Módulo 6

6.1 Clustering y Reglas de Asociación - ¿Qué, para qué y cómo?

Por **Elena Ruiz Sánchez**
Desarrolladora en Revvity.

Por **Carlos Cano Gutiérrez**
Profesor Titular en CCIA, Universidad de Granada.

Por **Jesús Alcalá Fernández**
Catedrático en CCIA, DaSCI, Universidad de Granada.

---

1. INTRODUCCIÓN A LAS TÉCNICAS DE APRENDIZAJE NO SUPERVISADO DE CLUSTERING Y REGLAS DE ASOCIACIÓN

La disponibilidad de grandes volúmenes de datos nos ofrece una oportunidad para descubrir patrones, relaciones o asociaciones previamente desconocidos en los datos. Las técnicas de Aprendizaje Supervisado (Regresión y Clasificación) que hemos visto en módulos anteriores nos permiten identificar de forma automática modelos predictivos a partir de las variables, pero siempre asumiendo que es necesario especificar *a priori* la variable de salida que queremos predecir a partir del resto.

El Aprendizaje No supervisado, que presentamos en este módulo, permite descubrir patrones o asociaciones en los datos sin especificar previamente una variable de salida. No hay una variable que establezca el valor a predecir. Se trata, simplemente, de *dejar que hablen los datos* para descubrir nuevas relaciones entre variables o nuevos grupos en los que se organizan los datos.

Por ejemplo, en el conjunto de datos de cáncer de melanoma que estamos estudiando en este MOOC, disponemos de una gran cantidad de variables clínicas y derivadas de análisis -ómicos con información sobre las muestras y pacientes. El análisis no supervisado permite identificar nuevos

subgrupos de pacientes de acuerdo con estas variables clínicas y -ómicas. La identificación de estos subgrupos puede tener relevancia clínica, permitiendo tratamientos más personalizados y eficaces. Los autores del estudio original identificaron tres nuevos grupos de pacientes en función de la expresión de sus genes y comprobaron que estos grupos de pacientes mostraban tiempos de supervivencia significativamente diferentes. Es decir, establecieron una nueva forma de agrupar a los pacientes de melanoma, y probaron su utilidad para el pronóstico y la supervivencia.

El análisis no supervisado también permite identificar qué relaciones existen entre las variables del estudio. Por ejemplo, permite estudiar si existe conexión entre los niveles de expresión genética, la presencia de mutaciones, y los niveles de metilación, y cuál es, si la hubiera, esta conexión. Esto resulta de interés, por ejemplo, para entender mejor las bases moleculares de una enfermedad y para identificar conexiones inadvertidas que pueden fraguar nuevas hipótesis para mejorar la prevención o tratamiento de la misma.

Dos de las técnicas de aprendizaje no supervisado más utilizadas para extraer conocimiento interesante a partir de conjuntos de datos con gran cantidad de información son el clustering y las reglas de asociación. A continuación, veremos una breve introducción a los conceptos básicos de cada una de estas técnicas.

2. CLUSTERING

Como se introdujo en la cápsula 3 del Módulo 3, el clustering es un conjunto de técnicas de Aprendizaje No Supervisado cuyo objetivo es la identificación de grupos en los datos. Un grupo (o *cluster*) es un conjunto de instancias que se parecen entre sí. El objetivo de un algoritmo de clustering es entonces agrupar las instancias disponibles de forma que instancias dentro del mismo grupo sean similares entre sí, y diferentes a las instancias de otros grupos.

Es importante notar aquí que el término *similitud* adquiere mucho protagonismo a la hora de definir los clusters. En computación, la similitud entre instancias se define por una función:

$$ff(a, b) \rightarrow R \quad \textit{(Ecuación 1)}$$

donde *a* y *b* son instancias de nuestro conjunto de datos y *R* representa el conjunto de los números reales. La función *f* mide o *cuantifica* la similitud entre estas instancias. Para determinar cómo de similares son dos instancias *a* y *b* se utilizan los valores de las variables de estas instancias. Existen numerosas funciones *tipo* para medir similitud (o su inversa: distancia) entre instancias, el uso de una u otra depende del tipo de variables y del problema. Revisaremos algunas de estas medidas en la sección 2.1.

Además de la definición de similitud/distancia que se proponga, existen multitud de algoritmos diferentes de clustering que pueden proporcionarnos clusters diferentes. En la cápsula 2 revisaremos los principios metodológicos de los algoritmos más populares.

Para explicar visualmente qué es el clustering, presentamos el siguiente ejemplo. La Tabla 1 representa un conjunto de pacientes (instancias) y sus valores para una serie de variables clínicas (columnas).

ID	Sexo	Edad	Estrés	Glucemia
1	M	27	2	165
2	M	52	0	92
3	M	12	1	110
4	V	25	0	115
5	M	48	0	97
6	V	54	2	180
7	V	67	0	107
8	M	72	0	89
9	V	43	2	136
10	M	31	2	165

Tabla 1. Conjunto de datos de variables clínicas de los pacientes de un hospital

Podemos reflexionar sobre distintas opciones para agrupar a los pacientes. Por ejemplo, empleando únicamente la variable *Sexo* podríamos identificar de forma trivial dos grupos de pacientes: varones y mujeres. Utilizando las variables *Edad* y *Glucemia* se pueden agrupar los datos en distintos clusters, aunque ya no resulta tan sencillo identificar estos grupos.

Las visualizaciones resultan muy útiles en esta labor. Por ejemplo, la Figura 1 muestra un gráfico de dispersión (*scatterplot*) de las observaciones en base a algunas de las variables de nuestra tabla: Edad (eje X), Glucemia (eje Y) y Nivel de estrés (Escala de color).

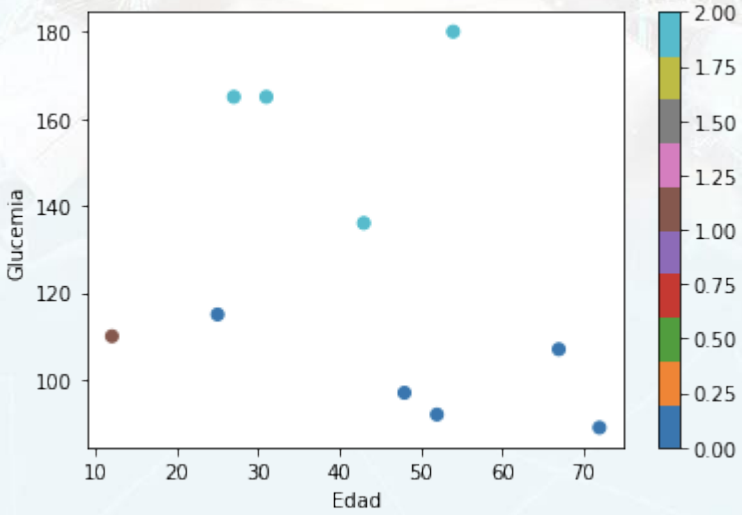

Figura 1. Gráfico de dispersión en base a las variables Edad, Glucemia y Nivel de estrés

Si afrontamos sobre estos datos una tarea de clustering, un resultado posible sería el de la Figura 2.

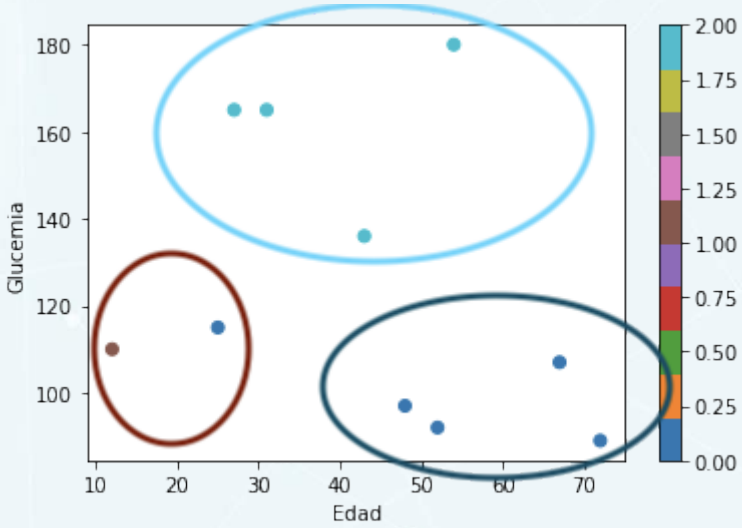

Figura 2. Distribución de clusters sobre el gráfico de dispersión

El grupo marrón representaría los pacientes caracterizados por edad (joven), Glucemia (baja-media) y Estrés (bajo-medio). El grupo azul claro representaría los pacientes caracterizados por Edad (media), Glucemia (alta) y Estrés (Alto). El grupo azul oscuro representaría los pacientes de Edad (Media-Alta), Glucemia (Baja) y Estrés (Bajo). En este caso, parece que la forma en que se distribuyen los objetos en el espacio *invita* a agruparlos de forma intuitiva de este modo. En esto justamente consiste el clustering: en dejar que los datos *revelen estas estructuras* e identificarlas con las técnicas adecuadas.

Es importante notar que el resultado del clustering dependerá de muchos factores: la naturaleza de los datos, las variables que tomemos en consideración, la medida con la que cuantifiquemos la similitud entre instancias y el propio algoritmo de clustering.

Mientras que los agrupamientos que hemos propuesto se pueden apreciar a simple vista con una representación adecuada, a medida que aumente el número de variables y de pacientes de la tabla la tarea de identificar grupos o clusters sobre los datos se convertirá en una tarea más y más difícil de resolver de forma manual.

2.1. MEDIDAS DE DISTANCIA

Existen numerosas funciones para medir la similitud o distancia entre instancias. El uso de una u otra depende del tipo de variables y del problema. Algunas de las más conocidas son la distancia Euclídea (cuando tenemos variables numéricas), la distancia de Levenshtein (para variables de tipo texto), y el coeficiente de Jaccard (para variables enteras).

Antes de calcular una medida de distancia entre instancias, es importante detenerse a considerar si las variables que definen dichas instancias están medidas a distintas escalas. Esto ocurre en los datos de la Tabla 1. Por ejemplo, una diferencia de 2 unidades en la variable *Estrés* es máxima, mientras que la misma diferencia en la variable Glucemia es insignificante. Sin embargo, si calculamos la distancia entre instancias sin tener en cuenta estas distintas escalas, la diferencia entre valores de la variable Glucemia dominará frente a la diferencia entre valores distintos de Estrés, pasando esta última variable a ser prácticamente irrelevante en el cálculo.

De este modo, para calcular adecuadamente la distancia entre instancias es importante que las variables estén normalizadas en la misma escala (por ejemplo [0,1]), para que se les atribuya a todas ellas el mismo peso en el cálculo de la distancia.

Para normalizar los valores de una variable pueden emplearse distintos métodos, desde métodos de conversión de rango (trasladar el valor de la variable de forma proporcional desde el rango [min, max] a [0,1]) hasta métodos más sofisticados. Explicaremos algunos de estos métodos en laCápsula 2 de este módulo.

2.2 ALGORITMOS DE CLUSTERING

Atendiendo a la metodología empleada, podemos clasificar los algoritmos de clustering más populares en las siguientes categorías:

- Algoritmos de clustering jerárquico
- Algoritmos de clustering por particionamiento: K-medias
- Otros algoritmos: Espectrales, Probabilísticos, basados en densidad, etc.

En las siguientes secciones, se describen brevemente las dos estrategias principales, es decir, clustering jerárquico y por particionamiento.

2.2.1 CLUSTERING JERÁRQUICO

Los algoritmos de **clustering jerárquico** son losmás intuitivos y sencillos de comprender. El objetivo de estos algoritmos es construir un ***dendrograma***: una estructura en forma de árbol que indica como ir agrupando instancias similares para formar clusters de mayor y mayor tamaño. Deeste modo, un dendrograma representa una jerarquía de clusters. El concepto de dendrograma tiene su origen en los árboles filogenéticos que representan la relación evolutiva entre especies (Figura 3).

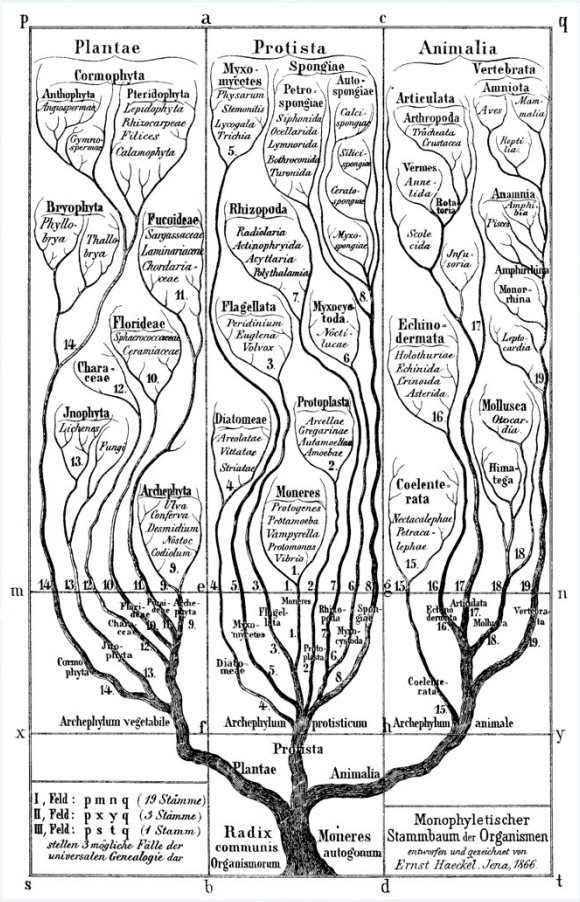

Figura 3. Árbol de la vida según Haeckel, E. H. P. A. (1866).Generelle Morphologie der Organismen : allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von C. Darwin reformirte Decendenz-Theorie. Berlin. Dominio Público.

En esta representación, la base del árbol filogenético representa un único cluster en el que se agrupan todas las especies, y este grupo se va dividiendo en subgrupos a medida que se recorre el árbol en altura y se ramifican sus ramas. De forma similar, un dendrograma suele representarse para clustering tal y como se ilustra en la figura (Figura 2). Es de destacar que la longitud de cada rama es proporcional a la distancia entre los dos grupos conectados con esa rama. A mayor longitud de la rama, mayor la diferencia entre los grupos. Por ejemplo, en este dendrograma (Figura 4) asociado a los datos de la Tabla 1, se agrupan directamente las instancias 10 y 1 por su alta similitud, también junto la instancia 6, constituyendo la rama coloreada de rojo. Otro cluster sería el compuesto por las instancias de la rama verde, que puede desglosarse a su vez en dos subclusters: por un lado, las instancias 4, 3 y 9 y, por otro, las instancias 8, 7, 5 y 2. Y así sucesivamente.

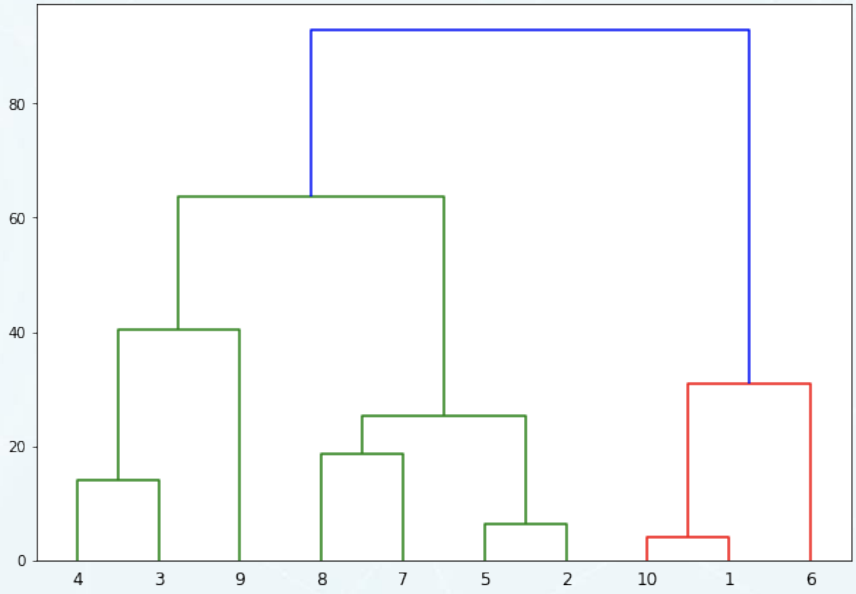

Figura 4. Dendrograma basado en los datos de la Tabla 1.

Existen dos metodologías fundamentales para construir un dendrograma utilizando estrategias de clustering jerárquico: **divisivas** o **aglomerativas**:

- La estrategia divisiva (también llamada *de arriba hacia abajo* o *top-down*) consiste en partir de un único grupo que engloba todas las instancias e ir haciendo iterativamente divisiones de este grupo en dos subgrupos disjuntos, hasta que cada instancia conforme un grupo independiente.

- Los enfoques aglomerativos (*hacia arriba* o *bottom-up*) parten del escenario opuesto: de inicio, cada instancia es un grupo independiente, y el algoritmo debe elegir en cada paso qué dos grupos se unen para formar un grupo más grande. Las decisiones respecto a cómo se dividen o se unen los grupos en cada paso se toman teniendo en cuenta la medida de distancia entre instancias y entre grupos de instancias.

Llegado este punto, es importante destacar que las medidas de distancia mencionadas anteriormente son funciones que se aplican sobre instancias individuales, no sobre grupos de instancias. Por tanto, ¿cómo podemos medir la distancia entre dos clusters?
Para ello, se proponen las medidas de la Tabla 2, que se calculan, a su vez, utilizando una medida de distancia entre dos instancias individuales.

Nombre	La distancia entre dos grupos se define como...
Distancia máxima (complete-linkage)	la distancia entre las instancias más diferentes de estos grupos
Distancia mínima (single-linkage)	la distancia entre las instancias más similares de estos grupos
Distancia promedio (average linkage, UPGMA)	la distancia promedio entre todas las parejas de instancias de un grupo y otro.
Distancia entre centroides (UPGMC)	La distancia entre un representante de cada grupo (centroide).

Tabla 2. Medias para calcular la distancia entre dos clusters

Para una definición matemática de cada criterio, consultar:
https://en.wikipedia.org/w/index.php?title=Hierarchical_clustering

En la Cápsula 2 utilizaremos algoritmos de clustering jerárquico con distintas medidas de distancia entre instancias y entre clusters.

2.2.2 CLUSTERING POR PARTICIONAMIENTO: K-MEDIAS

K-medias es un algoritmo de agrupamiento por particiones cuyo objetivo es separar los objetos en K conjuntos disjuntos minimizando la suma de distancias dentro de los objetos de cada grupo. Para lograr este objetivo, se define para cada uno de los K grupos un objeto denominado *centroide* que

representa las instancias de ese grupo. Los centroides no tienen por qué ser instancias del conjunto inicial de datos, de hecho, lo más habitual es que el centroide sea un objeto nuevo en el que el valor para cada variable es el "promedio" del valor de las instancias del cluster. El objetivo de K-medias es dividir las instancias en K conjuntos minimizando la distancia al cuadrado de las instancias con sus respectivos centroides. En la Cápsula 2 aplicaremos algoritmos de clustering K-medias e identificaremos los centroides de cada grupo.

2.3 EVALUACIÓN DE RESULTADOS Y DETERMINACIÓN DEL NÚMERO DE CLUSTERS

Un problema habitual para los algoritmos de clustering es determinar cuál es el mejor número de grupos en los datos. Numerosos algoritmos requieren este número como punto de partida, antes de ejecutarse. El clustering jerárquico no requiere conocer *a priori* el número de grupos que vamos a obtener en nuestros datos, pero tras obtener el dendrograma con las relaciones jerárquicas entre grupos es habitual *partir* el árbol a una cierta altura para obtener una descomposición en grupos. Plantearse la pregunta: ¿A qué altura partimos el dendrograma? es equivalente a preguntarnos: ¿cuántos grupos hay en nuestros datos?

Existen distintas métricas que nos permiten estimar el mejor número de grupos para nuestros datos. Una de las métricas más populares es el Índice Silueta (*Silhouette*). Se trata de una medida que estima cómo de similar es una instancia respecto a las instancias de su propio cluster (cohesión) comparado con instancias de otros clusters (separación). Este índice toma valores entre [-1, 1], a más alto el valor, más coherentes y mejor separados resultan los clusters obtenidos. En la Cápsula 2 calcularemos algunas de estas métricas y aprenderemos a utilizarlas para estimar *a priori* el número de grupos que mejor recoge la estructura de los datos.

3. REGLAS DE ASOCIACIÓN

Como comentamos en la cápsula 3 del Módulo 3, recordar que, en esta área, clásicamente a las instancias de un conjunto de datos se les llama transacciones, a los valores de cada instancia se le llama ítem, y a un conjunto de ítems se le llama itemset.

Las reglas de asociación son utilizadas para representar dependencias entre los ítems de un conjunto de datos. Estas reglas son expresiones del tipo A → C, donde A y C son itemsets cuya intersección es vacía. Estas reglas representan que cuando en una transacción aparecen los ítems de A, con una alta probabilidad también aparecen los items de C. Por ejemplo, la siguiente regla

podría ser extraída del repositorio de material sanitario de un hospital. Esta regla nos indica que cuando el personal médico utiliza una mascarilla también usa guantes. Este tipo de reglas nos permiten identificar asociaciones entre el material utilizado, lo que ayuda al responsable del material a planificar los pedidos que se realizan para que no haya déficit de ningún producto.

Antes de extraer estas reglas, debemos preprocesar el conjunto de datos (eliminando variables confounding, etc) y decidir qué vamos a considerar como ítems y transacciones, ya que dependiendo de como sea nuestro conjunto de datos puede haber diversas posibilidades. Por ejemplo, podemos distinguir los siguientes tipos de ítems:

- Si cada instancia del conjunto de datos es una lista de elementos (por ejemplo, una lista de productos de un supermercado), los ítems son los elementos que aparecen en las instancias (por ejemplo, un ítem podría ser pañales).
- Si el conjunto de datos contiene un número fijo de variables y cada instancia contiene un valor para cada variable, en este caso un ítem es un par (variable, valor). En el caso de que la variable sea continua, se suele dividir el dominio de la variable en intervalos y se remplaza el valor numérico por el nombre del intervalo al que pertenece.

3.1 MEDIDAS CLÁSICAS DE CALIDAD

Las reglas de asociación son comúnmente evaluadas haciendo uso de las medidas clásicas de soporte y confianza.

La medida de soporte se define como:

- Soporte de un itemset X: Es definido como la frecuencia con la que el itemset X aparece en el conjunto de datos. Se calcula como:

$$Sop(XX) = \frac{nº\ de\ ocurrencias\ de\ XX}{nº\ total\ de\ transacciones}$$

- Soporte de la regla A→C: Es definido como la frecuencia con la que la regla se cumple en el conjunto de datos. Se calcula como:

$$Sop(A \rightarrow C) = \frac{nº\ de\ ocurrencias\ de\ A\ y\ C\ juntos}{nº\ total\ de\ transacciones}$$

Esta medida toma valores en el dominio [0, 1], donde un soporte de 1 indica que aparece en todas las transacciones del conjunto de datos y un soporte de 0 que no aparece en ninguna.

La medida de confianza indica en cuántas transacciones del conjunto de datos en las que aparece el antecedente también aparece el consecuente de la regla. Se define como:

$$Conff(A \rightarrow C) = \frac{Sop(A \rightarrow C)}{Sop(A)} = \frac{nº\ de\ ocurrencias\ de\ A\ y\ C\ juntos}{nº\ de\ ocurrencias\ de\ A}$$

Esta medida toma valores en el rango [0, 1], donde una confianza de 1 indica que siempre que ocurre A también ocurre C, mientras que 0 representa que cuando ocurre A no ocurre C.

ID	Transacciones
1	(Sexo,M), (Estrés, Alto), (Tensión Arterial, Baja)
2	(Sexo,M), (Estrés, Medio), (Tensión Arterial, Baja)
3	(Sexo,V), (Estrés, Bajo), (Tensión Arterial, Baja)
4	(Sexo,V), (Estrés, Alto), (Tensión Arterial, Alta)

Tabla 3. Conjunto de datos de variables clínicas representada como transacciones

Por ejemplo, consideremos el conjunto de datos de la Tabla 3 y la regla (Sexo, M) → (Tensión Arterial, Baja). Los valores de soporte y confianza de esta regla son los siguientes:

- Sop(Sexo, M) = 2 / 4 = 0,5
- Sop((Sexo, M) y (Tensión Arterial, Baja))) = 2 / 4 = 0,5
- Conf((Sexo, M) → (Tensión Arterial, Baja)) = 0,5 / 0,5 = **1**

Destacar que, si en una regla intercambiamos el antecedente y el consecuente, la regla tendrá el mismo soporte, pero puede tener diferente confianza. En nuestro ejemplo, si eres hombre tendrás generalmente una tensión arterial baja, pero si tienes baja la tensión no tienes por qué ser hombre:

- Sop(Tensión Arterial, Baja) = 3 / 4 = 0,75
- Sop((Tensión Arterial, Baja) y (Sexo, M)) = 2 / 4 = 0,5
- Conf((Tensión Arterial, Baja) -> (Sexo, M)) = 0,5 / 0,75 = **0,66**

3.2 ALGORITMOS CLÁSICOS DE EXTRACCIÓN DE REGLAS DE ASOCIACIÓN

Para generar las reglas de asociación, el enfoque clásico primero genera todos los itemsets que tengan un soporte igual o superior a un umbral definido por el usuario, denominados itemsets frecuentes, y después genera a partir de ellos todas las reglas que tengan una confianza igual o superior a un umbral definido por el usuario.

A continuación, introducimos brevemente dos de los algoritmos clásicos más utilizados de la literatura: Apriori y FP-growth.

3.2.1 ALGORITMO APRIORI

Apriori es el primer algoritmo que se utilizó para obtener reglas de asociación a partir de un conjunto de datos. Este algoritmo va creando un árbol con todas las combinaciones posibles entre los ítems del conjunto de datos. En cada nivel del árbol se aumenta en 1 item el tamaño de los itemsets. Para hacer más eficiente el proceso, va eliminando del árbol aquellos itemsets que no son frecuentes, ya que cualquier otro itemset que lo contenga tampoco será frecuente. Una vez generado el árbol, el algoritmo genera a partir de los itemsets frecuentes todas las reglas de asociación con una confianza mayor que el umbral indicado.

La Figura 5 muestra un ejemplo de los itemsets que no se generan al eliminar un 2-itemset que no es frecuente.

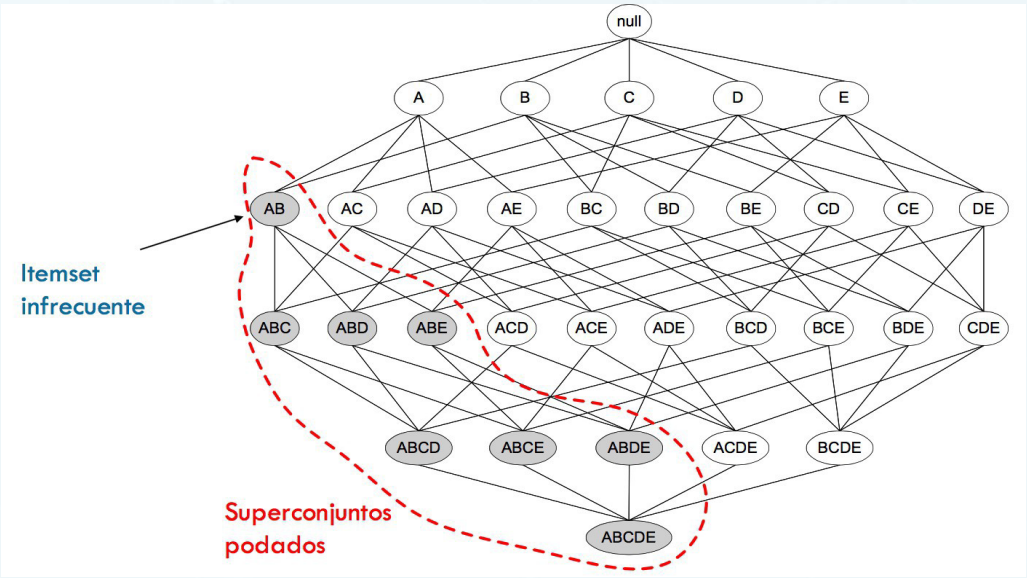

Figura 5. Ejemplo del proceso de generación de los itemsets frecuentes con Apriori.

3.2.2 ALGORITMO FP-GROWTH

FP-growth genera inicialmente una estructura en forma árbol, llamada FP-tree, que permite conocer fácilmente cuantas veces aparece cada itemset en el conjunto de datos. Para ello solo tenemos que descender por el árbol hasta el último ítem que contiene el itemset y consultar el contador del nodo. Por ejemplo, el itemset B,C,E aparece solo en 1 transacción del conjunto de datos según el árbol FP-tree de la Figura 6.

Una vez generada esta estructura, el algoritmo genera los itemsets frecuentes y, a partir de ellos, las reglas de asociación con una confianza mayor al umbral.

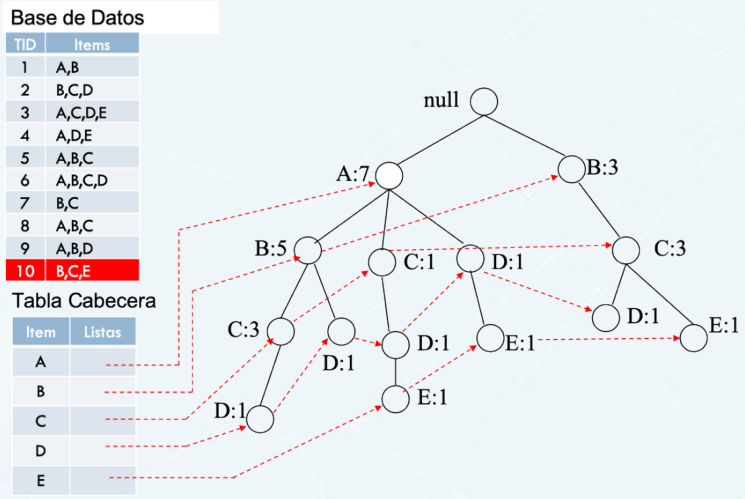

Figure 6. Ejemplo de un árbol FP-tree generado a partir de un conjunto de datos.

Ambos algoritmos (Apriori y FP-growth) generan exactamente las mismas reglas, pero el algoritmo FP-growth es más eficiente al necesitar normalmente menos memoria y tiempo de ejecución. Por esta razón, en la cápsula 3 de este módulo utilizaremos el algoritmo FP-growth para realizar el análisis.

3.3 OTRAS MEDIDAS DE CALIDAD.

Los algoritmos clásicos de extracción de reglas de asociación guían su búsqueda haciendo uso de las medidas clásicas de soporte y confianza. Sin embargo, estas medidas presentan varios problemas en el proceso. Por un lado, un ítem con un soporte muy alto puede ser añadido a cualquier itemset sin que su soporte se vea afectado, permitiendo la creación de un elevado número de itemsets frecuentes. Esto provoca que los algoritmos necesiten más memoria y

requieran un mayor tiempo de ejecución. Por otro lado, la medida de confianza no tiene en cuenta el soporte del consecuente, por lo que, si una regla tiene un itemset muy frecuente en el consecuente, la regla tendrá un valor alto para la confianza. Cualquier antecedente que tenga la regla parecerá un buen predictor del consecuente. Por ejemplo, supongamos las siguientes reglas:

- Regla 1: Conf (A→B) = 1,0
- Regla 2: Conf (C→D) = 0,8

En principio, según la medida de confianza, parece que la regla 1 es la mejor. Sin embargo, supongamos que el soporte del itemset A es 0,2, del itemset B es 0,8, del itemset C es 0,5 y del itemset D es 0,4. La Tabla 4 muestra las 10 transacciones del conjunto de datos.

ID	Transacciones
1	E, F, G
2	F, H, I
3	B, G, I
4	B, E, H
5	B, C, D
6	B, C, D
7	B, C, D
8	B, C, D
9	A, B, C
10	A, B, H

B
A
D
C

Tabla 4. Transacciones en las que aparecen los itemsets A, B, C y D.

Podemos ver en la Tabla 4 como el itemset B aparece en tantas transacciones del conjunto de datos que el itemset A parece un buen predictor. Sin embargo, el itemset A solo aparece en un conjunto reducido de ejemplos de los que contienen el itemset B. Por otro lado, el itemset D aparece en casi todas las transacciones en las que aparece el itemset C, por lo que el itemset C parece un buen predictor del itemset D. Por lo tanto, tras analizar las reglas, podemos concluir que la regla 2 parece ser más interesante que la regla 1.

Debido a estos problemas, durante los últimos años los investigadores han propuesto otras medidas para seleccionar y clasificar las reglas en función de su potencial interés para el usuario.

Algunas de las más utilizadas son:

- ***Lift:*** Esta medida representa la relación entre la confianza de la regla y la confianza esperada de la regla. Su dominio es [0,∞], donde los valores inferiores a 1 indican dependencia negativa, 1 indica independencia y los valores superiores a 1 indican dependencia positiva. Al no estar limitada superiormente, es difícil definir un umbral a partir del cual estudiar las reglas. Debido a ello, estudiaremos cualquier regla con un valor de lift superior a 1, sin que un valor mayor para la medida signifique que la regla es mejor.

- ***Leverage:*** Esta medida mide la diferencia entre la probabilidad conjunta observada y la esperada de AC suponiendo que A y C son independientes. Su dominio es [-1, 1], donde los valores inferiores a 0 representan dependencia negativa, 0 representa independencia y los valores superiores a 0 representan una dependencia positiva.

- ***Conviction:*** Mide el error esperado de la regla, es decir, con qué frecuencia A aparece en una transacción en la que C no aparece. Su dominio es [0,∞], donde los valores inferiores a 1 representan dependencia negativa, 1 representa independencia y los valores superiores a 1 representan dependencia positiva. Esta medida no está limitada superiormente, por lo que tiene el mismo problema que la medida lift. Destacar que si la confianza de la regla es 1 esta medida es ∞, ya que el error esperado es 0.

Todas estas medidas cumplen muchas de las propiedades que son deseables en las medidas de interés para las reglas, pero todas presentan algún problema particular. Debido a ello, se recomienda analizar las reglas considerando varias medidas de interés.

4. REFERENCIAS BIBLIOGRÁFICAS

- R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases. ACM SIGMOD International Conference on Knowledge Discovery and Data Mining, Washington DC (USA), 1993, 207-2016.
- P. Berkhin. A Survey of Clustering Data Mining Techniques, Springer, Berlin, 2006.
- J. Han, M. Kamber. Data Mining: Concepts and Techniques (3ª ed.). Morgan Kaufmann, 2011.
- P-N. Tan, M. Steinbach , A. Karpatne, V. Kumar. Introduction to Data Mining (2ª ed.). Pearson, 2019
- C. Zhang, S. Zhang. Association Rule Mining: Models and Algorithms. Lecture Notes in Computer Science 2307, Springer-Verlag, Berlin, 2002.
- Cancer Genome Atlas Network (2015). Genomic Classification of Cutaneous Melanoma. Cell 161:7 (2015) 1681–1696.
- A. Anguita-Ruiz, A. Segura-Delgado, R. Alcalá, C.M. Aguilera, J. Alcalá-Fdez. eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. PLoS Computational Biology 16:4 (2020) 1-34

REFERENCIAS ADICIONALES

- M.R. Berthold, Ch. Borgelt, F. Höppner, F. Klawonn. Guide to Intelligent Data Analysis, Springer-Verlag, 2010.
- F. Berzal, I. Blanco, D. Sanchez, M. Vila. Measuring the accuracy and interest of association rules: a new framework. Intelligent Data Analysis 6:3 (2002) 221–235.
- K.J. Cios, W. Pedrycz, R.W. Swiniarski, L.A. Kurgan. Data mining: A knowledge discovery approach, Springer, Boston, 2007
- L. Geng, H. Hamilton. Interestingness measures for data mining: a survey. ACM Computing Surveys 38:3 (2006) 1–32.
- H. Li, Y. Wang, D. Zhang, M. Zhang, E. Chang. PFP: parallel FP-growth for query
- Recommendation. ACM Conference on Recommender Systems, Lausanne (Switzerland), 2008, 107–114.
- P. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Alberta (Canada), 2002, 32–41.

- M. Zaki, W. Meira. Data Mining and Machine Learning: Fundamental Concepts and Algorithms (2ª ed.). Cambridge University Press, 2020.