Curso

3

MACHINE LEARNING & BIG DATA

Unsupervised Learning: Reglas de Asociación | Análisis de Canasta

Algoritmos No-Supervisados

JOSÉ NELSON ZEPEDA DOÑO

Cluster de Estudio: Advanced Analytics

Este material es el resumen de muchos autores que por medio de sus libros y documentos nos ofrecen fuentes riquísimas de conocimiento sobre los temas de Big Data y Machine Learning.
Algunas citas, figuras y tablas pueden ser encontradas de forma textual tal como lo indica el autor en su material original.

Nelson Zepeda
MIP • V 1.0
San Salvador El Salvador
Phone 503 79074137 • @nelsonzepeda733

Tabla de Contenido

Capítulo 1

Análisis de Canasta: Conceptos

El Market Basket Analysis es un análisis matemático que ayuda a encontrar patrones en la información de los tickets de venta de un conjunto de tiendas, durante un periodo determinado.

El Análisis de canasta es un una metodología muy utilizada ya que permite describir asociaciones entre diferentes items. Este método permite que fácilmente que por ejemplo que identifiquemos las asociaciones propias en un lanzamiento de un nuevo producto, y conocer cual producto juega como rol apalancador y cual de soporte, de tal manera que permita de una mejor forma describir la causalidad entre productos a analizar[1].

Según la revista Forbes, una de las competencias esenciales para cualquier organización moderna es el aprovechamiento de la información con la que cuenta para mejorar sus resultados de negocio. Cada vez es más barato almacenar y procesar información, capacidades que están al alcance de prácticamente todas las empresas, pero también es un hecho que pocas logran ser excelentes en el uso eficaz de la información disponible[2].

Usualmente la data para estos análisis proviene de las transacciones de venta e incluye elementos como:

- Lugar de la transacción o punto de venta (POS por sus siglas en ingles)
- Fecha y hora de la transacción
- Descripción de cada artículo comprado incluyendo precio
- Cantidad comprada de cada articulo
- Valor total del ticket o factura
- Forma de pago

[1] https://consultoriaestadistica.blogspot.com/2017/07/analisis-de-canasta.html

[2] https://www.forbes.com.mx/brand-voice/de-datos-dinero-en-retail-analisis-de-la-canasta-de-compra/

Si la descripción de cada artículo es completa, es decir incluye un ID único al cual se le pueda asociar una familia de productos, costo, SKU, etc., podremos empezar a buscar patrones en los datos y contestar de forma más acertada cuestionamientos asociados a los productos o bien al punto de venta.

Preguntas a contestar con el Análisis de Canasta

Un análisis de canasta se enfoca en contestar preguntas relacionadas con los aspectos siguientes:

1. Efectividad de precios y promociones y sus medios de comunicación (folletos, cupones, etc.)
2. Portafolio y surtido
3. Acomodación del producto en el punto de venta.

Figura 1-1 Preguntas básicas

Efectividad de Precios y Promociones

Este tipo de análisis permitirá entender cómo están respondiendo los clientes a las distintas promociones de manera clara pues se puede evaluar la efectividad comparativa de promociones de descuentos, promociones de tipo 2×1, empaquetados (bundles), entre otros, y determinar para qué conjunto de productos/promociones hay mejores resultados en generación de tráfico de clientes, optimización de márgenes de ganancia, intercambio de marcas, etc.

Con la identificación de reglas de afinidad se identifican aquellos productos cuya venta tiene distintos grados de correlación con otros productos y se determinan aquellos que conviene promocionar juntos, por ejemplo, por estar presentes en canastas de alto valor y evaluar el impacto en el ticket promedio posterior.

El área comercial puede también identificar cuando una promoción está teniendo un efecto contrario al deseado, por ejemplo al descontar un producto con la finalidad de generar tráfico y recuperar margen con artículos adicionales y descubrir que el producto descontado mayoritariamente forma parte de canastas de bajo valor.

También es posible influenciar la compra de productos equiparables de mayor margen, a través de cupones diseñados específicamente para cada cliente en función a su patrón de compra. En operaciones de retail con mostrador, por ejemplo en el caso de farmacias, es posible dar al despachador las sugerencias específicas para buscar cross y up selling en tiempo real mientras atiende al cliente.

Figura 1-2 Preguntas de Precios y Promociones

Portafolio y Surtido

El Market Basket Analysis aporta pruebas valiosas para determinar la efectividad del portafolio y surtido de un punto de venta.

- ¿Qué productos muestran nulo o lento movimiento?,
- ¿Cuáles de ellos son lentos pero están presentes en la mayoría de los tickets de alto valor?,
- ¿Cómo debe prepararse el punto de venta para garantizar abasto en temporadas/categorías específicas?

Estas preguntas se pueden responder de manera ágil y clara, aportando información accionable al área de compras y gestión de categorías para actividades de depuración de catálogos y búsqueda de productos alternos, especialmente de aquellos que son estacionales, promocionales y en general los clasificados como no-resurtibles.

Figura 1-3 Preguntas de Surtido

Acomodación del Producto

El Market Basket Analysis arroja información valiosa sobre la efectividad de distintos diseños de planogramas y sugerencias para su ajuste.

Antes de seguir, veamos el concepto de planograma:

Un planograma es la representación gráfica del acomodo de mercancías o productos en un área específica de un establecimiento comercial que puede ser una góndola, un expositor o un espacio seleccionado[3].

Figura 1-4 Planograma

- ¿Qué productos es conveniente posicionar en conjunto para incentivar venta cruzada?,
- ¿Qué productos están más presentes en tickets de un solo artículo (generan tráfico) y qué tan conveniente es ubicarlos al fondo de la tienda para aprovechar esta característica?

Mediante el análisis de afinidad mencionado se determinan los productos "B" que es conveniente colocar en ubicaciones adyacentes a los productos "A" para incentivar su rotación.

Haciendo distintos ejercicios de modificación en tiendas con patrones de compra comparables se puede determinar el diseño de planograma que mejores resultados arroja en ventas por metro cuadrado para el departamento del que se trate. De la misma forma, en función a la relevancia que se le quiere otorgar a los clústeres de compradores que forman la base de clientes de cada tienda, se pueden tomar decisiones de ajustar los planos de piso para dedicar más o menos metros cuadrados a distintos departamentos.

[3] https://es.wikipedia.org/wiki/Planograma

Capítulo 2

Reglas de Asociación

Rakesh Agrawal y Ramakrishnan Sikrant propusieron un algortimo para identificar asociaciones entre elementos en forma de reglas.

En el año 1994, Sikrant y Agrawal, presentaron un algoritmo cuya función es identificar las asociaciones entre elementos. El algoritmo se vuelve vital cuando las posibles combinaciones entre elementos alcanzan una cantidad considerable y generar todas las reglas por medio de un trabajo manual seria extremadamente complejo.

Las reglas de asociación son reglas que indican cierta relación entre sus conjuntos, sin que esto implique causalidad[4].

Generar estas reglas de asociación es un proceso muy costoso computacionalmente hablando ya que las reglas implican las combinaciones de productos.

La siguiente imagen muestra el resultado para 5 productos {A, B, C, D, E}

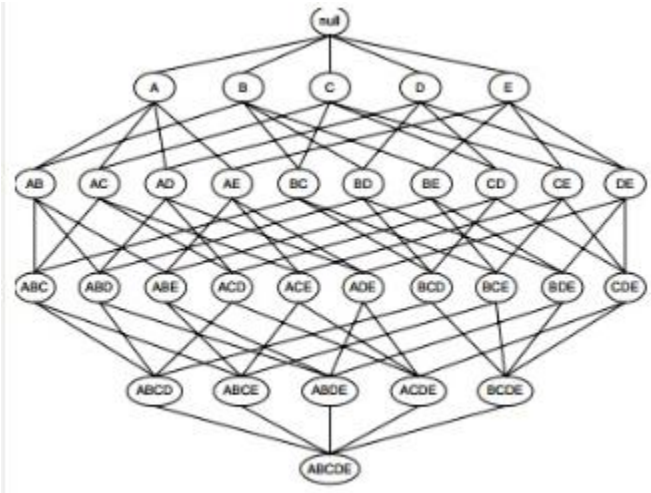

Figura 2-1 Combinaciones para 5 productos

[4] http://ferminpitol.blogspot.com/2014/05/reglas-de-asociacion-algoritmo-apriori.html

Gráficamente se puede observar la complejidad de los resultados, y entra en juego la necesidad de podar o "prunning" para eliminar todos aquellos conjuntos que no son frecuentes.

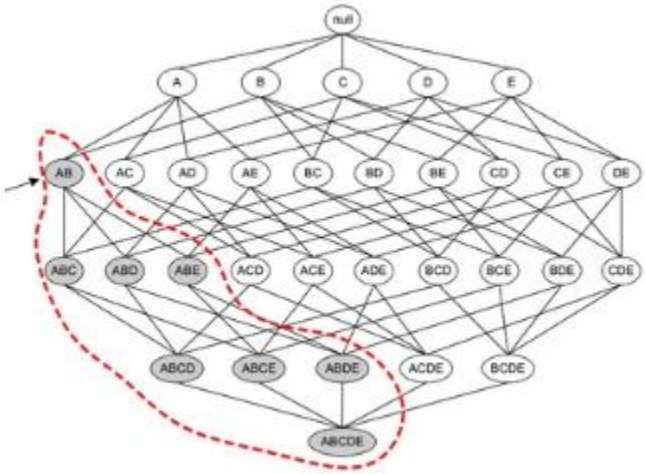

Figura 2-2 Podado de Conjuntos no frecuentes.

La forma de generar reglas de asociación consta de dos pasos:

1. Generación de combinaciones frecuentes: cuyo objetivo es encontrar aquellos conjuntos que sean frecuentes en la base de datos y a la vez considerando un umbral pre-establecido.

2. Generación de reglas: A partir de los conjuntos frecuentes se generan las reglas las cuales están basadas en el índice de confianza.

Antes de estudiar los algoritmos es necesario que se aclaren los conceptos básicos para las reglas de asociación[5]:

Conceptos.

- Elementos u Objetos (Items): dependiendo de la industria y el campo de aplicación, los elementos pueden ser pacientes, eventos, productos, clientes.

- Transacción: Es una operación identificada con un identificador único y que contiene como mínimo 1 elemento.

- Conjunto de elementos (Itemset): Un grupo de elementos que se pueden encontrar en una o varias transacciones.

[5] https://help.xlstat.com/customer/es/portal/articles/2062425-reglas-de-asociaci%C3%B3n-para-an%C3%A1lisis-cesta-de-compra?b_id=9283

- Soporte (Support): Probabilidad de encontrar un elemento o un conjunto de elementos en una transacción. Se estima por el número de veces que un elemento o conjunto de elementos se encuentra en todas las transacciones disponibles. Por ser una probabilidad este valor se encuentra entre 0 y 1.

- Regla (Rule): Una regla define una relación entre dos conjuntos de elementos (Itemsets) X e Y que no tienen elementos en común. X -> Y significa que, si tenemos el elemento X en una transacción, entonces podemos tener Y en la misma transacción.

- Soporte de una regla (Support of a Rule): Probabilidad de encontrar elementos o conjunto de elementos en una transacción. Se estima por el número de veces que ambos elementos o conjuntos de elementos se encuentran en todas las transacciones disponibles. Por ser una probabilidad este valor se encuentra entre 0 y 1.

- Confianza de una regla (Confidence of a Rule): Probabilidad de encontrar un elemento o conjunto de elementos Y en una transacción, sabiendo que el elemento o conjunto de elementos X está en la transacción. Se estima por la frecuencia correspondiente observada (número de veces que X e Y se encuentran en todas las transacciones, dividido por el número que se encuentra X). Este valor se encuentra entre 0 y 1.

- Importancia de una regla (lift of a rule): La importancia de una regla, que es simétrica (importancia(X->Y)=importancia(Y->X)), es el soporte del conjunto de elementos que agrupa X e Y, dividido por el soporte de X y el soporte de Y. Este valor puede ser cualquier número real positivo. Una lift mayor que 1 indica un efecto positivo de X en Y. un valor de 1 significa que no hay efecto, y es como si los elementos o conjuntos de elementos fueran independientes. Una lift menor que 1, significa que hay un efecto negativo de X en Y o viceversa, como si fueran excluyentes entre sí.

$$Rule: X \Rightarrow Y \begin{cases} Supprt = \dfrac{Frequency(X,Y)}{N} \\ Confidence = \dfrac{Frequency(X,Y)}{Frequency(X)} \\ Lift = \dfrac{Support}{Support(X) \times Support(Y)} \end{cases}$$

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B\&C \Rightarrow D$	1/5	1/3	5/9

Figura 2-3 Reglas de asociación[6]

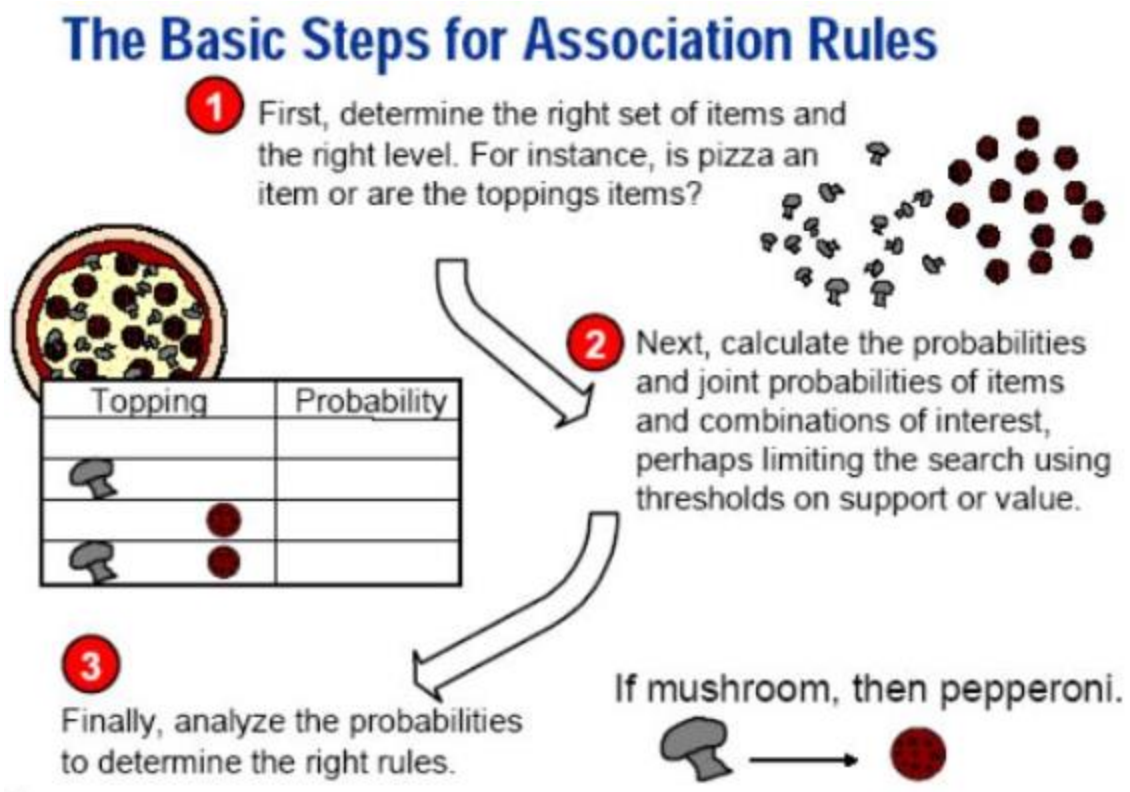

Figura 2-4 Pasos Básicos

[6] https://bicorner.com/2015/07/22/what-the-heck-are-association-rules-in-analytics/

Algoritmos.

Algoritmo AIS: fue el primer algoritmo orientado a encontrar reglas de asociación. La base de datos se escanea varias veces para obtener la frecuencia de aparición de cada elemento. Dado que evaluaba cada item, fue necesario introducir una rutina de podado o "Prunning" de aquellos elementos que no eran representativos y que podían ser eliminados del análisis, en forma resumida el algoritmo hacia lo siguiente:

- Generar los elementos candidatos y calcular en cada transacción si el elemento aparece o no para contarlo.
- Por cada transacción, se determina cual o cuales de los conjuntos de datos anteriores están contenidos en la transacción actual.
- Se genera un nuevo conjunto de elementos pues se extraen los conjuntos más representativos de esta transacción.

La desventaja de este algoritmo radica en su ineficiencia para descartar aquellos elementos representativos sin tener que iterar transacción por transacción.

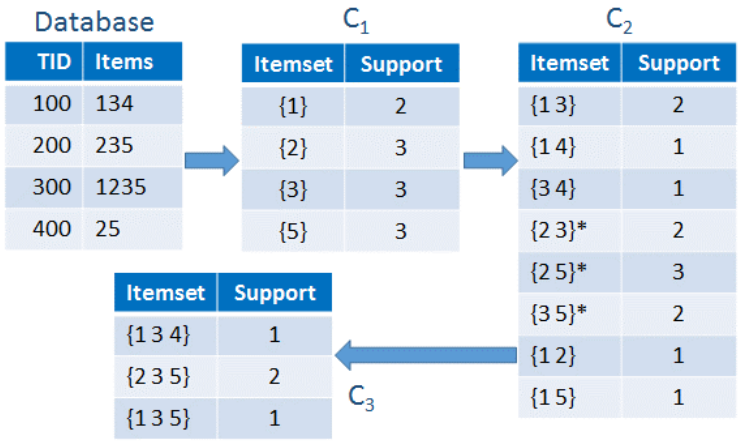

Figura 2-5 Aplicación Algoritmo AIS

Algoritmo SETM (Set Oriented Mining): Este algoritmo nace con el deseo de poder utilizar lenguaje SQL para procesar grandes conjuntos de datos.

Cada transacción es identificada por medio de un Transaction ID único y muchas veces hace el rol de llave primaria en una base de datos. Su forma de trabajar es muy similar al AIS.

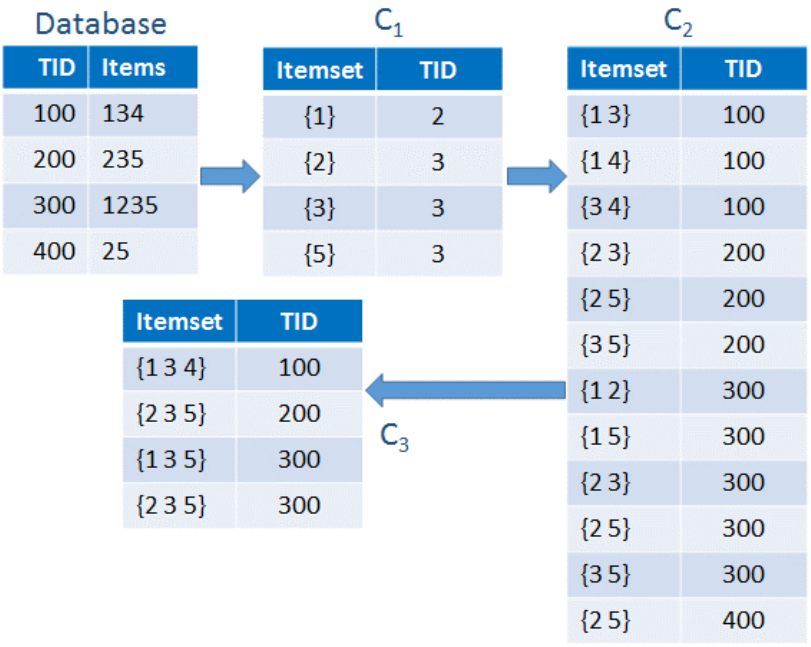

Figura 2-6 Algoritmo SETM

Algoritmo Apriori: Este algoritmo es el más conocido en el mundo de las reglas de asociación. Su estrategia se basa en los soportes de los diferentes conjuntos de elementos y luego por medio del uso de una función para generar candidatos realiza el cálculo del soporte.

Al igual que los algoritmos anteriores, el A priori recorre la base de datos en múltiples ocasiones, la primera iteración es importante por lo que se describe en el párrafo previo en donde calcula el soporte para cada conjunto o elemento identificando los elementos que tienen soporte mayor y menor, con esto se establece un valor de "soporte mínimo", este valor se compara con la siguiente iteración y de esa forma se van descartando elementos y se van convirtiendo los elementos frecuentes en reglas de asociación. El algoritmo se detiene cuando llega al conjunto vacío.

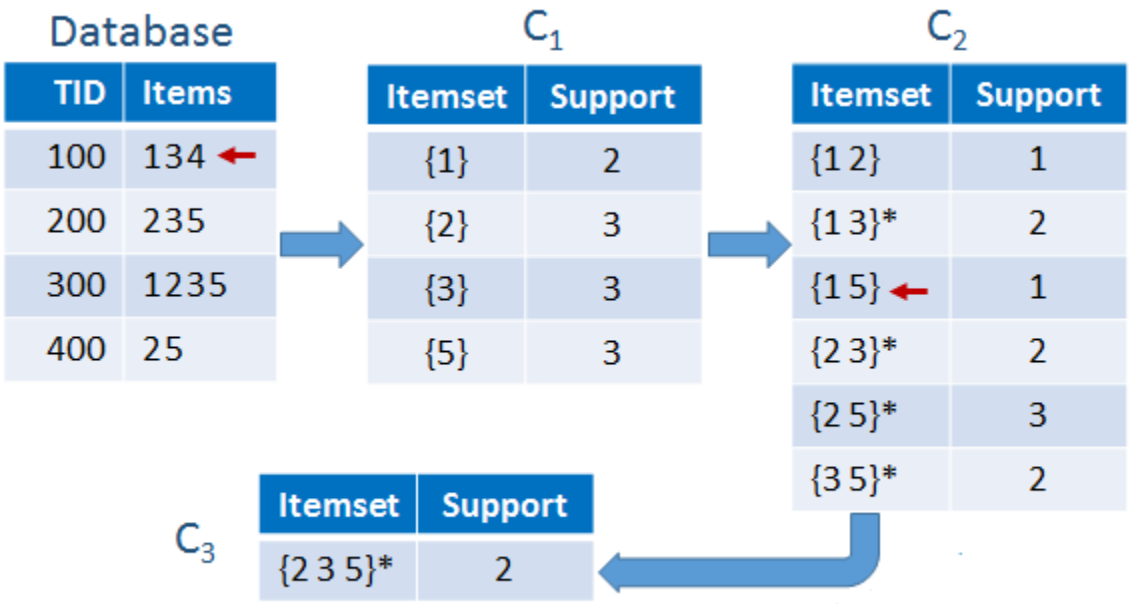

Figura 2-7 Algoritmo Apriori.

Algoritmo FP-Growth[7]**:** En términos generales, el algoritmo emplea una estructura de árbol (Frequent Pattern Tree) donde almacena toda la información de las transacciones. Esta estructura permite comprimir la información de una base de datos de transacciones hasta 200 veces, haciendo posible que pueda ser cargada en memoria RAM. Una vez que la base de datos ha sido comprimida en una estructura FP-Tree, se divide en varias bases de datos condicionales, cada una asociada con un patrón frecuente. Finalmente, cada partición se analiza de forma separada y se concatenan los resultados obtenidos. En la mayoría de casos, FP-Growth es más rápido que Apriori.

Algoritmo Eclat: En el 2000, Zaki propuso un nuevo algoritmo para encontrar patrones frecuentes (itemsets frecuentes) llamado Equivalence Class Transformation (Eclat). La principal diferencia entre este algoritmo y Apriori es la forma en que se escanean y analizan los datos. El algoritmo Apriori emplea transacciones almacenadas de forma horizontal, es decir, todos los elementos que forman una misma transacción están en la misma línea. El algoritmo Eclat, sin embargo, analiza las transacciones en formato vertical, donde cada línea contiene un item y las transacciones en las que aparece ese item.

[7] https://rpubs.com/Joaquin_AR/397172

Bibliografía

- Reglas de asociación y algoritmo Apriori con R
  By Joaquín Amat Rodrigo, 2018

- Association Rule Mining Models and Algorithms
  By Zhang, Chengqi, Zhang, Shichao, 2002

- R Data Analysis Cookbook
  by Kuntal Ganguly,2017