INDUCCIÓN DE REGLAS DE ASOCIACIÓN DE MINERÍA DE DATOS EN BASE DE DATOS DE ENTIDAD RETAIL

INDUCTION OF ASSOCIATION RULES OF DATA MINING IN DATABASE OF RETAIL ENTITY

Josué Galarreta Vásquez[1]

Fecha de recepción: 17 de mayo 2016
Fecha de aceptación: 20 de setiembre 2016

Resumen

El objetivo de este trabajo de investigación es descubrir reglas de asociación relevantes entre las categorías de productos en una base de datos de una entidad retail de electrodomésticos, a través del uso del algoritmo de inducción FP-Growth de la minería de datos. Cada regla de asociación determina un comportamiento de compra de los clientes, y permite conocer cuál es la probabilidad de que un determinado cliente compre un producto categorizado como Y dado que anteriormente compró un producto categorizado como X. El trabajo se aplicó en tres de las tiendas más representativas de la región norte del país de la entidad retail: Trujillo, Chiclayo y Piura. Se construyó un modelo en el software de minería de datos RapidMiner Studio conectado a una base de datos en Microsoft SQL Server que contiene la información histórica de diez años de la entidad retail. El modelo fue ejecutado una vez para cada tienda. Como resultado se encontraron reglas de asociación relevantes en Chiclayo y Piura que relacionan computadoras de escritorio e impresoras, reproductores y televisores. Los resultados también demostraron patrones de comportamiento distintos de clientes en cada tienda.

Palabras claves: *minería de datos, reglas de asociación, FP-Growth*

Abstract

The aim of this research work is to discover relevant association rules between product categories in a database from a retail entity of appliances, through the use of induction algorithm FP-Growth of data mining. Each association rule determines a buying behavior of customers and allows us to know what is the probability that a given customer buys a product categorized as Y given he bought a product categorized as X. Work was applied in three stores more representative of the northern region of the retail entity: Trujillo, Chiclayo and Piura. A model was built into the software of data mining RapidMiner Studio connected to a database in Microsoft SQL Server that contains historical information of ten years of the retail entity. The model was run once for each store. As a result, relevant association rules in Chiclayo and Piura linking desktop computers and printers, and players and TVs were found. The results also showed different behavior patterns of customers in each store.

Keywords: *data mining, association rules, FP-Growth*

[1] *Escuela Profesional de Ingeniería de Sistemas. Facultad de Ingeniería Civil, de Sistemas y Arquitectura. Ing. Universidad Nacional "Pedro Ruiz Gallo". Lambayeque. Lambayeque. Perú. josue.galarreta.v@gmail.com.*

1. Introducción

La entidad retail es una compañía que comercializa electrodomésticos y que ha logrado posicionarse como una de las compañías más exitosas de su sector en el país en los últimos años. Desde que informatizó sus procesos en el año 2004 y debido a la aplicación de una agresiva estrategia de crecimiento, la entidad retail ha experimentado un crecimiento vertiginoso tanto de ventas como de cantidad de información. Durante todos estos años la información ha sido gestionada a través de gráficos de líneas, gráficos de barra, gráficos circulares y tablas. Ésta forma de gestionar la información es importante, pero es limitada y no permite explotar el conocimiento que subyace en la base de datos que al año 2014 ya tenía más de diez millones de registros de ventas. En esta etapa de la entidad retail es importante aplicar otras medidas como lo es la minería de datos para encontrar nuevos conocimientos que guíen la gestión comercial.

Debido a que la minería de datos tiene muchos tipos de procedimientos, optamos por utilizar la inducción de reglas de asociación para encontrar relaciones entre distintas categorías de productos. El objetivo es encontrar reglas de asociación relevantes que nos permitan conocer que categorías de productos incentivan la compra de otras categorías de productos y en qué medidas. Esto para cada tienda de Trujillo, Chiclayo y Piura.

2. Materiales y Métodos

Para aplicar el procedimiento de inducción de reglas de asociación de la minería de datos se construyó un modelo en el software RapidMiner Studio el cual se conectó con el gestor de base de datos Microsoft SQL Server, el cual contiene la base de datos (esquema en estrella) de la entidad retail con la información del año 2004 al 2014.

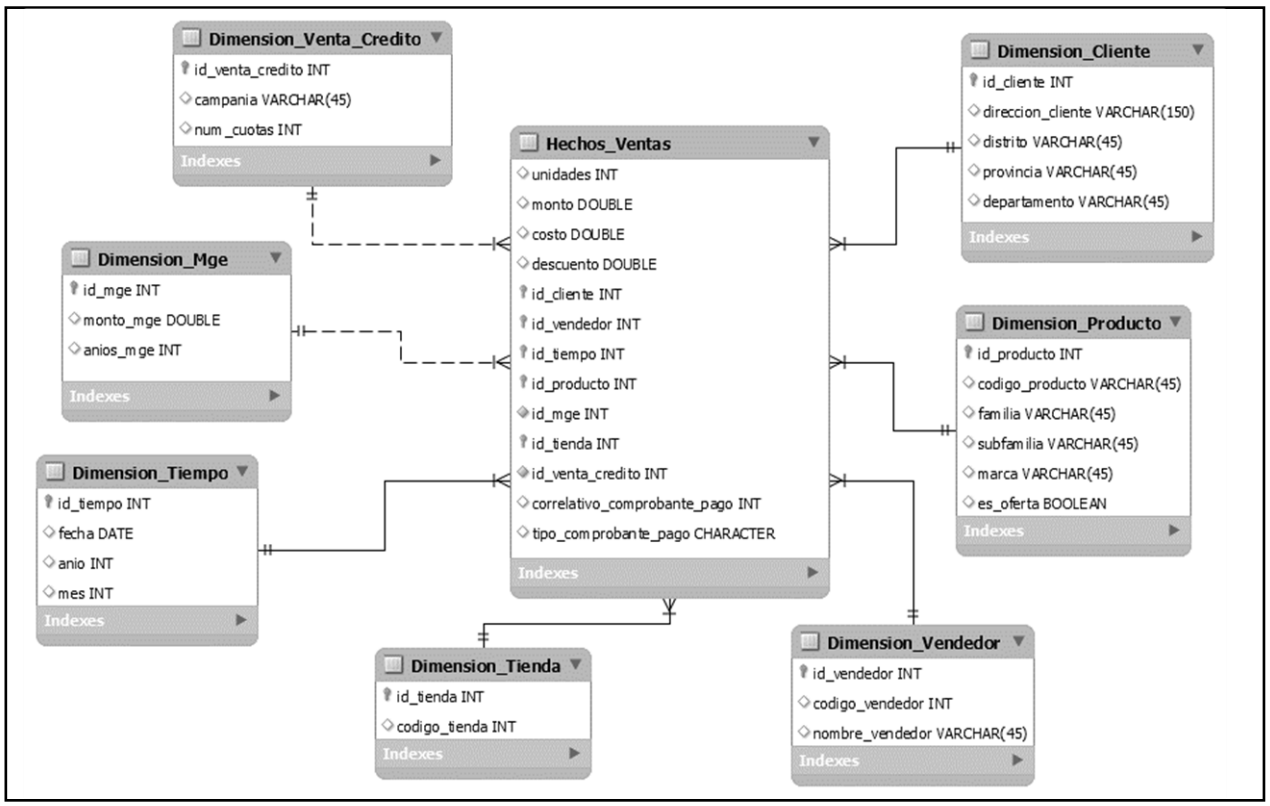

Figura 1. *Esquema estrella de base de datos de la entidad retail*

Fuente: *Elaboración propia*

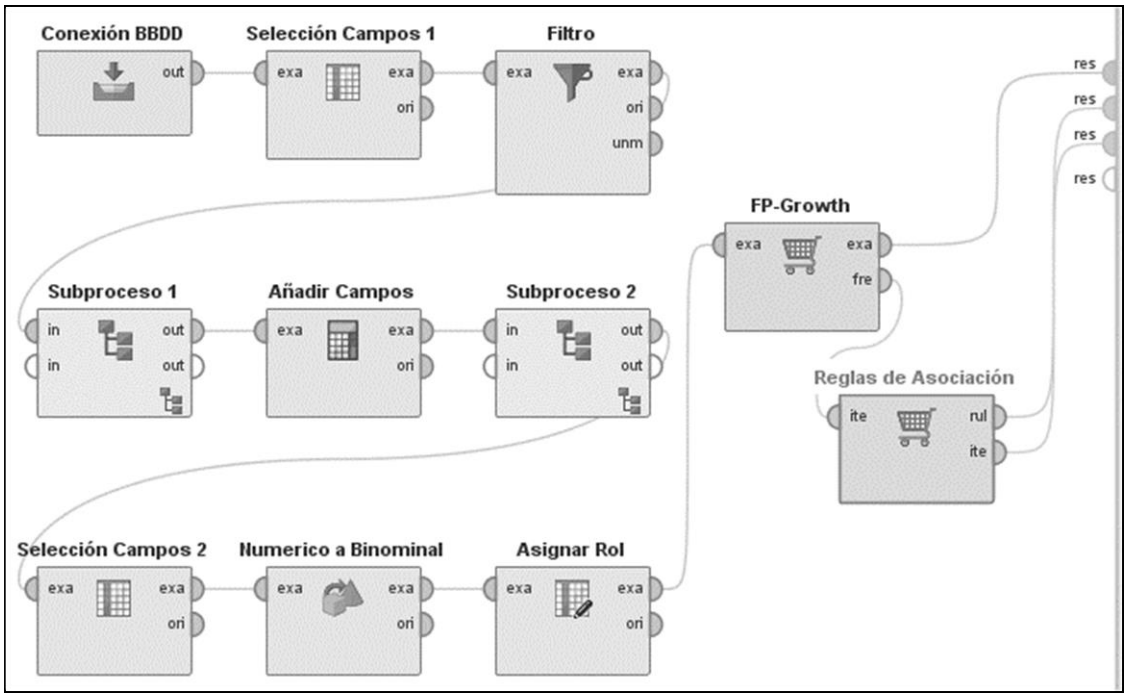

Figura 2. *Diseño del modelo de minería de datos en RapidMiner Studio*

Fuente: *Elaboración propia*

Se dispuso de todos los registros de la base de datos desde el año 2004 al año 2014 y se hizo el análisis de la población total de clientes para cada tienda.

Tabla 1.

Población total de clientes analizada por tienda

Tienda	Número de Clientes
Trujillo	144 520
Chiclayo	176 304
Piura	114 793

Fuente: *Elaboración propia*

Dentro del esquema estrella de la base de datos (Figura 1) la dimensión productos clasifica a los productos en *familia*, *subfamilia*, *marca* y *es_oferta*. En esta investigación se enfocó en encontrar las relaciones entre subfamilias de productos. El campo *es_oferta* se consideró para filtrar los productos que se entregaron como promocionales, ya que esto pudo distorsionar el resultado.

Tabla 2.

Clasificación de productos en familias y sub-familias

Familias	Sub-familias			
Cómputo	1 Desktop	3 Software	5 Lapto	7. Proyectores
	2 Impresoras	4 All in one	6 Tablets	
Digital	1 Consola	3 Audífonos	5 Juegos	7 Filmadoras
	2 Cámaras	4 MP3	6 Autoradio	

Línea blanca	1 Refrigeradoras 2 Campanas	3 Cocinas 4 Congeladoras	5 Hornos 6 Aire acondicionado	7 Secadoras 8 Lavadoras
Audio	1 Radio CD	2 Cine en casa	3 Equipos	
Video	1 Reproductores	2 TV		
Pads	1 Licuadora 2 Extractor 3 Tostadora 4 Batidora 5 Cuidado personal 6 Dispensador de agua 7 Grill	8 Thermas 9 Aspiradoras 10 Plancha 11 Horno eléctrico 12 Exprimidor 13 Waflera 14 Ventilación	15 Cafetera 16 Lustradoras 17 Parrilla eléctrica 18 Menaje 19 Combos 20 Calefacción 21 Sanguchera	22 Termoradiadores 23 Hervidor 24 Thermo 25 Deshumedecedores 26 Licuo-extractor 27 Olla 28 Freidoras

Fuente: *Elaboración propia*

Durante la ejecución del modelo se llevó a cabo una serie de transformaciones de la base de datos de entrada para generar la matriz adecuada que fue procesada por el algoritmo FP-Growth. Esta matriz cruza la información de todos los clientes que han comprado en la tienda y todas las subfamilias de productos. Si un cliente compró por lo menos una vez un producto de una subfamilia, entonces se registra un *True*, caso contrario se registra un *False*.

Tabla 3.

Formato de matriz lista para ser procesada por algoritmo FP-Growth

	Subfamilia_1	**Subfamilia_2**	**Subfamilia_3**	**Subfamilia_4**	**…**	**Subfamilia_n**
Cliente_1	True or False	True or False	True or False	True or False		True or False
Cliente_2	True or False	True or False	True or False	True or False		True or False
Cliente_3	True or False	True or False	True or False	True or False		True or False
Cliente_4	True or False	True or False	True or False	True or False		True or False
Cliente_5	True or False	True or False	True or False	True or False		True or False
Cliente_6	True or False	True or False	True or False	True or False		True or False
Cliente_7	True or False	True or False	True or False	True or False		True or False
…	…	…	…	…		…
Cliente_n	True or False	True or False	True or False	True or False		True or False

Fuente: *Elaboración propia*

El resultado del modelo es una lista de reglas de asociación, donde cada regla es una relación de causa-efecto, donde si un elemento *A* existe, otro elemento *B* también debe existir (Hornick, 2007).

Cada regla de asociación fue evaluada a partir de sus medidas de confianza, que es la medida de interés de una regla asociación más usada; y que permite medir la fuerza de una relación entre dos ítems (Giudici-Figini, 2010).

Del conjunto de algoritmos que abarca las reglas de asociación se escogió el algoritmo FP-Growth por ser más eficiente que su antecesor Apriori (Pinho, 2010).

3. Resultados

Tabla 4.

Resultados del proceso de minería de datos para Trujillo

Premisa	Conclusión	Confianza
Reproductores	TV	0.44
Equipos	TV	0.29
TV	Reproductores	0.29
Plancha	TV	0.26
Cocinas	TV	0.24
Licuadora	TV	0.23
Refrigeradoras	TV	0.22
Equipos	Reproductores	0.21

Fuente: *RapidMiner Studio*

Tabla 5.

Resultados del proceso de minería de datos para Chiclayo

Premisa	Conclusión	Confianza
Desktop	Impresoras	0.59
Impresoras	Desktop	0.51
Reproductores	TV	0.51
Equipos	TV	0.33
TV	Reproductores	0.31
Licuadora	TV	0.30
Plancha	TV	0.28
Cocinas	TV	0.27
Refrigeradoras	TV	0.25
Equipos	Reproductores	0.23
Cocinas	Refrigeradoras	0.23

Fuente: *RapidMiner Studio*

Tabla 6.

Resultados del proceso de minería de datos para Piura

Premisa	Conclusión	Confianza
Desktop	Impredoras	0.58
Impresoras	Desktop	0.57
Reproductores	TV	0.50
Equipos	TV	0.27
TV	Reproductores	0.27
Plancha	TV	0.26
Licuadora	TV	0.25

Cocinas	TV	0.23
Licuadora	Refrigeradoras	0.23
Cocinas	Refrigeradoras	0.22
Plancha	Licuadora	0.20
Refrigeradoras	TV	0.20
Plancha	Refrigeradoras	0.20
Licuadora	Plancha	0.20

Fuente: *RapidMiner Studio*

4. Discusión y Conclusiones

Los datos muestran que existen reglas de asociación con medidas de confianza relativamente altas (superiores a 0.5) en las tiendas de Chiclayo y Piura, lo que indica que en estas tiendas existen reglas de asociación relevantes y que se pueden considerar como un patrón de comportamiento válido. Por otro lado, en la tienda de Trujillo se han encontrado reglas de asociación con medidas de confianza relativamente bajas (inferiores a 0.5).

Con respecto a las reglas de asociación relevantes se observa que existen productos de bajo precio de venta que incentivan la compra de productos con un precio de venta mayor. Este es el caso de las *IMPRESORAS* que incentivan las compras de *DESKTOP*, y de los *REPRODUCTORES* que incentivan las compras de *TV*. En el caso de Chiclayo tenemos que *P* [*IMPRESORAS* | *DESKTOP*] = 0.59 y *P* [*TV* | *REPRODUCTORES*] = 0.51. En el caso de Piura tenemos que *P* [*IMPRESORAS* | *DESKTOP*] = 0.58 y *P* [*TV* | *REPRODUCTORES*] = 0.50. Este conocimiento puede ser utilizado por el área de marketing para incentivar las ventas de productos de mayor precio.

Los resultados también muestran que tanto Chiclayo como Piura muestran patrones de consumo semejante mientras que Trujillo difiere totalmente.

Existen también otras reglas de asociación con medidas de confianza relativamente bajas (inferiores a 0.5) que relacionan categorías de productos que no tienen ningún tipo de relación aparente. Por ejemplo, en el caso de Chiclayo tenemos que *P* [*TV* | *LICUADORA*] = 0.30. Estas relaciones no deben ser descartadas y deberían considerarse como puntos de partidas de futuras investigaciones.

Referencias

Fayyad, U. *et al* (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17(3).*

Giudici, P. y Figini, S. (2009). Applied Data Mining for Business and Industry. Wiley.

Han, J. y Kamber, M. (2006). Data Mining: Concepts and Techniques. Morgan Kaufmann.

Han, J. *et al* (2004). Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Mining and Knowledge Discovery*, 53-87.

Hornick, M. et al. (2007). Java Data Mining: Strategy, Standard, and Practice: A Practical Guide for Architecture, Design, and Implementation. Morgan Kaufmann Publishers Inc.

Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons.

Pinho, L. (2010). *Métodos de Clasificación Basados en Asociación Aplicados a Sistemas de Recomendación* (Tesis de Doctorado). Universidad de Salamanca, España.

Weiss, G. y Davison, B. (2010). Data Mining, in Handbook of Technology Management. John Wiley & Sons.