

# Universidad Nacional del Nordeste



## Facultad de Ciencias Exactas y Naturales y Agrimensura Licenciatura en Sistemas de Información

**Cátedra: Bases de Datos II Año: 2019**

**Grupo N° 2 Integrantes:**

* *Meza, Walter Leandro LU:48.957*
* *Aguirre Malagueño, Diego Andrés LU:43.291*
* *Virasoro, Desireé LU:46.384*
* *Billordo, Tania LU:45.971*
* *Domínguez, Cecilia LU:46.686*
* *Brull Valenzuela, Gloria Raquel LU:46.143*

# Técnicas descriptivas y predictivas de clasificación Clústers y árboles de decisión

**Índice**

[Introducción 4](#_TOC_250011)

[El análisis clúster como técnica descriptiva de clasificación 5](#_TOC_250010)

[Medidas de similitud 6](#_TOC_250009)

[Técnica en el análisis clúster 10](#_TOC_250008)

Clusters jerárquicos, secuenciales, aglomerativos y exclusivos (S.A.H.N.) 12

El dendograma en el análisis clúster jerárquico 13

Análisis clúster no jerárquico 13

[Los Árboles de decisión como técnica predictiva de clasificación 18](#_TOC_250007)

[Características de los árboles de decisión 19](#_TOC_250006)

[Herramientas para el trabajo con árboles de decisión 23](#_TOC_250005)

[Árboles CHAID 25](#_TOC_250004)

[Árboles CART. 26](#_TOC_250003)

[Árboles QUEST 27](#_TOC_250002)

Análisis de conglomerados y árboles de decisión como

métodos de segmentación 30

[Conclusión 32](#_TOC_250001)

[Bibliografía 33](#_TOC_250000)

**Índice de Figuras**

Figura [1] Mediadas de similitud. Distancia 7

Figura [2] Mediadas de similitud. Coeficientes de asociación 8

Figura [3] Mediadas de similitud. Coeficientes angulares 9

Figura [4] Dendograma 13

Figura [5] Árbol de clasificación del grado de supervivencia 19

Figura [6] Esquema general del análisis clúster 31

# Índice de Tabla

Tabla [1] Mediadas de similitud. Coeficientes de asociación 8

# INTRODUCCIÓN

Minería de datos es la exploración y análisis de grandes cantidades de datos con el objeto de encontrar patrones y reglas significativas.

Hoy día nuestra sociedad genera grandes cantidades de información que, unido al aumento de las capacidades de almacenamiento, han hecho que todo tipo de organizaciones puedan disponer de una gran cantidad y variedad de datos relativos a su actividad diaria. Esta información ofrece a la empresa una visión perspectiva y prospectiva, es por ello por lo que tiene una función vital en el proceso de toma de decisiones. Sin embargo, mucha de la información recogida en las bases de datos no se encuentra bien estructurada, resultando difícil de explotar desde el punto de vista estadístico por lo que para su utilización es necesario un proceso de tratamiento y análisis exhaustivo de los datos. A este recogido llamaremos minería de datos.

En un proceso de minería de datos se realizan diferentes tareas: **Descriptivas,** para identificar patrones que explican o resumen los datos y **Predictivas**, permiten estimar valores futuros o desconocidos de variables de interés, a partir de otras variables de la Base de Datos.

Para conseguir esos objetivos, el investigador puede utilizar diferentes técnicas como:

* Sistemas de agrupamiento: Consiste en obtener grupos naturales a partir de los datos. También se conoce como segmentación o clustering.
* Reglas de asociación: Su objetivo es identificar relaciones no explícitas entre variables categóricas.
* Reglas de asociación secuenciales: Caso particular de las reglas de asociación, en el que se buscan relaciones temporales en los datos.
* Correlaciones: Proporcionan información sobre el grado de similitud entre variables cuantitativas.
* Clasificación: Cada objeto pertenece a una clase, indicada por el valor de un atributo. El objetivo es predecir la clase a que pertenecen nuevos objetos a partir de las restantes variables.
* Regresión: Su objetivo es predecir el valor que toma una variable cuantitativa en nuevos objetos a partir de la información proporcionada por las restantes variables.

# El análisis clúster como técnica descriptiva de clasificación

El agrupamiento o (Clustering) es la **detección de grupos de individuos.** Se diferencia de la clasificación en el que no se conocen las clases ni su número, (aprendizaje no supervisado). El objetivo es determinar grupo o racimos (Clusters) diferenciados del resto. El análisis clúster es una técnica de Data Mining de clasificación automática de datos. Su finalidad esencial es revelar concentraciones de los datos (casos o variables) para su agrupamiento eficiente en clúster (o conglomerados) según su homogeneidad. El agrupamiento puede realizarse tanto para casos como para variables, pudiendo utilizarse variables cualitativas o cuantitativas. Los grupos de casos o variables de realizan basándose en la proximidad o lejanía de unos con otras, por lo tanto, es esencial el uso adecuando del concepto de distancia. Es fundamental que los elementos dentro de un clúster sean homogéneos y los más diferentes posibles de los contenidos en otros clústers.

El análisis clúster es por tanto una técnica de clasificación, conociéndose también con el nombre de taxonomía numérica. Otros nombres asignados al mismo concepto son análisis de conglomerados, análisis tipológico, clasificación automática y otros. El número de clúster no es conocido de antemano y los grupos se crean en función de la naturaleza de los datos. Se trata por lo tanto de una técnica de clasificación post hoc.

El análisis clúster define grupos tan distintos como sea posible en función de los propios datos sin especificación previa de los citados grupos (técnica de clasificación post hoc). Si las variables de aglomeración están en escalas muy diferentes será necesario estandarizar previamente las variables, o por lo menos trabajar con desviaciones respecto de la media. Es necesario observar también los valores atípicos y desaparecidos porque los métodos jerárquicos no tienen solución con valores perdidos y los valores atípicos deforman las distancias y producen clústers unitarios. También es nocivo para el análisis clúster la presencia de variables correlacionadas, de ahí la importancia del análisis previo de multicolinealidad. Si es necesario se realizan un análisis factorial previo y posteriormente se aglomeran las puntuaciones factoriales. La solución del análisis clúster no tiene por qué ser única, pero no deben encontrarse soluciones contradictorias por distintos métodos. El número de observaciones en cada clúster debe ser relevante, ya que en caso contrario puede haber valores atípicos que difuminen la construcción de los clústers. Los conglomerados deben tener sentido conceptual y no variar mucho al variar la muestra o el método de aglomeración. Los grupos finales serán tan distintos como permitan los

datos. Con estos grupos se podrán realizar otros análisis: descriptivos, discriminante, regresión logística, diferencia...

## Medidas de similitud

Según la clasificación de Sneath y Sokal existen cuatro grandes tipos de medidas de similitud.

* **Distancias:** Se trata de las distintas medidas entre los puntos del espacio definido por los individuos. Se trata de las medidas inversas de las similitudes, es decir, las disimilitudes. El ejemplo más clásico es la distancia euclídea.
* **Coeficientes** de asociación: se utilizan cuando trabajamos con datos cualitativos, aunque también se pueden aplicar a los datos cuantitativos si se está dispuesto a sacrificar alguna información proporcionada por los individuos o las variables. Estas medidas son, básicamente, una forma de medir la concordancia o conformidad entre los estados de las columnas de datos.
* **Coeficientes angulares:** se utilizan para medir la proporcionalidad e independencia entre los vectores que definen los individuos. El más común es el coeficiente de correlación aplicado a las variables continuas.
* **Coeficientes de similitud probabilística:** miden la homogeneidad del sistema por particiones o subparticiones del conjunto de los individuos e incluyen información estadística. La idea de utilizar estos coeficientes se basa en relacionarlos con diferentes clasificaciones utilizando para ellas criterios de bondad buenos ajustes estadísticos. Las principales propiedades de estos coeficientes es que son aditivos, se distribuyen como la Chi-cuadrado y son probabilísticas. Esta última propiedad permite, en aquellos casos en que es posible, establecer una hipótesis nula y contrastarla por los métodos estadísticos tradicionales.

A continuación, se presentan los ejemplos más característicos de cada uno de estos tipos de medidas de similitud.

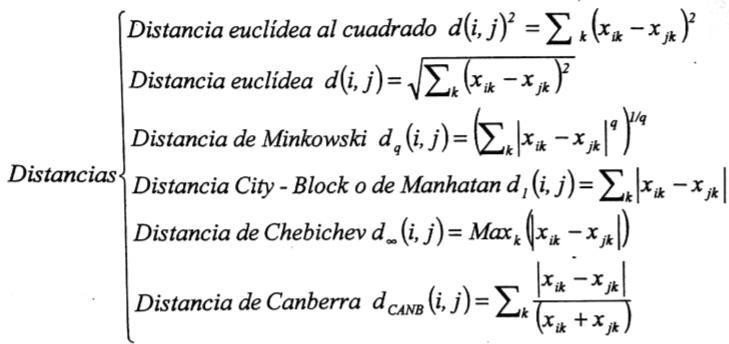


Figura 1.1 Ejemplos de tipos de medidas de similitud

Se observa que la distancia euclídea al cuadrado entre dos individuos se define como la suma de los cuadrados de las diferencias de todas las coordenadas de los dos puntos. La distancia euclídea se define como la raíz cuadrada positiva de la distancia anterior. La distancia de Minkowski es una distancia genérica que da lugar a otras distancias en casos particulares y se define como la raíz q-ésima de la suma de las potencias q-ésimas de las diferencias, en valor absoluto, de las coordenadas de los dos puntos considerados. La distancia City-Block o distancia de Manhatan, es un caso particular de la distancia o medida de Minkowski cuando q = 1 y resulta ser la suma de las diferencias, en valor absoluto, de todas las coordenadas de los dos individuos cuya distancia se calcula. El valor de esta medida es cero para la similitud perfecta y aumenta a medida que los objetos son más disimilares. La distancia de Chebychev se define como el caso límite de la medida de Minkowski para q tendiendo a infinito, es decir, es el máximo de las diferencias absolutas de los valores de todas las coordenadas. La distancia Canberra es una modificación de la distancia Manhatan que es sensible a proporciones y no sólo a valores absolutos.

Los coeficientes de asociación suelen utilizarse para el caso de variables cualitativas, y en general para el caso de datos binarios (o dicotómicos), que son aquéllos que sólo pueden presentar dos opciones (blanco - negro, sí - no, hombre – mujer, verdadero - falso, etc.). En este caso existen diferentes medidas de proximidad o similitud, que se verán a continuación,

partiendo de una tabla de frecuencias 2x2 en la que se representa el número de elementos de la población en los que se constata la presencia o ausencia del carácter (variable cualitativa) en estudio.

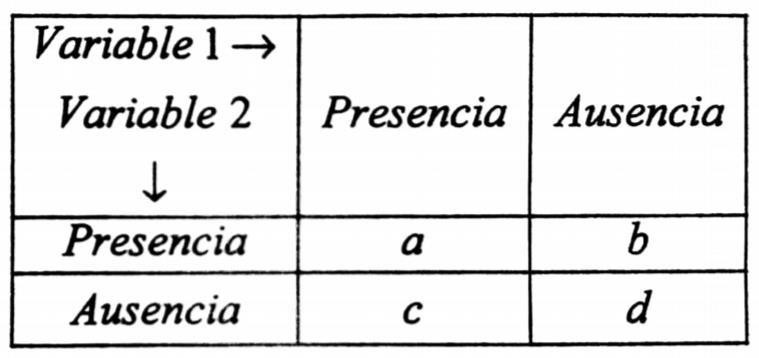


Tabla [1] Mediadas de similitud. Coeficientes de asociación

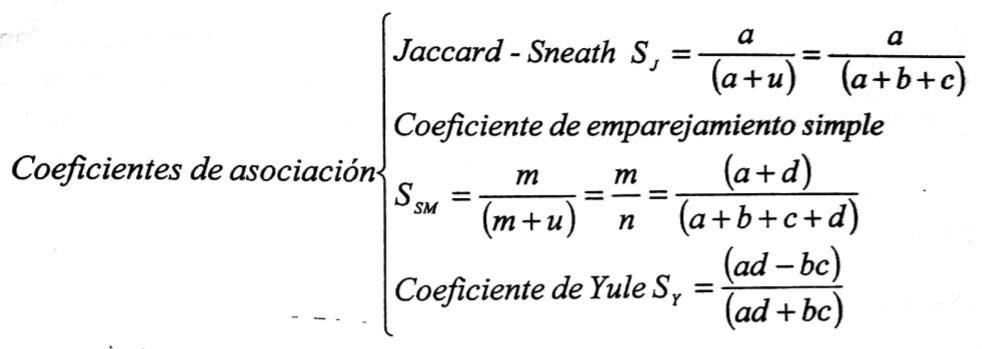
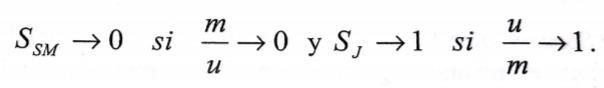


Figura [2] Medidas de similitud. Coeficientes de asociación

El coeficiente de Jaccard – Sneath no tiene en cuenta los emparejamientos negativos, y se define como el número de emparejamientos positivos entre la suma de los emparejamientos positivos y los desacuerdos

A partir de su expresión se deduce que SJ, tiende a cero cuando a/u tiende a cero, esto es, Sj es cero cuando el número de emparejamientos positivos coincide con el de desacuerdos, también SJ, tiende a uno cuando u tiende a cero, es decir, SJ vale uno cuando no hay desacuerdos. El coeficiente de Yule varía entre +1 y -1. El coeficiente de emparejamiento simple se define como el cociente entre el número de emparejamientos y el número total de casos considerados. De su expresión se deduce:



En el caso de los coeficientes angulares su campo de variación está entre -1 y +1. Los valores cercanos a 0 indican disimilitud entre los individuos y los valores que se acercan a +1 o a -1 indican similitud positiva o negativa respectivamente. El cálculo de este coeficiente entre los individuos i y j se realiza en función de Xi y XJ que son las medias correspondientes a los individuos i y j.

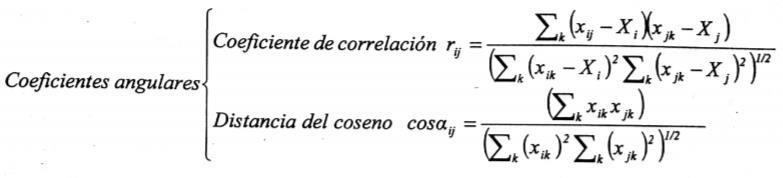


Figura [3] Medidas de similitud. Coeficientes angulares

Los coeficientes de similitud probabilística calculan la probabilidad acumulada de que un par de individuos i y j, sean tan similares, o más, que lo que empíricamente se puede afirmar sobre la base de la distribución observada.

## Técnica en el análisis clúster

Ya sabemos que el análisis clúster es un conjunto de métodos y técnicas estadísticas que permiten describir y reconocer diferentes agrupaciones que subyacen en un conjunto de datos, es decir, permiten clasificar, o dividir en grupos más o menos homogéneos, un conjunto de individuos que están definidos por diferentes variables.

El objetivo principal del análisis de conglomerado, consiste, por tanto, en conseguir una o más particiones de un conjunto de individuos en base a determinadas características de los mismos. Estas características estarán definidas por las puntuaciones que cada uno de ellos tienen con relación a diferentes variables.

Así, se podrá decir que dos individuos son similares si pertenecen a la misma clase, grupo o conglomerado o clúster. Si se consigue este objetivo, se tendrá que todos los individuos que están contenidos en el mismo conglomerado se parecerán entre sí, y serán diferentes de los individuos que pertenecen a otros conglomerados. Por tanto, los miembros de un conglomerado gozaran de unas características comunes que los diferencian de los miembros de otros conglomerados. Estas características deberán, ser genéricas, y es claro que difícilmente una única característica podrá definir un conglomerado.

El método para ejecutar un análisis de conglomerado comienza con la selección de los individuos objeto del estudio, incluyendo en algunos casos su codificación a partir de las variables o caracteres que los definen y su transformación adecuada para someterlos al análisis si es necesario (tipificación de variables, desviaciones respecto de la media, etc.).

Una vez determinadas las disimilitudes de los individuos, se proceden a ejecutar el algoritmo que formara las diferentes agrupaciones o conglomerados de individuos. Determinada ya la clasificación, el paso siguiente consiste en obtener una representación gráfica de los conglomerados obtenidos, de modo que se puedan visualizar los resultados alcanzados. Este proceso se lleva a cabo mediante un dendrograma. Conseguido el propósito de clasificación, la última fase a llevar a cabo es la interpretación de los resultados obtenidos.

Los diferentes métodos de análisis de conglomerados surgen de las distintas formas de llevar a cabo la agrupación de los individuos, es decir, dependiendo del algoritmo que se utilice para llevar a cabo la agrupación de individuos o grupos de individuos, se obtiene diferentes métodos de análisis de conglomerados. Una clasificación de los métodos de análisis de conglomerados basadas en los algoritmos de agrupación de individuos podría ser la siguiente:

* + Métodos Aglomerativos-Divisivos: un método es aglomerativo si considera tanto grupos como individuos y sucesivamente va fusionando los dos grupos más similares, hasta llegar a una clasificación determinada; mientras que un método es divisivo si parte de un solo grupo formado por todos los individuos, de modo que en cada etapa va separando individuos de los grupos establecidos anteriormente, formándose así nuevos grupos.
  + Métodos Jerárquicos-No jerárquicos: un método es jerárquico si consiste en una secuencia de g+1 clústers: G0….Gg en la que G0 es la partición disjunta de todos los individuos y Gg es el conjunto partición. El número de partes de cada una de las particiones disminuye progresivamente, lo que hace que estas sean cada vez más amplias y menos homogéneas. Por el contrario, un método se dice no jerárquico cuando se forman grupos homogéneos sin establecer relaciones de orden o únicas entre dichos grupos
  + Métodos Solapados-Exclusivos: un método es solapado si admite que un individuo pueda pertenecer a dos grupos simultáneamente en alguna de las etapas de clasificación, mientras que se dice exclusivo si ningún individuo puede pertenecer simultáneamente a dos grupos en la misma etapa
  + Métodos Secuenciales Simultáneos: un método es secuencial si a cada grupo se le aplica el mismo algoritmo en forma recursiva, mientras que los métodos simultáneos son aquellos en los que la clasificación se logra por una simple y no reiterada operación sobre los individuos
  + Métodos Monotéticos-Politéticos: un método se dice monotético si está basado en una característica única de los objetos a clasificar; mientras que es politético si se basa en varias características de los mismos, sin exigir todos los objetos las posean, aunque si las suficientes como para poder justificar la analogía entre los miembros de una misma clase
  + Métodos Directos-Iterativos: un método es directo si utiliza algoritmos en los que una vez asignado un individuo a un grupo ya no se saca del mismo mientras que los métodos iterativos corrigen las asignaciones previas volviendo a comprobar en posteriores iteraciones si la asignación de un individuo a un conglomerado es óptima, llevando a cabo un nuevo reagrupamiento de los individuos si es necesario.
  + Métodos Ponderados-No ponderados: los métodos no ponderados son aquellos que establecen el mismo peso a todas las características de los individuos a clasificar, mientras que los ponderados hacen recaer mayor peso en determinadas características.
  + Métodos Adaptativos-No adaptativos: los métodos no adaptativos son aquellos para los que el algoritmo utilizado se dirige hacia una solución en la que el método de formación de conglomerados es fijo y está predeterminado, mientras que los adaptativos (menos utilizados) son aquellos que de alguna manera aprenden durante el proceso de formación de los grupos y modifican el criterio de optimización o la medida de similitud a utilizar.

## Clústers jerárquicos, secuenciales, aglomerativos y exclusivos (S.A.H.N.)

Los métodos de análisis de conglomerados que más se usan son los que son a la vez secuenciales, aglomerativos, jerárquicos y exclusivos, y que reciben el acrónimo, en lengua inglesa, de S.A.H.N. (Sequential, Agglomerative, Hierarchic y Nonoverlaping). En todos los métodos de tipo S.A.H.N. se siguen dos pasos fundamentales en el proceso de elaboración de los conglomerados. El primero de ellos es que los coeficientes de similitud o disimilitud entre los nuevos conglomerados establecidos y los candidatos potenciales a ser admitidos se recalcula en cada etapa, y el otro es el criterio de admisión de nuevos miembros a un conglomerado ya establecido. Entre los diferentes métodos de análisis de conglomerados de tipo S.A.H.N. tenemos los siguientes:

* Método de unión simple, entorno o vecino más cercano o método del mínimo.
* Método de la distancia máxima o método del máximo.
* Método de la media o de la distancia promedio no ponderado.
* Método de la media ponderada o de la distancia Promedio Ponderado.
* Método de la mediana o de la distancia mediana
* Método del Centroide o de la Distancia Prototipo.
* Método de Ward o de mínima varianza

## El dendograma en el análisis clúster jerárquico

Es habitual en la investigación la necesidad de clasificar los datos en grupos con estructura arborescente de dependencia, de acuerdo con diferentes niveles de jerarquía.

Partiendo de tantos grupos iniciales como individuos se estudian, se trata de conseguir agrupaciones sucesivas entre ellos de forma que progresivamente se vayan integrando en clústers los cuales, a su vez, se unirán entre sí en un nivel superior formando grupos mayores que más tarde se juntaran hasta llegar al clúster final que contiene todos los casos analizados. La representación gráfica de estas etapas de formación de grupos, a modo de árbol invertido, se denomina dendograma y se representa a continuación:

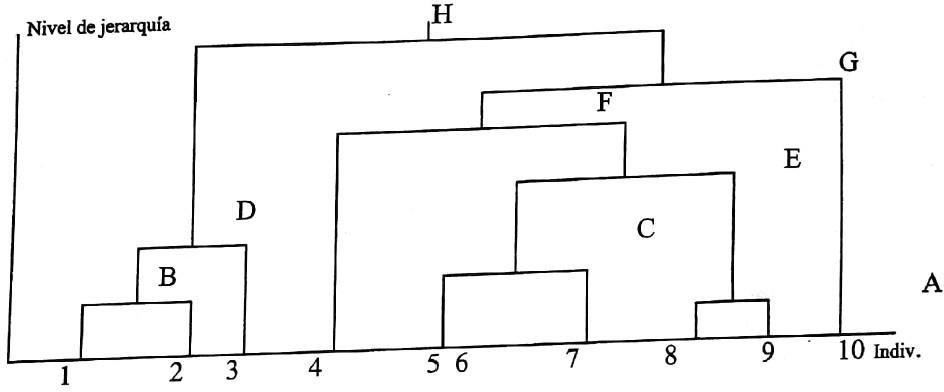


Figura [4] Dendograma.

La figura, que corresponde a un estudio de los individuos, muestra como el 8 y el 9 se agrupan en un primer clúster (A). En un nivel inmediatamente superior, se unen los individuos 1 y 2 (clúster B), y enseguida los 5, 6, y 7 (C). Un paso siguiente engloba el clúster B con el individuo 3 (D); y así sucesivamente hasta que todos ellos quedan estructurados al conseguir, en el nivel más alto, el clúster total (H) que reúne los 10 casos.

## Análisis clúster no jerárquico

La clasificación de todos los casos de una tabla de datos en grupos separados que configura el propio análisis proporciona clusters no jerárquicos. Esta denominación alude a la no existencia de una vertical de dependencia entre los grupos formados y, por consiguiente,

éstos no se presentan en distintos niveles de jerarquía. El análisis precisa que el investigador fije de antemano el número de clusters en que quiere agrupar sus datos.

Como puede no existir un número definido de grupos o, si existe, generalmente no se conoce, la prueba debe ser repetida con diferente número a fin de tantear la clasificación que mejor se ajuste al objetivo del problema, o la de más clara interpretación.

Los métodos no jerárquicos, también se conocen como métodos partitivos o de optimización, dado que, como hemos visto, tienen por objetivo realizar una sola partición de los individuos en K grupos. Esto implica que el investigador debe especificar a priori los grupos que deben ser formados. Esta es, posiblemente, la principal diferencia de los métodos jerárquicos. La asignación de individuos a los grupos se hace mediante algún proceso que optimice el criterio de selección. Otra diferencia está en que estos métodos trabajan con la matriz de datos original, no requieren su conversión en una matriz de proximidades. Pedret agrupa métodos no jerárquicos en las cuatro familias siguientes: reasignación, búsqueda de la densidad, directos y reducción de dimensiones.

Los métodos de reasignación permiten que un individuo asignado a un grupo en un determinado paso del proceso sea reasignado a otro grupo en un paso posterior si esto optimiza el criterio de selección. El proceso termina cuando no quedan individuos cuya reasignación permita optimizar el resultado que se ha conseguido. Algunos de los algoritmos más conocidos dentro de estos métodos son el método K-means (o K-medias) de McQueen (1967), el Quick Cluster Analysis y el método de Forg, los cuales se suelen agrupar bajo el nombre de métodos centroides o centros de gravedad. Por otra parte, está el método de las nubes dinámicas, debido a Diday.

Los métodos de búsqueda de la densidad presentan una aproximación tipológica y una aproximación probabilística. En la primera aproximación, los grupos se forman buscando las zonas en las cuales se da una mayor concentración de individuos. Entre los algoritmos más conocidos dentro de estos métodos están el análisis modal de Wishart, el método de Taxmap de Carmichael y Sneath, y el metodo de Fortin. En la segunda aproximación, se parte del postulado de que las variables siguen una ley de probabilidad según la cual los parámetros varían de un grupo a otro. Se trata de encontrar los individuos que pertenecen a la misma distribución. Destaca en esta aproximación el método de las combinaciones de Wolf.

Los métodos directos permiten clasificar simultáneamente a los individuos y a las variables. Las entidades agrupadas, ya no son los individuos o las variables, sino que son las observaciones, es decir, los cruces que configuran la matriz de datos. Los métodos de

reducción de dimensiones, como el análisis factorial de tipo Q, guardan relación con el análisis clúster. Este método consiste en buscar factores en el espacio de los individuos, correspondiendo cada factor a un grupo. La interpretación de los grupos puede ser compleja dado que cada individuo puede corresponder a varios factores diferentes.

Resulta muy intuitivo suponer que una clasificación correcta debe ser aquélla en que la dispersión dentro de cada grupo formado sea la menor posible. Esta condición se denomina criterio de varianza, y lleva a seleccionar una configuración cuando la suma de las varianzas dentro de cada grupo (varianza residual) sea mínima.

Se han propuesto diversos algoritmos de clasificación no jerárquica, basados en minimizar progresivamente esta varianza, que difieren en la elección de los clústers visiónales que necesita el arranque del proceso y en el método de asignación de individuos a los grupos. Aquí se describen los dos más utilizados.

El algoritmo de las H-medias parte de una primera configuración arbitraria de grupos con su correspondiente media, eligiendo un primer individuo de arranque de cada grupo y asignando posteriormente cada caso al grupo cuya media es más cercana. Una vez que todos los casos han sido ubicados, calcula de nuevo las medias centroides y la toma en lugar de los primeros individuos como una mejor aproximación de los mismos, repitiendo el proceso mientras la varianza residual vaya disminuyendo. La partición de arranque define el número de clusters que lógicamente, puede disminuir si ningún caso es asignado a alguno de ellos.

El algoritmo de las K-medias, el más importante desde los puntos de vista conceptual y práctico, parte también de unas medias arbitrarias y, mediante pruebas sucesivas, contrasta el efecto que sobre la varianza residual tiene la asignación de cada uno de los casos a cada uno de los grupos. El valor mínimo de varianza determina una configuración de nuevos grupos con sus respectivas medias. Se asignan otra vez todos los casos a estos nuevos centroides en un proceso que se repite hasta que ninguna transferencia puede ya disminuir la varianza residual; o se alcance otro criterio de parada: un número limitado de pasos de iteración simplemente, que la diferencia obtenida entre los centroides de dos pasos consecutivos sea menor que un valor prefijado. El procedimiento configura los grupos maximizando, a su vez, la distancia entre sus centros de gravedad. Como la varianza total es fija, minimizar la residual hace máxima la factorial o intergrupos. Y puesto que minimizar la varianza residual es equivalente a conseguir que sea mínima la suma de distancias al

cuadrado desde los casos a la media del cluster al que van a ser asignados, es esta distancia euclídea al cuadrado la utilizada por el método.

Como se comprueban los casos secuencialmente para ver su influencia individual, el cálculo puede verse afectado por el orden de los mismos en la tabla; pese a lo cual es el algoritmo que mejores resultados produce. Otras variantes propuestas a este método llevan a clasificaciones muy similares.

Como cualquier otro método de clasificación no jerárquica, proporciona una solución final única para el número de clusters elegido, a la que se llegará con menor número de iteraciones cuanto más cerca estén las "medias" de arranque de las que van a ser finalmente obtenidas. Los programas automáticos seleccionan usualmente estos primeros valores, tantos como grupos se pretenda formar, entre los puntos más separados de la nube.

Los clusters no jerárquicos están indicados para grandes tablas de datos, y son también útiles para la detección de casos atípicos: Si se elige previamente un número elevado de grupos, superior al deseado, aquéllos que contengan muy escaso número de individuos servirían para detectar casos extremos que podrían distorsionar la configuración.

Es aconsejable realizar el análisis definitivo sin ellos, ya con el número deseado de grupos para después, opcionalmente, asignar los atípicos al cluster adecuado que habrá sido formado sin su influencia distorsionante. Un problema importante que tiene el investigador para clasificar sus datos en grupos es, como se ha dicho, la elección de un número adecuado de clusters. Puesto que siempre será conveniente efectuar varios tanteos, la selección del más apropiado al fenómeno que estudia ha de basarse en criterios tanto matemáticos como de interpretabilidad. Entre los primeros, se han definido numerosos indicadores de adecuación como el Criterio cúbico de clusters y la Pseudo F que se describen en el ejemplo de aplicación práctica. El uso inteligente de estos criterios, combinado con la interpretabilidad práctica de los grupos, constituye el arte de la decisión en la clasificación multivariante de datos.

Matemáticamente, un método de clasificación no jerarquizado consiste en formar un número prefijado K de clases homogéneas excluyentes, pero con máxima divergencia entre las clases. Las K clases o clusters forman una única partición (clustering) y no están organizadas jerárquicamente ni relacionadas entre si. La clasificación no jerárquica o de reagrupamiento tiene una estructura matemática menos precisa que la clasificación

jerárquica. El número de métodos existentes ha crecido excesivamente en los últimos años y algunos problemas derivados de su utilización todavía no han sido resueltos.

Supongamos que N es el número de sujetos a clasificar formando K grupos, respecto a n variables X1.,... .Xn. Sean W, B y T las matrices de dispersión dentro grupos, entre grupos y total respectivamente. Como T = B + W y T no depende de la forma en que han sido agrupados los sujetos, un criterio razonable de clasificación consiste en construir K grupos de forma que B sea máxima o W sea mínima, siguiendo algún criterio apropiado. Algunos de estos criterios son

1. Minimizar Traza(W)
2. Minimizar Determinate(W)
3. Minimizar Det(W)/Det(T)
4. Maximizar Traza(W-1B)
5. Minimizar ∑ki=1∑Ni

Los criterios a) y b) se justifican porque tratan de minimizar la magnitud de la matriz W. El criterio e) es llamado criterio de Wilks y es equivalente a b) porque det(T) es constante. El caso d) es el llamado criterio de Hottelling y el criterio e) representa la suma de las distancias de Mahalanobis de cada sujeto al centroide del grupo al que es asignado.

Como el número de formas de agrupar N sujetos en K grupos es del orden de k ^(n) \* k!, una vez elegido el criterio de optimización, es necesario seguir algún algoritmo adecuado de clasificación para evitar un número tan elevado de agrupamientos.

El método ISODATA, introducido por Ball y Hall (1967), es uno de los más conocidos. Esencialmente consiste en partir de K clases (construidas por ejemplo aleatoriamente) y reasignar un sujeto de una clase i a una clase j si se mejora el criterio elegido de optimización.

# Los Árboles de decisión como técnica predictiva de clasificación

Los arboles de decisión, también llamados arboles de clasificación, presentan de hecho un aspecto similar a los dendrogramas del análisis de conglomerados jerárquico, aunque se construyen e interpretan de forma completamente distinta. Se trata de una seria de métodos muy flexibles, que pueden manejar un gran número de variables y complicadas iteraciones en ellas, y cuyos resultados resulta fácilmente interpretables para cualquier persona. Los arboles de clasificación son particiones secuenciales del conjunto de datos realizadas para maximizar las diferencias de las variables dependientes o criterio base; por tanto conllevan la división de las observaciones en grupos que difieren respecto a una variable de interés. Estos métodos, se caracterizan además por desarrollar un proceso de división de forma arborescente. Mediante diferentes índices y procedimientos estadísticos se determina la división más discriminante de entre los criterios seleccionados, es decir, aquella que permite diferenciar mejor a los distintos grupos del criterio base, obteniéndose de este modo la primera segmentación. A continuación, se realizan nuevas segmentaciones de cada uno de los resultados de los segmentos, y así sucesivamente hasta que el proceso finaliza con normas estadísticas preestablecidas o interrumpidas voluntariamente en cualquier momento por el investigador. Además, los criterios descriptores no tienen por qué aparecer en el mismo orden para todos los segmentos, y un criterio puede aparecer más de una vez para un mismo segmento. Al final, enumerando los criterios a través de los cuales se ha llegado a un segmento determinado se obtiene el perfil del mismo.

Por ejemplo, supongamos que deseamos conocer que pasajeros del Titanic tuvieron más probabilidades de sobrevivir a su hundimiento, y que características estuvieron asociadas a la supervivencia al naufragio. En este caso, la variable de interés (VD) es el grado de supervivencia. Podríamos entonces dividir a los pasajeros en grupos de edad, sexo y clase en la que viajaban y observar la proporción de supervivientes de cada grupo.

Un procedimiento arborescente selecciona automáticamente los grupos homogéneos con la mayor diferencia en proporción de supervivientes entre ellos, este caso, el sexo (hombres y mujeres). El siguiente paso es subdividir cada uno de los grupos en función de otra característica, resultando que los hombres son divididos en adultos y niños, mientras que las mujeres se dividen en grupos basados en la clase en que viajan en el barco. Usar diferentes predictores en cada nivel del proceso de división supone una forma sencilla y elegante de manejar interacciones que a menudo complican en exceso los modelos lineales tradicionales. Cuando se ha completado el proceso de subdivisión, el resultado es un

conjunto de reglas que puede verse fácilmente por medio de un árbol. Por ejemplo: si un pasajero del Titanic es un hombre y es adulto, entonces tiene una probabilidad de sobrevivir del 20 por ciento. Además, la proporción de supervivencia en cada una de las subdivisiones se puede usar con fines predictivos para vaticinar el grado de supervivencia de los miembros de ese grupo. Un árbol de clasificación del grado de supervivencia de los pasajeros del Titanic podría ser el que se observa en la figura siguiente.

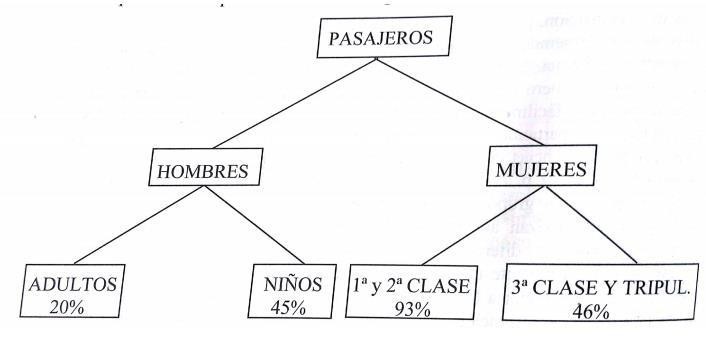


Figura [5] Árbol de clasificación del grado de supervivencia de los pasajeros del Titanic.

## Características de los árboles de decisión:

Las características más importantes en el trabajo con los árboles de decisión son la especificación de los criterios para minimizar los costes, la selección del método de división y la elección del tramo de árbol adecuado o problema del sobreajuste.

En cuanto a la especificación de los criterios para minimizar los costes, el objetivo de cualquier análisis arborescente es clasificar o predecir con el mínimo coste. En la mayoría de las situaciones, los costes hacen referencia simplemente a la proporción de casos mal clasificados. Sin embargo, no siempre es así, ya que al menos los factores más importantes también pueden afectar los costes finales de una clasificación: las probabilidades a priori y los costes de una clasificación errónea.

Las probabilidades a priori, también conocidas como ponderaciones de clase, especifican la probabilidad, sin tener ningún conocimiento previo de los valores de los predictores, de que un caso caiga en cada una de las clases de la variable dependiente. Por ejemplo, en un estudio educacional, se observa que en general en secundaria existen muchos menos abandonos escolares que chicos que siguen estudiando, en consecuencia, la probabilidad a priori de que un estudiante abandone la escuela es menor que la probabilidad de que permanezca en ella. Las probabilidades a priori constituyen elementos centrales de cualquier árbol de decisión y prácticamente todos los programas actuales ofertan la opción de utilizar ponderaciones estimadas según las proporciones de cada clase. Pero no siempre resulta adecuado utilizar este tipo de ponderaciones. Así, si disponemos de un fichero de datos con 400 personas que contestaron y 9600 que no contestaron y se supone que las dos submuestras son representativas, estos árboles aplican automáticamente probabilidades del 0,04 para los que respondieron y del 0,96 para os que no respondieron. El problema es que, dado que el árbol trata de maximizar el número de casos que clasifica correctamente, dedicará la mayor parte de su esfuerzo a clasificar a los que no respondieron. Incluso, aunque el árbol clasificase mal a todos los que respondieron, sólo se equivocaría en el 4 por ciento de los casos. Semejante modelo sería teóricamente preciso, pero prácticamente inútil, por eso los paquetes ofrecen también la posibilidad de especificar probabilidades iguales para cada clase, tratando a todas ellas como si fuesen del mismo tamaño.

Otro factor que influye en el coste de una clasificación son los costes de una acción errónea. La mayor parte de los árboles de decisión permiten al investigador especificar también costes variables de una clasificación errónea. Un ejemplo de coste variable es que para un banco sea 10 veces más costoso no ser capaz de detectar una transacción fraudulenta que detener una transacción legal, o que sea 200 veces más costoso para un hospital no detectar una enfermedad contagiosa en un paciente que diagnosticar como contagiosa una enfermedad que lo es. Normalmente, estos costes se computan cuando él árbol ya ha sido desarrollado completamente y, por tanto, no tienen impacto sobre su estructura básica. En cuanto a la selección del método de división, se trata de escoger el método con el que seleccionar, en cada uno de los niveles del proceso de división, la mejor división posible del mejor predictor. En la actualidad predominan fundamentalmente dos enfoques:

**Métodos exhaustivos**: El más conocido y conceptualmente más simple consiste en examinar todas las posibles divisiones de los datos según cada predictor y seleccionar la división que produce clasificaciones más puras. No existe una medida universalmente

preferible para cualquier problema, por lo que resulta conveniente disponer de una amplia gama de coeficientes para poder seleccionar el que mejor se adapta a cada problema concreto. Sin embargo, la mayor parte del software actual cubre un número muy limitado de índices y varios de los principales paquetes sólo incluyen un único coeficiente. Dos de los principales problemas de los métodos exhaustivos son su complejidad computacional (cuando contamos con un gran número de predictores con muchos niveles cada uno, el número total de divisiones posibles que deben ser examinadas por el programa llega a ser enorme), y sobre todo el sesgo que muestran a la hora de seleccionar variables, ya que estos métodos tienden a seleccionar primero los predictores con más categorías. Los siguientes métodos logran solventar ambos problemas.

**Métodos de tipo discriminante:** Este tipo de métodos siguen un proceso diferente, computacionalmente más sencillo. En vez de buscar a la vez la mejor variable y su mejor punto de división, abordan estos dos problemas por separado. En cada nodo, calculan primero un test χ2 (para cada predictor categórico) o un ANOVA (para cada predictor métrico), seleccionándose de entre todas las variables significativas, la que proporciona probabilidades asociadas menores. En una segunda fase, se aplica un análisis discriminante sobre el predictor con el fin de encontrar la mejor división posible de la variable. Estos procedimientos son utilizados en los árboles QUEST.

En cuanto a la elección del tamaño adecuado o problema del sobreajuste, tenemos que una característica de los árboles de clasificación es que si no se establece ningún límite en el número de divisiones a ejecutar, se consigue siempre una clasificación pura, en la que cada nodo contiene únicamente una sola clase de objetos. Las clasificaciones puras presentan varios inconvenientes porque suelen ser poco realistas, se corre el riesgo de encontrarnos con muy pocos elementos en cada clase y, además, llegan a extraer toda la información de los datos, incluida aquella información idiosincrática (ruido), propia de la muestra particular que estamos utilizando. Esta falta de generalización, de replicación a otras muestras, se conoce como sobreajuste (o sobreaprendizaje en el marco de las redes neuronales), y para combatirla se han planteado numerosas estrategias diferentes y en ocasiones complementarias. Dos de las principales son las reglas de parada y la poda, que se especifican a continuación:

**Reglas de parada:** Una primera estrategia consiste en detener la generación de nuevas divisiones cuando éstas supongan una mejora muy pequeña de la predicción.

Por ejemplo, si con diez divisiones clasificamos correctamente al 90 por ciento de los sujetos y con 11 divisiones el porcentaje sube al 90,1 por ciento, no tiene mucho sentido añadir más al árbol. Hay muchas reglas de parada directa como ésta para detener automáticamente el proceso de construcción del árbol. Las principales son las siguientes:

* + Extensión máxima del árbol, es decir, numero de niveles máximos permitidos por debajo del nodo raíz.
  + Mínimo número de casos en un nodo, especifica que los nodos no sobrepasen un número determinado de casos.
  + Mínima fracción de objetos, que consiste en que los nodos no contengan más casos que una fracción determinada del tamaño de una o más clases.

La regla de parada la establece a priori el propio investigador, en función de investigaciones pasadas, análisis previos, o incluso en función de su propia experiencia o intuición.

Una fuente de información diagnostica para determinar si el tamaño de nuestro árbol es el adecuado consiste en evaluar, una vez detenido el proceso de división, su calidad predictiva en muestras distintas a las utilizadas para su cálculo. Se han descripto al menos 3 formas opcionales de llevar a cabo una validación cruzada de este tipo:

* + Validación cruzada en dos mitades: consiste en dividir los datos disponibles en 2 partes, las muestras de estimación y las muestras de validación, desarrollar un árbol a partir de las muestras de estimación y utilizarlo para predecir la clasificación de la muestra de validación.
  + Validación cruzada en v partes: de las muestras disponibles se extrae aleatoriamente v submuestras y se calculan v arboles de clasificación, cada vez dejando fuera una de las v submuestras para validar el análisis, de tal manera que cada submuestra se utilice v-1 veces para obtener el árbol y una sola vez para validarlo. Opción sumamente útil con muestras pequeñas.
  + Validación cruzada global: aquí, se replica el análisis completo un número determinado de veces apartando una fracción de los casos (casos holdout) para validar el árbol seleccionado. Resulta especialmente útil en combinación con las técnicas automáticas de selección de árboles, lo que enlaza con la segunda de las estrategias para evitar el sobreajuste: la poda.

*Poda.* Tras analizar diferentes reglas de parada durante años, Breiman, Firedman, Losen y Stone (1984) llegan a la conclusión de que resulta imposible especificar una regla que sea totalmente fiable. Existe siempre el riesgo de no descubrir estructuras relevantes en los datos debido a una finalización prematura del análisis. Por ello, sugieren un enfoque alternativo en 2 fases. En una primera fase se desarrolla un enorme árbol que contenga cientos o incluso miles de nodos. En la segunda fase, el árbol es podado, eliminándose las ramas innecesarias hasta dar con el tamaño adecuado del árbol. Este proceso automático y retrospectivo, compara simultáneamente todos los posibles subárboles resultados de podar en diferentes grados el árbol original.

El primer y principal algoritmo de poda se debe precisamente a Breiman et al. (1984). Consta de 2 pasos: primero se construye una secuencia de subárboles cada vez más pequeños, todos ellos como resultado de podar cada vez más el árbol original. A continuación, cada subárbol de la secuencia es validado en una nueva muestra (mediante una validación cruzada en v partes), escogiéndose el que menores costes de validación cruzada (VC) presenta. Para calcular los costes de validación se utiliza una función que penaliza la progresiva complejidad del árbol a medida que este va teniendo más ramas. Esta función entra en funcionamiento cuando se alcanza un valor crítico que sobrepasa los costes del proceso de división (cada vez menores). En ese momento, los costes dejan de descender y comienzan a ascender ligeramente, y es alrededor de ese punto de inflexión donde se localiza el tamaño idóneo del árbol.

Como suele ser común que existan varios árboles con costes VC cerca del mínimo, Breiman et al. (1984) sugieren utilizar la regla 1ET: seleccionar el árbol de menor complejidad de entre todos los árboles que no superen el mínimo coste VC más una medida de su error típico. Se han propuesto otros métodos de poda (desvianza-complejidad, error reducido y poda pesimista), aunque ninguno ha demostrado ser claramente superior a los demás (Murthy, 1998), porque la elección de uno u otro queda en manos del investigador (y de la disponibilidad del paquete estadístico correspondiente).

## Herramientas para el trabajo con árboles de decisión.

Diversas empresas han desarrollado software específico de árboles de clasificación. Una clasificación de uno de los principales programas podría ser la siguiente:

Familia CART: CART, Tree(S), etc. su propósito inicial es la predicción estadística. Realiza únicamente divisiones binarias. Recurre a la validación cruzada y a la poda para determinar el tamaño correcto del árbol. La variable dependiente puede ser cuantitativa o nominal. Las variables predictoras pueden ser nominales u ordinales, aunque las últimas versiones también admiten variables continuas.

Familia CLS: CLS, ID3, C4.5, C5.0, etc. Su propósito inicial es detectar relaciones estadísticas complejas. El número de ramas puede originar varía entre dos y el número de categoría del predictor. Para determinar el tamaño del árbol utiliza tests de significación estadística (con ajuste de multiplicidad en las últimas versiones). AID, MAID y XAID trabajan con variables dependientes cuantitativas. RHAID, CHAID y TREEDISC con nominales, aunque las versiones del CHAID desarrollada por Statistical Innovation y que distribuye el SPSS puede manejar variables dependientes cuantitativas (categorizándolas).

Métodos de tipo discriminante. FACT y QUEST. Su propósito inicial es solucionar problemas de los métodos exhaustivos. En concreto, tratar de eliminar el denominado sesgo de selección de la variable, que presentan métodos como CART y que consiste en la tendencia a seleccionar en primer lugar las variables con más categorías. FACT elimina este sesgo solo cuando utiliza variables dependientes ordinales. QUEST logra eliminar este sesgo, sea la VD nominal u ordinal. En primer caso, diseñados para trabajar como variables dependientes categóricas como continuas. FACT divide a la población en tantos grupos como categorías tiene la variable seleccionado, QUEST establece divisiones binarias.

Combinaciones lineales: OC1, arboles SE, etc. Su propósito inicial es detectar relaciones lineales combinadas con el aprendizaje de conceptos. El número de ramas varía entre dos y el número de categorías del predictor.

Modelos híbridos: IND, Knowledge Seeker, etc. Su propósito inicial es combinar métodos de otras familias. IND combina el CART y C4.5, así como métodos bayesianos y de codificación mínima. Knowledge Seeker combina CHAID y el ID3 con un novedoso ajuste de multiplicidad.

Pero tres de los procedimientos arborescentes que actualmente gozan de una mayor aceptación tanto en los ámbitos teóricos como aplicativos son: los arboles CHAID (Kass, 1980), CART (Breiman et al., 1984) y QUEST (Loh y Shih, 1997).

## Árboles CHAID.

CHAID o Chi-square Automatic Interaction Detector (Kass, 1980, Biggs, de Ville y Suen, 1991) representa la culminación de una serie de métodos basados en el detector Automático de Interaccione (AID) de Morgan y Sonquist (1963). CHAID es un método exploratorio de análisis de datos útil para identificar variables importantes y sus interacciones con fines de segmentación, análisis descriptivos o paso como previo a otros análisis posteriores.

La medida dependiente puede ser cualitativa (nominal u ordinal) o cuantitativas. Para variables cualitativas, el análisis lleva a cabo una serie de análisis x2 entre las variables dependientes y predictoras. En el caso de variables dependientes cuantitativas se recurre a métodos de análisis de varianza, en lo que los intervalos (divisiones) se determinan óptimamente para las variables independientes de forma que maximicen la capacidad para explicar la varianza de la medida dependiente.

Para dividir cada nodo, Kass (1980) comienza localizando el par de categorías permisible del predictor (el conjunto de pares permisibles viene dado por el tipo de predictor que está siendo analizado) con el menor valor de x2. Si el nivel de significación es menor que un cierto nivel crítico, se unen ambas categorías y repite el proceso. Si es mayor, se convierten 2 candidatas a la división de la variable. Este proceso continua con cada par de categorías hasta que dejan de producirse uniones y posibles divisiones. La última candidata a la división (que generalmente no suele coincidir con la división más significativa) es la que se elige para dividir al predictor. El proceso se repite de forma recursiva en cada uno de los nodos hasta que se activa cualquiera de las reglas de parada del proceso. Par mitigar el sesgo de selección de la variable, Kass recurre a un ajuste de Bonferroni.

Este enfoque original de Kass ahorra bastante tiempo de computación. Sin embargo, no garantiza que sea capaz de encontrar realmente la mejor división posible en cada modo. Únicamente una búsqueda exhaustiva en cada nodo de todos los conjuntos de categorías candidatos a una división puede lograr esto. El CHAID exhaustivo, propuesto por Biggs, de Ville y Suen (1991), selecciona siempre la división más significativa de todas. Los autores encontraron también que el ajuste de Bonferroni que utiliza Kass penaliza en exceso a las variables con muchos niveles, y aunque logra controlar el error tipo I, presenta un error tipo II demasiado alto. Por ello, CHAID exhaustivo trata a todas las variables por igual, independientemente del tipo de variable y del número de categorías. Además, permite trabajar tanto con variables dependientes categóricas como métricas. Las variables

categóricas utilizan el estadístico χ2 y dan lugar a un árbol de clasificación. Las variables métricas utilizan el estadístico F y dan lugar a lo que se conoce como árboles de regresión. También permite utilizar predictores de tipo métrico, mediante su conversión previa en variables categóricas. Tanto el CHAID clásico como el exhaustivo pueden producir divisiones de la VD en más de dos grupos (al contrario que CART, que sólo produce divisiones binarias).

## Árboles CART.

CART (Classification And Regression Trees), También llamados C&RT en algunos programas y textos de estadística, constituyen una alternativa al CHAID exhaustivo para desarrollar árboles de clasificación. De hecho, el algoritmo fue desarrollado por Breiman, Friedman, Losen y Stone a principios de los ochenta (1984) para intentar superar algunas de las deficiencias y debilidades que por entonces mostraba la formulación original del CHAID.

Hasta la aparición de la versión exhaustiva, CHAID parecía limitado a VDs nominales y VIs categóricas, y sin embargo existía la necesidad de un método que permitiese utilizar criterios y predictores de cualquier nivel de medida.

Como su propio nombre indica, CART resulta apropiado para árboles de clasificación (VD cualitativa) o de regresión (VD cuantitativa) y, como característica diferencial, genera árboles binarios. Se construye dividiendo la muestra en subconjuntos de datos. En cada división, se evalúa cada predictor para encontrar el mejor punto de corte (con predictores cuantitativos) o las mejores agrupaciones de categorías (con predictores categóricos). A continuación, se comparan también los predictores, seleccionándose el predictor y la división que produce la mayor bondad de ajuste. Por tanto, como regla de división, este programa utiliza medidas de bondad, también denominadas medidas de impureza de nodos (Breiman et al., 1984 Murthy, 1998). Para variables cuantitativas, algunas de las medidas propuestas han sido la reducción del error cuadrático o la desviación media absoluta de la mediana.

Con variables cualitativas, una de las medidas más utilizadas es el coeficiente Gini, que evalúa la probabilidad de una mala clasificación para los casos de un nodo y alcanza un valor de 0 cuando la clasificación es perfecta. Otros coeficientes similares (y que, por tanto, comparten esencialmente las mismas ventajas e inconvenientes) incluyen la medida χ2 de

Bartlett, y la medida G-cuadrado, similar al χ2 de máxima verosimilitud y del que hablaremos en el modelado de muestras finitas. Muy diferentes a los anteriores es el coeficiente twoing, se trata de una medida basada en agrupamiento de la VD en las dos mejores subclases y posterior cálculo de su varianza. Twoing tiende a producir árboles completamente distintos a los que se consiguen utilizando Gini. En concreto, mientras que los árboles twoing tienden a equilibrar más el reparto de los datos en cada una de las dos categorías, los coeficientes Gini y las medidas de entropía tienden a producir árboles que dividen los datos en subgrupos excesivamente descompensados.

¿Qué árbol es mejor? Desde el punto de vista de su precisión predictiva, puede ser que ambos tengan un comportamiento similar, aunque probablemente la mayor parte de los investigadores de mercado se decanten por el árbol twoing. En cualquier caso, es fundamental que los diferentes programas ofrezcan una buena selección de medidas (entre las que no debe faltar twoing) para que sea el investigador el que decida cuál utilizar. CART fue el primer programa que tuvo en cuenta esto, ofertando hasta nueve diferentes.

Relacionado con esto, conviene señalar también que, aunque algunos autores consideran que los árboles construidos a partir de distribuciones de probabilidad son más fiables que los construidos directamente a partir de los valores de los atributos (Shang y Breiman, 1996), otros autores han descrito varios aspectos problemáticos del hecho de utilizar tests de significación estadística como reglas de parada (Neville 1999), opción que es la que utilizan CHAID o QUEST.

## Árboles QUEST.

El acrónimo QUEST proviene de Quick, Unbiased, Efficient, Statistical Tree (Loh y Shih, 1997). Se trata de un algoritmo de clasificación arborescente creado específicamente para solventar dos de los principales problemas que presentan métodos como CART y CHAID exhaustivo a la hora de dividir un grupo de sujetos en función de una variable independiente. En primer lugar, la complejidad computacional. Una variable ordinal con n valores en un nodo conlleva (n-1) divisiones, por lo que el número de cálculos en cada nodo aumenta de forma proporcional el número de valores. En el caso de variables categóricas, el número de cálculos aumenta exponencialmente al número de categorías, siendo en general de (2m-1-

1) para una variable con m valores. En segundo lugar, los sesgos en la selección de variables. Pero un problema más serio desde el punto de vista interpretativo y de

generalización de resultados es que los métodos exhaustivos tienden a seleccionar aquellas variables que cuentan con un mayor número de categorías. Doyle (1973) fue el primero en darse cuenta de ello en el contexto de los algoritmos AID y THAID, y, más recientemente, Loh y Shih (1997) han encontrado evidencia empírica de este sesgo en los métodos exhaustivos.

El algoritmo QUEST emplea un enfoque cuyo cálculo resulta mucho más sencillo que los métodos exhaustivos. En vez de intentar seleccionar a la vez el mejor predictor y su mejor punto de corte, QUEST aborda estos dos problemas por separado. En cada nodo, calcula la asociación entre cada predictor y la VD mediante el estadístico F del ANOVA o la F de Levene (en el caso predictores continuos y ordinales) o mediante una χ2 de Pearson (en el caso de predictores nominales). Para asegurarse divisiones binarias de la VD, se aplica un algoritmo conglomerativo que fuerza a crear siempre dos superclases en el predictor. Finalmente, se selecciona el predictor que presenta la mayor asociación con la VD, es decir, el menor valor p dentro de los significativos (corregidos según Bonferroni para eliminar el sesgo en la selección de variables).

En una segunda fase, para encontrar el mejor punto de corte se recurre a un análisis discriminante cuadrático (cuya ventaja sobre el lineal es que permite acomodar varianzas distintas). El proceso se repite de forma recursiva hasta que salta alguna de las reglas de parada. Gracias a este procedimiento de cálculo, QUEST no presenta sesgos de respuesta, muestra una mayor simplicidad de cálculo, permite incluir métodos de validación mediante poda y permite incorporar combinaciones lineales de variables.

Comparando QUEST con los métodos exhaustivos, se ha comprobado que cuando se utilizan divisiones univariadas no existe un ganador claro. Sin embargo, los Árboles QUEST basados en divisiones que resultan de combinaciones lineales han demostrado ser más rápidos y, sobre todo, precisos que los procedimientos exhaustivos, los cuales utilizan típicamente divisiones univariadas (Loh y Shih, 1997). Una reciente comparación de los principales métodos arborescentes (Song y Yoon, 2000) ha demostrado que QUEST es el único método que no muestra sesgos serios a la hora de seleccionar una variable u otra. CHAID presenta un suave sesgo y CART está claramente sesgado hacia predictores continuos y/o con muchas categorías. Por su parte, el C5.0 de Quinlan (1993) es el que presenta el problema más serio de sesgo, dando como resultado un poder estadístico muy pobre.

En líneas generales, QUEST parece ser superior a CART y éste, a su vez, superior a CHAID. Los resultados dependen en gran medida del tipo de problema concreto que se esté abordando.

# Análisis de conglomerados y árboles de decisión como métodos de segmentación

El análisis de conglomerados, también llamado análisis clúster, constituye uno de los procedimientos estadísticos más utilizados hoy en día para la segmentación. La definición de la propia técnica se asemeja, de hecho, a los fines genéricos que persigue la segmentación: identificar grupos de sujetos lo más heterogéneos posible entre sí y lo más homogéneos posible dentro de cada grupo.

El análisis establece dichos grupos basándose en la similitud que presentan un conjunto de entidades (por ejemplo, turistas) respecto a una serie de características que el investigador ha de especificar previamente (motivaciones, necesidades, beneficios buscados, etc). No obstante, es el análisis, y no el analista, el que finalmente extrae los grupos de sujetos y sus características definitorias: número de segmentos, número de integrantes de cada segmento, etc. El análisis de conglomerados constituye, por tanto, el ejemplo paradigmático del enfoque de segmentación post hoc. Asimismo, se trata de un método descriptivo de segmentación.

Por otro lado, los árboles de decisión constituyen métodos predictivos de segmentación y son la herramienta más utilizada hoy en día para segmentar.

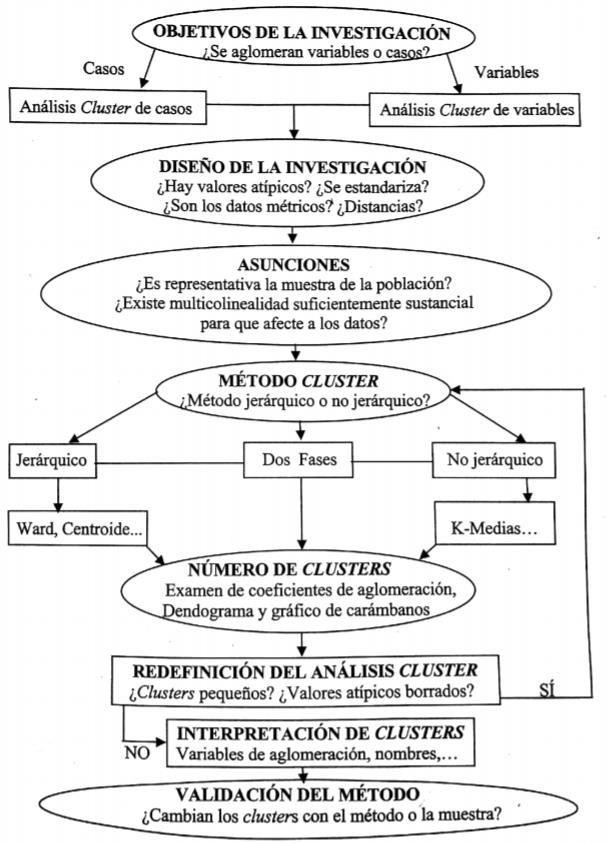


Figura [6] Esquema general del análisis clúster.

31

# CONCLUSIÓN

La minería de datos surge a partir de la necesidad de analizar grandes volúmenes de datos, con el objetivo de obtener modelos y patrones descriptivos o predictivos, siendo esta una parte de lo que se conoce como KDD: descubrimiento de conocimiento a partir de base de datos.

Dentro de las técnicas descriptivas tenemos el análisis de cluster, que es una técnica de clasificación de información automática cuya finalidad es revelar concentraciones en los datos para su agrupamiento eficiente en clusters según su homogeneidad, haciendo uso de diferentes medidas y coeficientes para clasificar los datos.

Dentro de las técnicas predictivas tenemos los arboles de decisión o clasificación, que son particiones secuenciales del conjunto de datos, cuyo objetivo es maximizar las diferencias de la variable dependiente o criterio base, es decir dividir las observaciones en grupos que difieren de acuerdo a este criterio y con ello poder obtener las diferentes probabilidades de que un grupo cumpla con los criterios previamente establecidos. Los árboles de decisión proveen de una herramienta de clasificación muy potente ya que son fáciles de comprender y modificar, y el modelo desarrollado se puede expresar como un conjunto de normas de decisión. En comparación con otros métodos, los árboles de decisión se pueden crear de una forma relativamente rápida.

# BIBLIOGRAFÍA

C. Pérez López, D. Santín Gonzáles. Minería de Datos: Técnicas y Herramientas. Thomson Paraninfo S.A. España. 2007.

[http://exa.unne.edu.ar/depar/areas/informatica/dad/BDII/Presentaciones\_Proyector/Min eria\_de\_Datos\_Tecnicas\_Descriptivas\_y\_Predictivas\_de\_Clasificacion.pdf](http://exa.unne.edu.ar/depar/areas/informatica/dad/BDII/Presentaciones_Proyector/Min%20eria_de_Datos_Tecnicas_Descriptivas_y_Predictivas_de_Clasificacion.pdf)

<http://exa.unne.edu.ar/depar/areas/informatica/dad/BDII/Presentaciones/Mineria_de_Datos_Introduccion.pdf>