CLASIFICACIÓN NOSUPERVISADA

SALVADOR PINTOS

SETIEMBRE/2000

Índice General

Capítulo 1

Clasificación no supervisada

1.1 Qué es clasificar

Clasificar ha sido, y es hoy en día, un problema básico en un amplio espectro de disciplinas que se extiende de las ciencias básicas a la ingeniería . Dependiendo de la ciencia y del periodo histórico, el problema de clasificar lleva consigo su propia terminología: desde taxonomía al tan actual reconocimiento de patrones.

El propósito fundamental, común a todas las disciplinas, consiste en hacer una partición de un conjunto de objetos en categorías. Estas categorías, (o sus sinónimos: clases, conglomerados, grupos, etc.), se construyen de manera tal que un objeto en un grupo dado es similar, en algún sentido, a cualquier otro del mismo grupo; y objetos en distintos grupos tienden a ser diferentes.

Cada objeto es observado mediante un conjunto de variables cuantitativas que reflejan las cualidades fundamentales del mismo. Cada objeto tiene asociado entonces un conjunto de valores sobre un conjunto de p variables, que en lo sucesivo se llamará una observación. El conjunto de observaciones se agrupa en una matriz X de dimensión $(n \times p)$.

Luego, el proceso de clasificar, que se lleva a cabo sobre la matriz X, consiste en: dado un conjunto de n observaciones y sus características dadas por p variables, se requiere agruparlos basándose en las semejanzas que existan entre sí.

Metodologías para abordar la clasificación

Las métodologías de clasificación provienen fundamentalmente de dos fuentes: El análisis estadístico multivariado y el área de la inteligencia artificial llamada computación emergente. Los métodos pueden organizarse así:

- Análisis estadístico multivariado
  - Análisis de conglomerados (cluster)
  - Análisis discriminante
- Computación emergente

- Redes neuronales
  - Perceptrón multicapa
  - Mapas auto-organizativos
- Lógica difusa

Gran parte de la teoría estadística del análisis multivariado, que constituye el núcleo de los procesos clasificatorios fue desarrollada en la primera mitad de este siglo. Sin embargo, dadas las dificultades de cálculo, sólo podían abordarse pequeños problemas: limitados tanto en el número de observaciones como en el de variables que caracterizaban a los objetos. Los algoritmos de computación emergente, que no exigen conocimiento previo del tipo de distribución de probabilidad, han probado ser muy eficientes para abordar problemas de data compleja.

En las ultimas décadas los algoritmos de clasificación se implementan eficientemente sobre un computador y proveen los resultados sin intervención humana. Sin embargo, en la mayoría de las aplicaciones tecnológicas, el procesamiento obtenido es sólo un instrumento de soporte en la toma de decisiones, y es el usuario que conduce el proceso de Data Mining quien deberá decidir, por ejemplo: ¿en cuántas categorías clasificar la población de objetos?; ¿debe la clasificación ser jerárquica?; ¿un objeto alejado estadísticamente de las clases existentes es el anuncio del descubrimiento de una nueva clase o debe forzársele a pertenecer a una de las clases existentes?

1.1.1 Extractores de características.

Si bien la capacidad de cálculo de los actuales computadores permite resolver eficientemente gran parte de los problemas de clasificación no es menos cierto que cada vez la complejidad de los problemas de clasificación va en aumento: tanto en el número n de observaciones a clasificar como en la dimensión p del espacio de variables que definen el objeto.

El síndrome de la dimensionalidad.

La mayoría de los algoritmos clasificatorios padecen del síndrome de la dimensionalidad: probada eficiencia para problemas de dimensión reducida pero se vuelven ineficientes en problemas de gran escala. Es así que para espacios donde la dimensión p es excesiva se vuelve indispensable reducir la dimensionalidad del mismo. Los procedimientos que llevan a cabo esa función se denominan extractores de características.

Propósito

El objetivo fundamental de un extractor de características en procesos de clasificación es encontrar una transformación desde el espacio de dimensión p de las variables asociadas a cada observación en un espacio de dimensión inferior, denominado espacio de las características, que retenga de cada observación lo esencial de la información necesaria para el proceso de clasificación. Más precisamente: que el proceso clasificador de las observaciones en el espacio de la totalidad de las variables y en el espacio de las características conduzca a una

división de las observaciones en las mismas clases o con diferencias insignificantes.

Obviamente, la terminología de espacio de las características obedece a que de las numerosas variables que representan la observación se extraen las características esenciales de las mismas.

Existen tres razones principales para aplicar un extractor de características. La primera, la complejidad computacional de los algoritmos de clasificación se reduce sensiblemente al trabajar sobre un espacio de dimensión inferior. La segunda, los métodos estadísticos de estimación se vuelven más confiables en un espacio de dimensión reducida. La tercera, la posibilidad de que la dimensión del espacio de las características no exceda de tres, para permitir una visualización gráfica de las clases en juego.

Los métodos para extraer características son los vistos en reducción de la dimensionalidad y otros que se presentarán en el capítulo siguiente.

1.1.2 Tipos de clasificación.

Existe una división primaria en el concepto de clasificar:

- clasificación supervisada
- clasificación no supervisada.

La diferencia fundamental entre ambos métodos estriba en si se conoce o no la clase a la cual pertenece cada patrón (observación) de la data.

A continuación se aclararán estos conceptos.

La clasificación es supervisada si ya existe un conjunto de observaciones clasificadas en un conjunto de clases dado, y se conoce la clase a la que cada observación pertenece.

En la clasificación supervisada se distinguen dos fases fundamentales bien diferenciadas: la primera, consiste en el desarrollo o creación de una o varias regla de decisión (diseño del clasificador), y la segunda, el proceso en sí de clasificación de nuevas observaciones.

En la primera fase, el conjunto cuyas clases ya están bien definidas se desglosa en un conjunto de entrenamiento y otro de validación. Se diseña el clasificador con el conjunto de entrenamiento y se observa su capacidad para clasificar con el conjunto de validación. En la segunda fase se procede a clasificar nuevas observaciones de las que se desconoce la clase a la que pertenecen.

La clasificación es no supervisada cuando se dispone de un conjunto de objetos (observaciones), donde se desconoce tanto el número de clases en que es razonable particionarlo así como a qué clase pertenece cada observación.

Este proceso de clasificación no supervisada, es significativamente más complejo que el de la supervisada ya que se desconocen las clases naturales, y dependerá de la habilidad para seleccionar:

- las características que representan al objeto (elección de las variables que constituyen una observación)

- la metodología de clasificación

1.2 Clasificación no supervisada

Definición

Este proceso de clasificación consiste en: agrupar un conjunto de n objetos, definidos por p variables, en c clases, donde en cada clase los elementos posean características afines y sean más similares entre sí que respecto a elementos pertenecientes a otras clases.

La similaridad entre observaciones se establece en términos de distancias tal como se expondrá en esta sección. El número, c, de clases puede estar preestablecido o no, y depende del método elejido.

Varios son los própositos que pueden conducir a este tipo de clasificación:

- Graficar grupos afines, como es el caso de los dendrogramas de las taxonomías.
- Clasificar, simplemente, información abundante y compleja
- Hallar el número c de clases adecuado
- Encontrar subclases dentro de clases naturales
- Conceptualizar, interpretar los patrones analizando las causas intrínsecas de la formación de los mismos
- Hallar clases ocultas no previstas
- Preprocesar datos complejos con la finalidad de reducir la información a la aportada por los centros de las clases, para posteriormente realizar otros análisis con esta información simplificada, (Caras de Chernoff, por ejemplo).

Análisis de conglomerados

Los métodos de clasificación no supervisada tradicionales en estadística se encuentran en la literatura como análisis de conglomerados (Cluster analysis).

1.2.1 Formulación matemática

Toda la información disponible reside en la matriz X de dimensión $(n \times p)$ de las observaciones. Esta información puede ser relevante tanto en el espacio de dimensión p, R^p de las variables, (espacio de los vectores filas), así como en el de dimensión n , R^n, de las observaciones (vectores columnas).

Por ejemplo, la distancia entre dos observaciones (que es un concepto fundamental en clasificación no supervisada), se refiere a la distancia entre vectores fila y se determina en el espacio de dimensión p. Por el contrario, la correlación muestral entre las p variables -decisivo para la extracción de características, ya

que dos variables altamente correlacionadas contienen casi la misma información y puede eliminarse una de ellas- se determina en el espacio de dimensión n de las columnas.

Se supondrá en lo que sigue que cada observación es un punto del espacio de dimensión p de las variables y que dado dos observaciones, X_i y X_j, la distancia entre ambas es la distancia Euclídea habitual:

$$dist(X_i, X_j) = \|X_i - X_j\|$$

Objetivos duales en la clasificación

Existen dos objetivos duales en el proceso de obtener una clasificación óptima:

- Minimizar las desviaciones entre las observaciones que pertenecen al mismo grupo
- Maximizar las distancias entre los centros de los grupos

Se formulan como duales ya que es posible probar que basta lograr uno de los objetivos para que simultáneamente se logre el otro.

Uno de los objetivos del Análisis de conglomerados es hacer que la dispersión sea mínima, pero de nada sirve que una clase este muy concentrada en torno a su centro y otras no. Es por ello que es necesario establecer una medida de concentración global de la totalidad de las clases:

Definición

Se llamará SW_j, dispersión en la clase j, de N_j elementos, a la suma de las distancias al cuadrado de cada observación X_i al centro m_j de la clase (j) que la contiene:

$$SW_j = \sum_{i=1}^{N_j} \|X_i - m_j\|^2 \tag{1.1}$$

Si m es el centro de la data, la dispersión total de la data está dada por:

$$ST = \sum_{i=1}^{N} \|X_i - m\|^2 \tag{1.2}$$

Uno de los objetivos duales es hacer que la dispersión sea mínima, pero de nada sirve que una clase este muy concentrada en torno a su centro y otras no. Es por ello que es necesario establecer una medida de concentración global de la totalidad de las clases, PW , que es.la suma de las dispersiones, SW_j, de cada clase. Es esta medida la que debemos minimizar para optimizar la clasificación. Puesto que ST es constante, ya que sólo depende de la muestra, cada clasificación podrá ser valorada comparando su dispersión PW asociada con el valor de ST. Si el proceso de aglomeración ha sido eficaz, PW debe ser considerablemente inferior a ST. Es posible redefinir ahora el objetivo de la clasificación

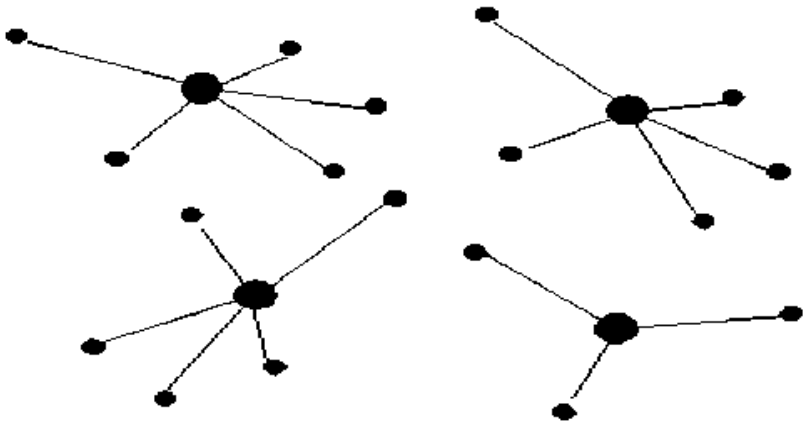

Figura 1.1: Dispersiones de las clases

no supervisada: fijado el número c de clases, distribuir las observaciones en c clases de modo de minimizar:

$$\min PW = \sum_{j=1}^{c} SW_j \tag{1.3}$$

Se afirmó ya que una buena clasificación exige clases muy concentradas, así mismo otro objetivo es lograr que las clases estén bien separadas. Se establecerá a continuación una medida de la separación entre las clases, la idea más simple para definir esta separación es medir la suma de las distancias al cuadrado entre los centros de cada cluster (m_j) y el centro general (m), ponderada cada distancia por el número de elementos de cada cluster. Esto es equivalente a calcular la dispersión de un cluster que este constituido por el mismo número de elementos de la población pero donde en cada centro m_j del cluster W_j, se encuentran superpuestas todas las observaciones X_i pertenecientes a ese cluster W_j.

$$SB = \sum_{j=1}^{c} N_j \left\| m_j - m \right\|^2$$

El teorema fundamental de descomposición

Este teorema, que no se demostrará aquí, vincula la dispersión dentro de las clases (PW) con la separación entre clases SB y prueba que independientemente del número de clases elegidas y de la constitución de las clases, la suma de PW y SB es constante y es igual a la dispersión total ST.

El Teorema de descomposición queda expresado como:

$$ST = PW + SB$$

La deteminación de una clasificación no es nada trivial y es necesario complejos algoritmos para hallar simplemente soluciones satisfactorias.

Para tener una medida de bondad de ajuste de la clasificación que sea comparable con otras clasificaciones, se propone, el indicador:

$$R^2 = 1 - \frac{PW}{ST} \tag{1.4}$$

El indicador, que es análogo con el de los modelos lineales, cumple: $0 \leq R^2 \leq 1$. Si la clasificación es adecuada, PW, debe ser pequeño. De modo que cuanto mayor sea R^2 mejor es la clasificación.

1.2.2 ¿Cuántas clasificaciones son posibles?

Anteriormente se expresó la dificultad de encontrar un algoritmo óptimo ¿a qué es debida esta dificultad, si para una partición dada el cálculo de PW es tan simple?. La razón es que el cálculo exhaustivo de PW para cada una de las particiones posibles se torna inabordable, aún para los computadores más eficientes del mundo, cuando inclusive los valores de n y c son relativamente pequeños.

El número de particiones de un conjunto de n elementos en c clases está dado por los números de *Stirling de segunda clase* S_c^n que están dados por la siguiente relación de recurrencia:

$$S_c^{n+1} = kS_c^n + S_{c+1}^n \qquad S_n^n = 1 \qquad S_1^n = 1$$

Por ejemplo:

n	c	particiones
8	3	966
12	4	611.501
15	4	42.355.950
20	5	749.206.090.500

Vista la dificultad intrínseca de la optimización, se han propuesto diversos métodos para obtener soluciones razonables.

El *análisis de conglomerados*, de extenso uso durante el siglo pasado en problemas de reconocimiento de patrones admite una clasificación primaria en:

- Métodos que clasifican a partir de la matriz de distancia entre todas las observaciones de la data. Entre estos métodos se cuentan los *jerárquicos*.
- Métodos *directos*, que sólo calculan distancias de las observaciones a posibles centros de las clases para luego modificar estos últimos sin necesidad de usar, las distancias entre las observaciones.

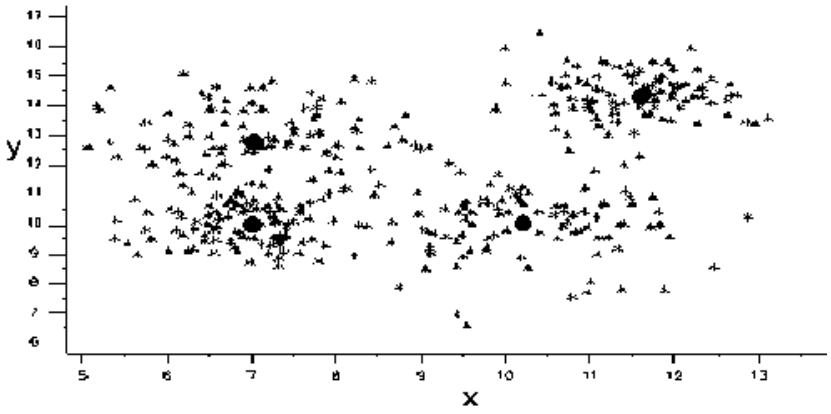

Figura 1.2: K-means: centros finales

1.3 Métodos directos

Los métodos directos se caracterizan por:

- Son iterativos
- Calculan las distancias de las observaciones a posibles centros de las clases, para luego modificar estos últimos siguiendo el criterio de optimización
- No hacen uso de las distancias entre los elementos
- El número de clases se fija de antemano

Usados principalmente cuando n es grande (n >5000, por ejemplo).
Algoritmo K-means
El método directo más ampliamente usado es el algoritmo iterativo de evolución de los centros (k-means). Consta de las siguientes etapas:

1. Ubicación tentativa de los centros iniciales de las clases
2. Asignación de las observaciones a la clase más cercana
3. Determinación de los nuevos centros de las clases
4. Verificar si se cumple alguno de los criterios de finalización del algoritmo. En el caso de no satisfacerse el criterio de convergencia se vuelve a la etapa 2.

Nótese que si las observaciones son 10.000 y las clases 10 en cada iteración se determinan 100.000 distancias; sin embargo, en un método jerárquico, donde se toman las distancias entre todas las observaciones, sería necesario calcular 5000x9999 distancias!!

La figura 1.2 muestra los centros finales de una clasificación en 4 clases, de una data de 400 observaciones de dos variables, obtenida con el algoritmo *FASTCLUS* del paquete estadístico SAS. Se observa en la nube de puntos que una clasificación en tres clases podría ser también válida.

1.3.1 Ubicación de los centros iniciales

El procedimiento *FASTCLUS* del SAS difiere de otros métodos directos en la importancia que le otorga a la etapa de ubicar los centros iniciales. Salvo que el usuario desee simplificar la etapa inicial haciendo uso de la opción REPLACE. El algoritmo debido a Anderberg (1973) para asignar los centros determina, siempre, excelentes resultados iniciales, reduciendo así el número de iteraciones de la etapa de modificación de los centros.

Este algoritmo tiene la siguiente propiedad:
si existe un número m de clases "naturales" lo suficientemente separadas como para que: la distancia máxima entre elementos de una misma clase sea menor que la distancia mínima entre elementos de clases distintas, entonces proporcionándole al procedimiento el número m, el algoritmo garantiza que hallará estas clases en la etapa inicial, sin necesidad de realizar iteración alguna de los centros.

Los centros iniciales son elegidos entre la totalidad de las observaciones luego de analizar de manera secuencial, observación por observación, si éstas califican para ser centros. Para explicar como procede el algoritmo para determinar los m centros iniciales se necesitarán las siguientes definiciones:

dmi= mínima distancia entre los centros.

C^j = centro más cercano a la observación X_j .

D_j =distancia entre C^j y X_j .

Las primeras m observaciones se constituyen en los centros iniciales. Luego, las observaciones restantes se analizan de a una y se convertirán en centros sustituyendo a uno de los centros existentes si satisfacen las siguientes pruebas.

Considérese la observación X_j:

Prueba 1. si D_j es mayor que dmi, X_j pasa a ser centro y sustituirá a uno de los dos centros cuya distancia entre ellos es dmi. ¿A cuál?. A aquel que si se le sustituye deja una mayor dmi. Obsérvese que luego de esta sustitución dmi aumenta. A continuación se considerará una nueva observación. Si esta prueba no se satisface se somete a la observación X j a la prueba 2.

Prueba 2. Si la mínima distancia de X_j a los restantes centros, (esto es excluyendo al centro C^j), es mayor que la de C^j a los restantes centros entonces X_j es un nuevo centro y se descarta a C^j. En caso contrario se descarta X_j y se considerará una nueva observación.

Si la muestra es:

OBS	X	Y
1	4	7
2	2	5
3	7	2
4	6	1

5	1	4
6	5	6
7	8	6
8	3	8

y se consideran 3 clases el algoritmo seleccionará las observaciones 6, 5 y 3.

Algoritmos iterativos que modifican los centros

Luego de explicar detalladamente la importancia de la etapa inicial, se tratará ahora el proceso iterativo de modificación de los centros. Fijados los centros iniciales existen dos algoritmos bien definidos:

1. Clasificar todas las observaciones asignándolas a sus centros más cercanos, para luego hallar el centro de cada clase y a partir de estos nuevos centros volver a iterar. Nótese que cada modificación de centros es posterior a la asignación de todas las observaciones.

2. Cada vez que una observación es trasladada de una clase a otra se vuelve a calcular los nuevos centros de estas dos clases, (la que pierde y la que gana una observación), que han sido modificadas. Luego, si se tienen n observaciones, este algoritmo implica calcular un máximo de $2n$ centros por iteración. Se ha indicado un máximo ya que es posible que al considerar un nuevo elemento este siga perteneciendo a la misma clase y no sea necesario recalcular los centros.

El primer algoritmo es el definido en el FASTCLUS por defecto. Suele ser más rápido y produce resultados razonables.

Nótese que en el primer algoritmo, ubicados ya los centros iniciales, el orden de las observaciones no altera el resultado. Sin embargo, el segundo es totalmente dependiente del orden en que se presenten las observaciones.

Sugerencia: presente el conjunto de observaciones en distinto orden ("barájelas") y vuelva a correr el algoritmo desde los mismos centros iniciales si desea analizar distintas soluciones.

1.3.2 Determinación del número óptimo de clases

¿Cómo saber cuántas clases deben hallarse? Un criterio similar al existente para modelos lineales es estudiar el indicador R^2, (ecuación 1.4) en función del número de clases. Es obvio que R^2 crece al aumentar el número de clases y que si $c = n$ entonces $R^2 = 1$, luego no se busca el máximo de R^2 sino analizar su tasa de cambio.

En la figura 1.3 se analiza la evolución de R^2, en el proceso de clasificar la misma data vista en la figura 1.2. Se observa que incrementar el número de clases más allá de 4 no redunda en una disminución significativa de la dispersión. Es cuatro (4) el número óptimo de clases.

Otro estadístico de ámplio uso es el *Seudo-F* (con una estructura similar al F de Fischer de la comparación de varianzas). A partir del cociente de la suma de

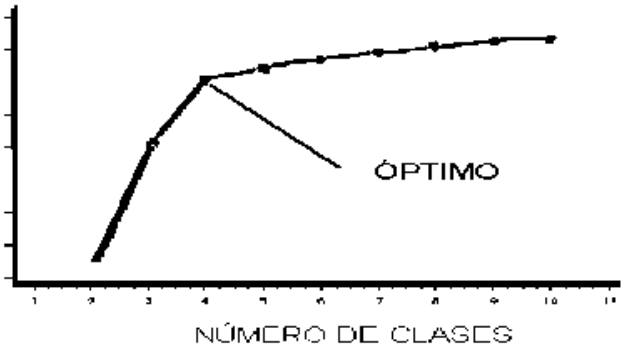

Figura 1.3: Evolución de R-cuadrado

cuadrados entre clases (SB) y la suma de cuadrados dentro de las clases (PW), se define el estadístico $SeudoF$ dividiendo las respectivas sumas de cuadrados entre sus "grados de libertad".

$$SeudoF = \frac{\frac{SB}{c-1}}{\frac{PW}{n-c}}$$

Se desea maximizar $SeudoF$. Nótese que se penaliza el exceso en el número de clases.

Como es obvio, tanto la curva de $SeudoF$ como la de R^2 dependen de las mejores soluciones obtenidas para cada c, y es razonableble preguntarse como obtener de manera eficiente estos estadísticos.

Estrategia de Kendall modificada

Los métodos directos se caracterizan por clasificar en un número fijo, predeterminado de clases. Pero surge la pregunta ¿cuál es el número adecuado de clases para un conjunto de observaciones dadas? Se propone aquí la siguiente estrategia que es una modificación de la propuesta por Kendall, (1980):
1.- Optar inicialmente por un número q de clases superior al número de clases que el usuario considere adecuadas.
2.- Aplicar el algoritmo FASTCLUS y obtener los q centros y los estadístico $SeudoF$ y R^2.
3.- Reducir q en una unidad ($q = q - 1$) y reunir las dos clases que al unirlas produzca un mínimo aumento del valor de PW y determinar el nuevo centro de dichas clases.
4.- Volver a la parte 2.

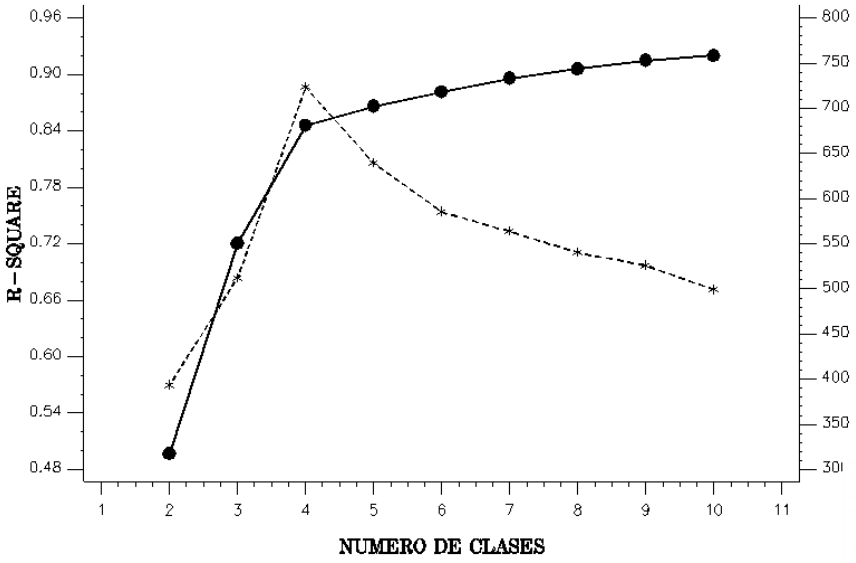

Figura 1.4: Determinación del número de clases

Una vez terminado el proceso, realizar un estudio comparativo de los estadísticos de bondad de ajuste: $SeudoF$ y R^2 en función de q, graficando ambas funciones, para determinar el número óptimo de clases (Figura 1.4).

1.3.3 Clases más cercanas

El algoritmo anterior propone unir las clases más cercanas, entendiéndose por más cercanas aquellas que al unirlas el increnento de PW sea mínimo. Si se aplica para dos clases de centros m_i y m_j con N_iy N_j elementos respectivamente el teorema de descomposición ($ST = SB + PW$). Luego la pérdida por unir dos clases está dada por SB. Para el caso de dos clases SB se expresa como $B_{i\ j}$ indicando las clases que se unen por los subíndices i, j. La media total de las clases, es en este caso:

$$m = \frac{(\ m_i\ N_i\ + m_{jj}\ )}{(i\ +\ Nj\ )}$$

Como se recordará esta varianza SB entre clases está dada por:

$$B_{i\ j} = N_i\ \|\ m_i\ - m\ \|^2 + N_j\ \|\ m_j\ - m\ \|^2$$

Sustituyendo el valor de m en la expresión anterior y luego de algunos cálculos queda:

$$B_{i\ j} = \frac{\parallel\ m_i\ -\ m_j\ \parallel^2}{(\ \frac{1}{N_i}\ +\ \frac{1}{N_i}\ )}$$

Obsérvese que para determinar las clases que producen un menor aumento de PW, o lo que es lo mismo una pérdida mínima en R^2, se debe tener en cuenta no sólo la distancia entre los centros sino también el número de elementos en cada clase. Privilegiándose la unión de clases con pocos elementos como se observa en el denominador de $B_{i\ j}$.

1.4 Jerarquias

Una clasificación jerárquica parte de un conjunto Ω cuyos elementos deben ser clasificados. Se trata de obtener sucesivas particiones, organizadas en diferentes niveles jerárquicos, formadas por clases disjuntas.

Una jerarquía H sobre el conjunto Ω está formada por un subconjunto H del conjunto de todas las partes de Ω, que cumple con dos axiomas:

1. Axioma de la intersección
   Dados dos elementos de H, estos son disjuntos o uno de ellos está contenido en el otro.

2. Axioma de la reunión.
   Todo elemento de H es el resultado de la reunión de los elementos de H que contiene o bien no contiene ningún elemento de H.

Ambos axiomas reflejan la noción de que dos géneros deben ser siempre disjuntos y que todo género es la unión de las especies que lo constituyen.
La jerarquía es total si cada elemento, (individuo), de Ω está en H y si el propio Ω está en H. Los elementos de H se llaman: clases, conglomerados o *clusters*.

En general las clasificaciones jerárquicas son totales ya que queremos que la jerarquía parta desde los individuos y vaya creciendo agrupando especies similares hasta llegar al conjunto total.

Por ejemplo sí $\Omega = \{a, b, c\}$ entonces, el conjunto de todas sus partes está formado por los 8 conjuntos: $\{(a), (b), (c), (a, b), (a, c), (b, c), (a, b, c), \emptyset\}$
$H = \{(a), (b), (c), (a, b), (a, b, c)\}$ es una jerarquía (formada por 5 clases). Compruebe que satisface los axiomas. Obsérvese que se unen primeramente los individuos a y b para formar una clase y luego esta clase se une con c para llegar al conjunto total.

En cambio $F = \{(a), (b), (c), (a, b), (b, c), (a, b, c)\}$ no lo es. Las clases (a,b) y (b,c) no cumplen con el axioma 1, tienen intersección (b) no vacía y sin embargo no está una dentro de la otra.

1.4.1 Jerarquías indexadas

Definición

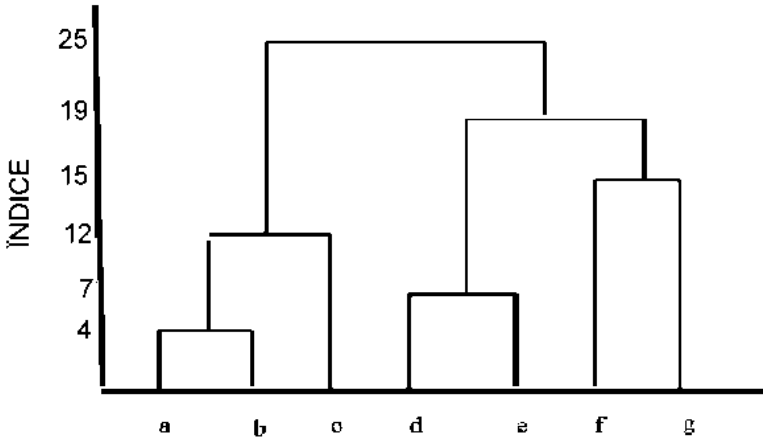

Figura 1.5: Jerarquía indexada

Una función, **ind**, es un índice sobre la jerarquía H si satisface estas dos condiciones:

1. $ind(\{a\}) = 0$ para todo elemento a $\in \Omega$.
2. dadas dos clases x e y de la jerarquía, si $x \subset y$ entonces $ind(x) < ind(y)$.

Definición
H es una *jerarquía indexada* si existe un índice sobre H.

La figura 1.5 muestra una jerarquía indexada sobre el conjunto $\Omega = \{a, b, c, d, e, f, g\}$. Sobre dicho conjunto pueden establecerse $2^7 = 128$ partes. De estas 128 sólo unas pocas forman la jerarquía H. El tipo de representación de una jerarquía expuesto en dicha figura es un *dendrograma,* el cual es un diagrama en forma de árbol que tiene asociado una escala donde el nivel 0 representa los individuos y el nivel más alto la raíz.. Como se ilustra en la figura 1.5 las 14 clases de la jerarquía H son:
$H = \{a, b, c, d, e, f, g, (a, b), (d, e), (a, b, c), (f, g), (d, e), (d, e, f, g), (a, b, c, d, e, f, g)\}$

Cuando dos clases son disjuntas se dirá que son clases al mismo nivel, como (d,e) y (a, b, c) en el ejemplo anterior.

El índice de la jerarquía permite generar una distancia, pero antes se revisará el concepto de distancia.

1.4.2 Distancias

Cuando en un conjunto de individuos se tiene una medida, $dist(i, j)$, de la discrepancia entre los mismos, se dice de una manera genérica que se tiene una distancia o disimilaridad sobre dicho conjunto. Según sea el tipo de exigencia

que se tenga sobre esta medida se pueden clasificar distintos tipos de distancia. Algunas de las propiedades de $dist$ resultan naturales y evidentes, otras son más complejas.

I $dist(i,j) \geq 0$ Positividad
II $dist(i,i) = 0$
III $dist(i,j) = dist(j,i)$ Simetría
IV $dist(i,j) \leq dist(i,k) + dist(k,j)$ Desigualdad triangular
V $dist(i,j) = 0 \rightarrow i = j$
VI $dist(i,j) = \max\{dist(i,k), dist(k,j)\}$ desigualdad ultramétrica
VII $dist(i,j)$ es euclídea. Proviene de un espacio euclídeo

VI implica **IV** ya que si se consideran los tres elementos i, j, k como vértices de un triángulo la IV indica que un lado es menor que la suma de los otros dos, y la VI es más fuerte ya que implica que un lado es igual al mayor de los otros dos. La VI implica que todo triángulo es isósceles y que la base es el menor y los lados iguales son los mayores salvo que los tres sean iguales.

A partir de estas propiedades surgen algunas definiciones de uso frecuente.

DENOMINACIÓN	PROPIEDADES
Disimilaridad	I, II, III
Distancia métrica	I, II, III, IV, V
Distancia ultramétrica	I, II, III, VI
Distancia euclídea	I, II, III, V, VII

Distancia asociada a una jerarquía indexada

Para construir una distancia entre los elementos de Ω asociada a la jerarquía H se procederá de la siguiente manera:
dados dos elementos $i \in \Omega$ y $j \in \Omega$ entonces $dist(i,j) = ind(c)$, donde c es la menor clase de la jerarquía H que contiene a (i,j).

Aplicando la definición anterior a la jerarquía H del dendrograma anterior se tiene, por ejemplo, $dist(d,g) = 19$ ya que la menor clase que los contiene es (d,e,f,g) que está definida a una altura de 19 sobre el índice. Completando las distancias obtenemos la matriz de distancias siguiente:

DIST	a	b	c	d	e	f	g
a	0	4	12	25	25	25	25
b	4	0	12	25	25	25	25
c	12	12	0	25	25	25	25
d	25	25	25	0	7	19	19
e	25	25	25	7	0	19	19
f	25	25	25	19	19	0	15
g	25	25	25	19	19	15	0

Si el lector observa con detenimiento esta matriz comprobará que la distancia es ultramétrica. Por ejemplo si se toma el triangulo a c e, las longitudes de sus lados son ac =12, ae=25, ce=25.

Para cualquier valor del índice, $x > 0$ la relación $iR_x j \Leftrightarrow dist(i,j) \leq x$ es una relación de equivalencia e induce una partición sobre los elementos de Ω.

Las clases de esa partición son las clases especie, genero, familia etc. "Clusters" a ese nivel. Por ejemplo, si: $x = 13$ las clases son 4: $\{(a,b,c),(d,e),f,g\}$. Si $x = 16$ las clase son 3: $\{(a,b,c),(d,e),(f,g)\}$. Cuando la distancia crece el número de clases disminuye.

Es posible probar el siguiente teorema fundamental que relaciona las jerarquías y la ultramétrica:
TEOREMA Si $dist$ es una ultramétrica en Ω, se construye a partir de $dist$ una jerarquía indexada sobre Ω. Y recíprocamente, una jerarquía indexada sobre Ω define una ultramétrica sobre Ω.

Por supuesto que la distancia inducida por la jerarquía creada coincide con la distancia original si esta distancia ya es de por sí una ultramétrica.

1.4.3 Cómo crear una jerarquía a partir de una similaridad

La dificultad fundamental de una clasificación jerarquía es que dado el conjunto Ω las distancias $dist$ entre los elementos de Ω suelen ser, la mayoría de las veces, una simple similaridad sin necesidad de ser una ultramétrica. Luego, es necesario crear una distancia ultramétrica $distu$ con la cual sí se genere una jerarquía. Se desea que la distancia ultramétrica $distu$ sea lo más próxima a la distancia $dist$.

Existen diversos métodos para generar $distu$, que difieren en la forma de unir las clases de modo de conservar a $distu$ como una ultramétrica. Una propiedad deseable de esos métodos es la de poder calcular las distancias para un número dado de clases a partir de la matriz de disimilaridades anterior sin necesidad de volver a la matriz original. Más adelante se expondrán los métodos que tienen esta propiedad con detalle y se mostrarán otros que no la tienen.

Aquí se expondrá como ejemplo el método del promedio ponderado de las distancias entre los elementos de las clases a unir (average linkage). Sea la matriz de disimilaridades (sólo se incluye la parte inferior de la matriz debido a la simetría)

a	a	b	c	d	e
a	0				
b	4	0			
c	6	8	0		
d	**2**	10	4	0	
e	12	8	14	6	0

Se unen **a** con **d**.

	a d	b	c	e
a d	0			
b	7	0		
c	**5**	8	0	
e	9	8	14	0

Se unen **ad** con **c**.

Para calcular la distancia de la clase **adc** a **e**, por ejemplo debe ponderarse por el número de elementos que pertenecen a cada clase incluida en el cluster adc, es decir, **ad** tiene 2 elementos y c tiene uno solo. $distu(\mathbf{adc}, \mathbf{e}) = (2 * 9 + 14)/3$. Lo mismo con el resto:

	a d c	b	e
a d c	0		
b	**22/3**	0	
e	32/3	8	0

Luego de unir **adc** con **b**, ya que es la menor distancia de la tabla anterior, $distu(\mathbf{adcb}, \mathbf{e})$ es 10.

La tabla que sigue incluye las disimilaridades originales y en negrita las distancias que surgen de la jerarquía creada.

	a	b	c	d	e
a	0				
b	4 - **7.33**	0			
c	6 - **5**	8 - **7.33**	0		
d	2 - **2**	10 - **7.33**	4 - **5**	0	
e	12 - **10**	8 - **10**	14 - **10**	6 - **10**	0

considerando 3 elementos cualesquiera, el lector comprobará que la distancia es ultramétrica (el triángulo que se forma siempre es isósceles).

Como ya se mencionó, la bondad del método consiste en que la distancia ultramétrica *distu*, sea lo más próxima a la disimilaridad *dist*.

1.5 Métodos jerárquicos

La finalidad de los métodos jerárquicos consiste en agrupar clases para formar una nueva, de tal manera que se maximice alguna medida de similitud entre clases.

Las ciencias biológicas proporcionan los ejemplos clásicos de agrupación jerárquica, así los individuos se agrupan en especie, éstas en género, los géneros en familia, etc.

Los métodos jerárquicos se dividen en aglomerativos y divisivos,. Los procedimientos que utiliza el SAS en el PROC CLUSTER son del tipo aglomerativo.

Los métodos jerárquicos crean una jerarquía entre las clases que se construyen a partir de las observaciones.

Propósito*:*

Dado un conjunto inicial donde cada elemento es una clase, crear un arbol jerárquico agrupando en cada etapa las dos clases ubicadas a mínima distancia, ésta indica la altura sobre el árbol.

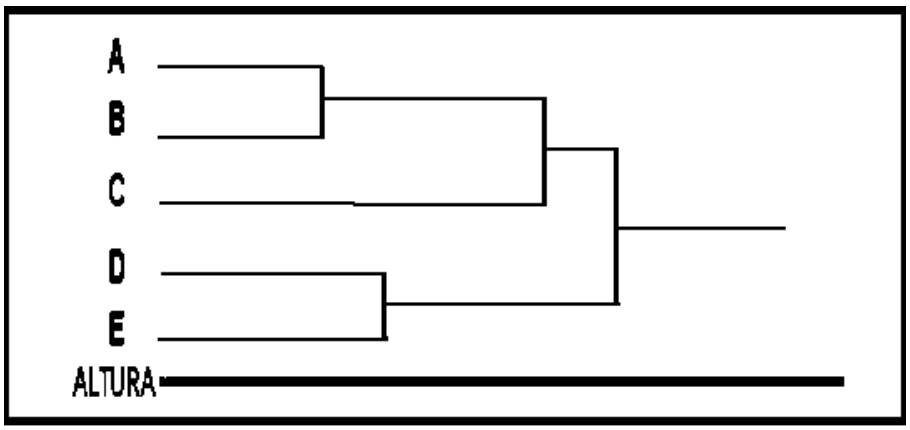

Figura 1.6: Clasificación jerárquica

Llámese n al número total de observaciones y c al número de clases formado. Inicialmente cada clase está constituida por una sola observación, partiéndose de tantas clases como observaciones ($c = n$); y a continuación se selecciona alguna medida de distancia entre clases agrupándose las dos clases con menor distancia. Se repite sucesivamente el procedimiento de unir las clases más cercanas hasta que se cumpla alguna de estas condiciones:

- Se forma un solo grupo $c = 1$ con todas las observaciones.
- Se alcanza un número prefijado C_f de clases.
- Se detecta a través de un test de significación que existen razones estadísticas que hace que el número de clases existente sea óptimo en algún sentido.

Los métodos jerárquicos se caracterizan por:

- Clasifican a partir de la matriz de distancia entre las observaciones
- Construyen una distancia entre clases
- No se fija el número de clases
- Se determina el número óptimo de clases a partir del árbol jerárquico
- Apropiados sólo si el tamaño del conjunto es pequeño, en cuyo caso son más eficicientes que los métodos directos

La posibilidad de comparar a partir de un dendrograma distintos números de clases y decidir cual se considera adecuado en lugar de trabajar con un número prefijado, o mejor aún poder detectar el número óptimo de clases apartir de

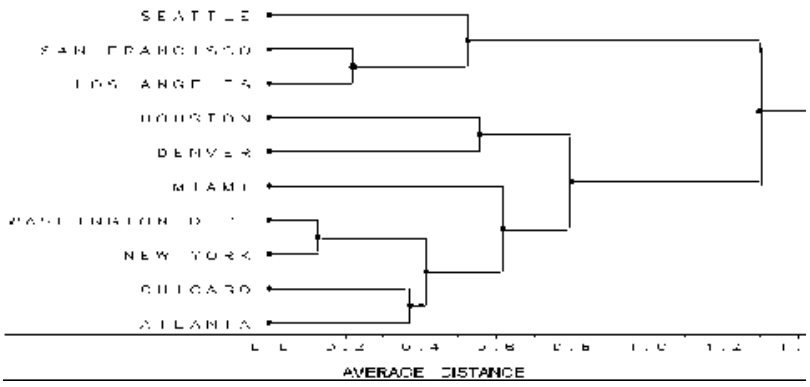

Figura 1.7: Arbol jerárquico

análisis estadísticos, le otorga a los métodos jerárquicos una evidente supremacía sobre los métodos directos.

Otra ventaja de los métodos jerárquicos es que en lugar de partir de las distancias de cada observación al centro asignado (como en los métodos directos), el análisis parte de una matriz de distancia entre todas las observaciones, que relaciona dos a dos todas las observaciones. sin embargo, esta virtud, puede convertirse en desventaja ya que significa un número de distancias a calcular y guardar en memoria muy superior a las distancias involucradas en un método directo.

La figura 1.7 es un arbol jerárquico de las distancias entre algunas ciudades en EEUU.

Puesto que es muy importante diferenciar los dos conceptos de distancia utilizados en la clasificación jerárquica: **entre elementos** y **entre clases.** se expondrán a continuación:

1.5.1 Distancias entre elementos

- Euclídea
- Estándar
- Mahalanobis

Las agrupaciones obtenidas usando la distancia euclídea son invariantes a traslaciones o rotaciones, pero no son invariantes a transformaciones lineales o en general a cambios de escala de las variables. Una forma de lograr esta invarianza consiste en normalizar los datos, tal como es habitual en estadística.

La distancia de Mahalanobis, incluye la información completa de la estructura de correlación; estimada su media, vector μ, y su matriz de covarianza Σ,

la distancia entre dos observaciones x e y, está dada por:

$$dist(x, y) = (x - y)^T \Sigma^{-1} (x - y). \tag{1.5}$$

1.5.2 Distancia entre clases

Distancia máxima. (Complete linkage)

Esta estrategia considera que la distancia entre dos grupos viene definida por la máxima distancia entre una observación de una clase y una observación de otra clase.

Distancia promedio ponderado. (Average linkage)

En esta estrategia se considera que la distancia entre dos grupos es el promedio de las distancias de todas las observaciones de un grupo respecto a todas las observaciones del otro. Debido a sus características ésta distancia es la más usada en las ciencias aplicadas.

Distancia prototipo. (Centroide)

Esta estrategia se diferencia de la anterior en que en esta se considera la distancia promedio (prototipo) de un grupo con respecto al prototipo de otro grupo, es decir, es la distancia entre los centros de ambos grupos.

Estrategia de la mínima varianza (Ward)

Este método une las clases tales que el aumento en PW (suma de la dispersión dentro de las clases) sea mínimo. Esto es equivalente a minimizar varianzas. Otra forma de expresarlo es que éste es el método que produce un mínimo deterioro en el estadístico R^2. Este método tiende a unir clases poco numerosas y a producir clases con un número similar de observaciones.

Es importante señalar, que a pesar de las diferencias que existen entre estas estrategias, si los grupos son compactos y bien definidos, todas ellas conducen a resultados prácticamente análogos.

La estrategia de la distancia máxima tiende a "alargar" los dendrogramas y a tomar en cuenta elementos alejados, no representativos, del grupo; mientras que la estrategia de la distancia mínima se va al otro extremo, pudiendo utilizar elementos fuera de grupo como puente de unión entre dos grupos, este efecto es denominado "efecto de cadena". Por estas razones el uso de estas dos estrategias ha ido decreciendo. La estrategia de la distancia promedio ponderado (average linkage) puede ser obtenida en muy poco tiempo de CPU. Esta cualidad aunado al hecho de no ser tan sensible a la presencia de elementos alejados hacen de este método el preferido en las aplicaciones. Al tener la tendencia de unir clases con varianza pequeña, éste método en consecuencia genera clases con varianza similar.

La aparente virtud del método del centroide consiste en no ser demasiado sensible a la presencia de observaciones alejadas, (nótese que cuatro puntos vértices de un cuadrado pueden tener el mismo centro, ya sea que el cuadrado tenga 1 o 100 observaciones por lado), pero a pesar de esto es inferior a los métodos de Ward y de promedio ponderado.

Es un principio básico de la estadística que un método es más eficiente en la medida que use la totalidad de la información muestral disponible. Los métodos

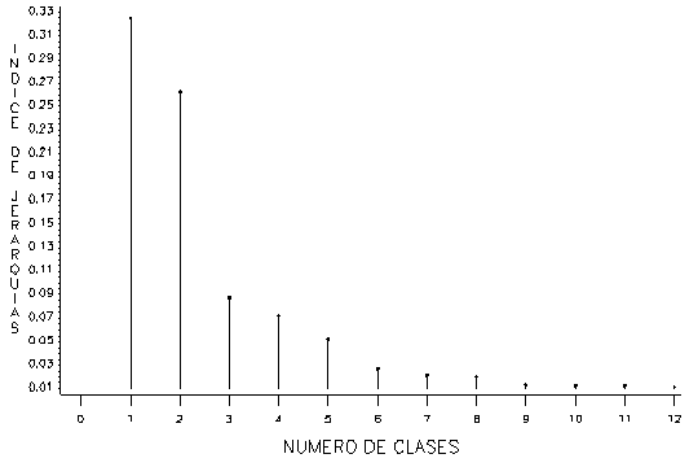

Figura 1.8: Indice de jerarquía vs número de clases

de la distancia mínima y máxima, al tomar en cuenta la relación entre dos elementos privilegiados de cada clase, desprecia el resto de la información existente en las clases. Es por ésta razón que el método del promedio ponderado, o el de Ward, que utilizan las distancias entre los elementos de ambas clases, son más eficientes.

1.5.3 Determinación del número óptimo de clases

La clasificación óptima se basa en el estudio de la evolución del índice de jerarquía. En rasgos generales se observa el incremento relativo en dicho índice al reducir las clases. Esto es, el "costo" en la pérdida de la dispersión dentro de las clases más cercanas como consecuencia de unirlas y convertirlas en una sola clase. Recuérdese que el propósito de cualquier técnica de agrupamiento es, precisamente, la obtención de grupos de forma tal que se maximice la divergencia entre ellos y se minimice la dispersión dentro de las clases.

La evolución del índice de jerarquía en función del número de clases se vuelca en un gráfico que permite visualizar adecuadamente el número óptimo de clases (ver figura 1.8).

Existen diversas estrategias para determinar la distancia entre las clases, para cada una de ellas existe un índice de jerarquía y en consecuencia el criterio sobre el número de clases óptimo, que depende de dicho índice, puede variar.

Es necesario pues un soporte estadístico que aporte nuevos elementos a la decisión. Los más relevantes son el R^2 y el Seudo-F. Al igual que con el índice de jerarquía se grafican los estadísticos R^2 y Seudo-F versus el número de clases.

Se ha insistido en la aplicación de varios métodos ya que el usuario determinará el número de clases óptimo luego de un exhaustivo análisis, usando diversos

índices de jerarquía así como de los estadísticos asociados.

1.6 Apéndices

Proc Fastclus del SAS

PROC FASTCLUS MAXCLUSTERS = n Sentencias requeridas
| RADIUS = t <opción>;
VAR variables;
ID variables;
FREQ variables; Sentencias opcionales
WEIGHT variables;
BY variables;

La instrucción PROC FASTCLUS inicia el procedimiento del FASTCLUS, y se debe especificar la instrucción MAXCLUSTER o la instrucción RADIUS. La sintaxis es la siguiente:

PROC FASTCLUS MAXCLUSTER = n | RADIUS = t <opción>;
MAXCLUSTER = n

Proc Cluster del SAS

El procedimiento CLUSTER produce cluster jerárquicos de observaciones en un conjunto de datos usando 11 métodos. La matriz de entrada puede ser de coordenadas numéricas o de distancias. CLUSTER produce un archivo de salida con el cual el procedimiento TREE puede dibujar un diagrama de árbol o una salida de cluster a un nivel especifico del árbol.

Los métodos de agrupamiento disponibles son: promedio ponderado (average linkage), del centroide (CENTROID), distancia máxima (COMPLETE), densidad (DENSITY, incluyendo los métodos Wong´s hybrid y Kth-nearest-neighbor), máxima probabilidad (EML), flexible-beta (FLEXIBLE), análisis de similaridad de McQuitty (MCQUITTY), mediana (MEDIAN), distancia mínima (SINGLE) mínima varianza (WARD).

Especificaciones del PROC CLUSTER

PROC CLUSTER METHOD = nombre <opciones>; sentencia requerida
BY variables;
COPY variables;
ID variable; sentencias opcionales
RMSSTD variable;
VAR variables;

Cálculo de distancias entre clases.

La distancia entre dos cluster puede ser definida directamente o por combinatoriedad, esto es, por una ecuación que actualice la matriz de distancia cuando dos cluster se unan. En todas las fórmulas combinatorias que siguen, se asume que los cluster C_ky C_L, de tamaño N_k y N_L, se unen para formar C_M, y la fórmula dá la distancia entre el nuevo cluster C_M y cualquier otro cluster C_j.

AVERAGE

En éste método la distancia entre dos cluster es la distancia promedio entre

pares de observaciones, una en cada cluster. Este, tiende a unir cluster con varianzas pequeñas y a producir cluster con varianza similar.

$$D_{(\,j\,M\,)} = \frac{(N_k\, D_{j\,K} + N_L\, D_{j\,L})}{N_M}$$

CENTROID
La distancia se define como el cuadrado de la distancia euclídea entre los centroides de los dos cluster a unir.

$$D_{(j\,M)} = \frac{(N_k\, D_{j\,K} + N_L\, D_{j\,L})}{N_M} - \frac{N_k\, N_L\, D_{k\,L}}{N_M^2}$$

COMPLETE
La distancia utilizada es la distancia máxima de una observación en un cluster a una observación en otro cluster. Este método tiende a producir cluster con igual diámetro y puede ser distorsionado severamente por outliers.

$$D_{(\,j\,M\,)} = \max(\,D_{j\,K}\,,\,D_{j\,L}\,)$$

SINGLE
Este método usa la distancia mínima que exista entre dos observaciones de clusters diferentes.

$$D_{(\,j\,M\,)} = \min(\,D_{j\,K}\,,\,D_{j\,L}\,)$$

WARD

$$D_{(\,j\,M\,)} = \frac{((N_j + N_k)D_{j\,K} + (N_j + N_L)\,D_{j\,L}) - N_j\,D_{k\,L})}{(N_j + N_M)}$$