente y el consecuente y todas están relacionadas con la localidad y la provincia de Buenos Aires.

En cuanto a los demás grupos (Descripción general, subcategoría y categoría), las correlaciones encontradas son negativas mostrando una "repelencia" entre los ítems y las ciudades en las cuales se encuentran, como por ejemplo el caso del jarro térmico en la provincia de Buenos Aires y el termo. Tabla 1

Table 1: Reglas con información demográfica

Reglas	**Soporte**	**Confianza**	**Lift**	**Coseno**	**Kulczinsky**	**IR**	**Grupo**
BTP4S79-5RC,BTP4S79-75RC,BTP4S79-75SC,Capital Federal-Prov =>BTP4S79-5SC	0,006	1,000	48,438	0,526	0,638	0,723	Producto
BTP4S79-5RC,BTP4S79-75RC,BTP4S79-75SC,C.A.B.A.-Loc =>BTP4S79-5SC	0,005	1,000	48,438	0,504	0,627	0,746	Producto
BTP4S79-5BLC,BTP4S79-75RC,BTP4S79-75SC,Capital Federal-Prov =>BTP4S79-5SC	0,006	0,972	47,093	0,512	0,621	0,718	Producto
LCM1406NB,Buenos Aires-Prov =>LCM1406NR	0,006	0,947	48,501	0,527	0,620	0,680	Producto
REELS PEJERREY,REELS VARIADA,C.A.B.A.-Loc =>CAÑAS TELESC. PEJERREY	0,010	0,693	6,684	0,254	0,393	0,831	Subcategoría
LINEAS FLY,Capital Federal-Prov =>CAÑAS FLY	0,006	0,609	17,363	0,328	0,393	0,638	Subcategoría
LINEAS FLY,C.A.B.A.-Loc =>CAÑAS FLY	0,006	0,625	17,808	0,315	0,392	0,682	Subcategoría
REEL CHARGER,Buenos Aires-Prov =>REEL BELLUS	0,007	0,652	11,313	0,284	0,388	0,760	Desc. General
BINOCULAR TRAVEL II,C.A.B.A.-Loc =>BINOCULAR ORBITAL	0,008	0,615	9,271	0,266	0,365	0,759	Desc. General
JARRO TÉRMICO,Buenos Aires-Prov =>TERMO	0,009	0,628	6,668	0,239	0,359	0,811	Desc. General
EL PALOMAR-Loc =>PESCA	0,008	0,613	1,853	0,120	0,318	0,947	Categoría
TORTUGUITAS-Loc =>CAMPING	0,005	0,618	1,113	0,078	0,314	0,978	Categoría
MONTE GRANDE-Loc =>CAMPING	0,005	0,607	1,093	0,077	0,308	0,978	Categoría

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

3.4 Reglas generadas de un año a otro

Las siguientes reglas presentadas a continuación fueron tomadas con las siguientes condiciones:

- El periodo de comparación fue todo el 2014 con respecto a los meses de marzo a Mayo del 2015

- Las reglas generadas y catalogadas como interesantes reflejan un periodo de comparación bastante diferente entre un año y otro por lo que muchas encontradas en el año 2015 no se cumplían en 2014 y viceversa.

Es muy probable que tomando el mismo periodo de meses del 2015 para el 2014 resulten en mayor cantidad de reglas similares en caso de existir comportamientos de venta en común.

Table 2: Reglas por Año

Reglas	support	confidence	lift	chiSquared	cosine
Gasto Alto =>Muy Frecuente	0.394919	0.989869754	1.588929	658.8846364	0.792148
LCM1406NG,Muy Frecuente =>LCM1406NR	0.008661	1	49.48571	733.6394043	0.654654
MG1970,CASA DE OPTICA =>MG1968	0.008083	1	52.48485	726.6615868	0.651339
MG1968,Gasto Bajo =>MG1970	0.009238	0.888888889	49.66308	784.9937245	0.677334
Muy Frecuente,CASA DE PESCA =>Gasto Alto	0.245958	0.742160279	1.860234	421.7077793	0.676417
BTP4S79-5RC,BTP4S79-75RC,BTP4S79-75SC,Muy Frecuente =>BTP4S79-5SC	0.01097	0.95	30.47037	565.5188873	0.578152
BTP4S79-75BLC,BTP4S79-75RC,Muy Frecuente =>BTP4S79-75SC	0.021363	0.860465116	17.32937	614.3191105	0.608441
BTP4S79-5RC,BTP4S79-75BLC,BTP4S79-75SC,Muy Frecuente =>BTP4S79-5SC	0.010393	0.947368421	30.38596	533.8581048	0.561951
LCM1406NR,CASA DE PESCA =>LCM1406NB	0.006928	0.923076923	47.02262	555.5195607	0.570782
LCM1406NG,Muy Frecuente =>LCM1406NS	0.007506	0.866666667	46.90833	600.2787385	0.593366
BTP4S79-75RC,BTP4S79-75SC,Gasto Medio =>BTP4S79-75BLC	0.017321	0.9375	17.27394	495.5050869	0.546994
BTP4S79-5RC,BTP4S79-75SC,Muy Frecuente =>BTP4S79-5SC	0.01097	0.904761905	29.0194	537.0771176	0.564218
LCM1406NG,Muy Frecuente =>LCM1406NB	0.007506	0.866666667	44.14902	564.0946071	0.57565
BTP4S79-5SC,BTP4S79-75BLC,Muy Frecuente =>BTP4S79-75SC	0.011547	1	20.13953	387.2625516	0.482243
BTP4S79-75SC,CASA DE CAMPING =>BTP4S79-75BLC	0.017321	0.882352941	16.25782	463.3276079	0.530662
BTP4S79-5RC,BTP4S79-75RC,CASA DE CAMPING =>BTP4S79-5BLC	0.008661	0.833333333	32.80303	479.5421825	0.533002
TN4X30-1,Gasto Bajo =>CASA DE OPTICA	0.014434	0.961538462	9.972363	226.7555922	0.379398
Poco Frecuente,CASA POLIRUBRO =>Gasto Bajo	0.011547	1	4.463918	70.08767704	0.227038
TN4X30-3,Gasto Bajo =>CASA DE OPTICA	0.01097	0.904761905	9.383519	159.4307079	0.320838
HPR50600,Gasto Bajo,CASA DE OPTICA =>HPR50360	0.006928	0.705882353	30.56471	354.7639184	0.460179
INCA200,Gasto Medio,Muy Frecuente =>CASA DE CAMPING	0.008083	0.875	4.735938	51.08166831	0.195656
MG1968,TN4X30-3 =>CASA DE OPTICA	0.006351	0.846153846	8.775679	84.50729205	0.236082
TERRA =>Muy Frecuente	0.016166	0.848484848	1.361979	7.283571541	0.148385
MG1968,TN4X30-3 =>Gasto Bajo	0.005774	0.769230769	3.433783	22.39811261	0.140803
PATAGONIAPRO =>Gasto Alto	0.010393	0.75	1.879884	12.50689906	0.139774
Gasto Medio,CASA POLIRUBRO =>Muy Frecuente	0.0306	0.716216216	1.149663	2.86109861	0.187564
TERRA,Muy Frecuente =>Gasto Alto	0.011547	0.714285714	1.790366	11.80103417	0.143784

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.