Cluster

Eliseo Martínez H.

1. ¿Para qué sirven los Cluster o Conglomerados?

Por lo genera se dice que la finalidad de la formación de cluster o conglomerados es agrupar elementos en grupos homogéneos en función de "alguna similitud o similaridades" entre ellos. Como podemos ver esta finalidad parece autoreferente. Puesto que si tenemos elementos homogeneos ya por sí estarían formados los cluster. En rigor, quien realiza una formación de cluster lo hace mediante una técnica de homogeneidad, y posteriormente encontrado los cluster o conglomerados realiza el reconocimiento de patrones de formación. ¿ah, por esto y por esto otro estos grupos se formaron?. Tanto es así que a menudo al análisis de cluster se le llama *métodos de clasificación automática no supervisada*. En primer lugar ¿qué se clasifica o agrupa? Se pueden agrupar tanto las observaciones como las variables.

Las observaciones, o los individuos que responden (en términos matriciales, las filas) se pueden agrupar y de esta forma particionarlos en "patrones de respuestas". Además se pueden agrupar las variables, y de esta forma también estudiar los patrones de suimilaridades que existen entre las preguntas (en términos matriciales, las columnas).

En segundo lugar ¿cómo se clásifica? El método de clásificación deberá obedecer a tres postulados básicos, ya sea que se clasifiquen las observaciones o las variables, y estos son

1 cada elemento (observación o variable) pertenece a uno y sólo uno de los grupos;
2 todo elemento queda clasificado, y;
3 cada grupo sea internamente homogéneo

Como podemos observar, el tercer punto parece indicar el cómo se clasifica, puesto que debemos espicificar qué entendemos por grupo homogéneo. El primer punto, como veremos más adelante se va a violar en benficio de los dos puntos posteriores.

Podemos estructurar los grupos en *jerarquías*. Esto significa formar grupos jerarquicos por similitud, lo que implica que los datos se agrupan en niveles de jeraquía, de tal forma que los niveles superiores contengan a los niveles inferiores. El ejemplo paradigmático es la clasificación de animales y plantas, en general las especies, en las ciencias biológicas. De tal forma que no se puede decir que cada elemento pertenecerá a uno y solo un grupo, sino más bien se definen asociaciones de pertenencias en cadenas que permitan reconocer un patrón de ubicación de cada observación.

Podemos *clasificar o particionar las variables*. Esto significa que ante muchas variables podemos hacer una división entre ellas y formar grupos, de tal forma que nos permitirá hacer una reducción en la dimensión de ellas. De igual forma esta clasificación de variables se puede realizar buscando grupos entre ellas o jerarquizándolas.

Los métodos de partición o de formación de grupos utilizan directamente la matriz de

datos, y para la clasificación jerárquica se utiliza la matriz de distancia o de similitudes entre las observaciones (o las variables, según lo que se quiera clasificar). En la agrupación de variables se debe considerar el tipo de variable. Si las variables son continuas se utilizará la matriz de correlación, y si son discretas se utilizará la matriz de distancia ji-cuadrado.

2. Métodos de partición

2.1 Algoritmo de K-medias

Supongamos que tenemos nuestra matriz de datos $\mathbf{X} = (x_{i\,j})_{n\times p}$, donde tenemos n observaciones y p variables. Nuestro objetivo es particionar la muestra de tamaño n, esto es las observaciones, en K grupos. Elegimos estos K grupos en forma arbitraria o mediante algún criterio. De esta forma las observaciones x_{ij} las reescribimos como x_{ijk}, entendiendo que x_{ijk} es el dato x_{ij} que está asignado al grupo k, con $k = 1, ..., K$. Ahora definimos el valor de $\overline{x}_{jk}$ como el promedio de la j-ésima variable que pertenece al grupo k, esto es

$$\overline{x}_{jk} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ijk} \tag{1}$$

donde n_k es el tamaño del grupo k, donde es claro que $n_1+\cdots+n_K = n$. Calculamos ahora la suma de las distancias cuadráticas de todas las observaciones del grupo k correspondiente a la j-ésima columna respecto de la media dentro del grupo

$$\sum_{i=1}^{n_k}(x_{ijk} - \overline{x}_{jk})^2$$

Luego sumamos estas distancias cuadráticas para cada columna y luego para cada grupo, esto es

$$\sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{i=1}^{n_k}(x_{ijk} - \overline{x}_{jk})^2 \tag{2}$$

A esta suma la llamaremos *suma de cuadrados dentro del grupo*, $SCDG$, y sirve para medir la homogeneidad entre la partición de los K grupos elegidos. De tal forma que debemos efectuar una buena partición en K grupos si la suma $SCDG$ es mínima. De otra forma nos quedamos con aquella partición en que se obtenga el valor de

$$\min SCDG = \min \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{i=1}^{n_k}(x_{ijk} - \overline{x}_{jk})^2 \tag{3}$$

Podemos observar que $\sum_{i=1}^{n_k}(x_{ijk} - \overline{x}_{jk})^2$ es esencialmente la varianza de la j-ésima columna dentro del grupo k, en efecto si designamos tal varianza por s^2_{jk}, se tiene que

$$s^2_{jk} = \frac{1}{n_k}\sum_{i=1}^{n_k}(x_{ijk} - \overline{x}_{jk})^2$$

entonces (2) queda como

$$\sum_{k=1}^{K}\sum_{j=1}^{p} n_k\, s_{jk}^2$$

De tal forma que minimizar (2) es equivalente a

$$\min SCDG = \min \sum_{k=1}^{K}\sum_{j=1}^{p} n_k\, s_{jk}^2$$

Observe que la construcción de todos los valores de $SCDG$ se realiza para cada una de las particiones posibles de K grupos, lo cual si teóricamente es posible significa un esfuerzo de cómputo formidable. En efecto, considere $n = 130$ variables y $K = 3$. Y apenas estamos interesados en las particiones del tipo siguiente: que el primer grupo tenga 30 variables ($n_1 = 30$), que el segundo grupo tenga 30 variables ($n_2 = 30$), y el tercer grupo tenga las 70 variables restantes ($n_3 = 70$). Para esta particular condición hay

$$\frac{130!}{30!30!70!} = 76730\,48982\,46752\,01267\,36536\,46382\,21253\,74298\,47956\,86699\,64800$$

¡particiones posibles! Sin considerar las otras formas de particionar las observaciones en 3 grupos, como por ejemplo $n_1 = 1$, $n_2 = 1$ y $n_3 = 128$.

De manera que intentaremos buscar un camino más corto y eficiente. Supongamos que tenemos particionadas la muestra en K grupos. Para un grupo $k \in K$ calculamos el vector de medias $\overline{\mathbf{x}}_k$, es decir

$$\overline{\mathbf{x}}_k = \begin{pmatrix} \overline{x}_{1k} \\ \vdots \\ \overline{x}_{pk} \end{pmatrix}$$

Luego medimos las distancias cuadráticas entre los puntos de cada frupo con su media, esto es para un particular $k \in K$

$$\sum_{i=1}^{n_k} (\mathbf{x}_{ik} - \overline{\mathbf{x}}_k)^t\ (\mathbf{x}_{ik} - \overline{\mathbf{x}}_k)$$

donde el vector $\mathbf{x}_{ik}$ está constituido por la observación i-ésima que pertenece al grupo k [1]. Luego aplicamos este mismo cálculo para todos los grupos, y obtenemos el mínimo sobre todas las formas de particionar en K grupos las n onservaciones. Esto es

$$\min \sum_{k=1}^{K}\sum_{i=1}^{n_k} (\mathbf{x}_{ik} - \overline{\mathbf{x}}_k)^t\ (\mathbf{x}_{ik} - \overline{\mathbf{x}}_k) = \min \sum_{k=1}^{K}\sum_{i=1}^{n_k} d^2(i,k) \tag{4}$$

entendiendo que $d^2(i,k)$ es la distancia cuadrática entre el elemento i del grupo k y la media de dicho grupo. No es complicado probar que los criterios establecidos en (3) y (4) son equivalentes.

Notemos además que, por ser $d^2(i,k)$ un escalar entonces trivialmente se tiene que

[1] En rigor se debe reindiciar los elementos de las filas que conformen el grupo k, puesto que estas filas deberán tener sus índices entre 1 y n_k.

$traza\left[d^2(i,k\right] = d^2(i,k)$, de modo que

$$\min \sum_{k=1}^{K} \sum_{i=1}^{n_k} traza\left[d^2(i,k)\right] = \min\, traza \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{x}_{ik} - \overline{\mathbf{x}}_k)\ (\mathbf{x}_{ik} - \overline{\mathbf{x}}_k)^t$$

y si llamamos $\mathbf{W}$ a la matriz de la suma de cuadrados dentro de los grupos, esto es

$$\mathbf{W} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{x}_{ik} - \overline{\mathbf{x}}_k)\ (\mathbf{x}_{ik} - \overline{\mathbf{x}}_k)^t$$

tenemos que

$$\min\, traza\,(\mathbf{W}) = \min\, SCDG \tag{5}$$

El criterio (5) se conoce como el criterio de la traza.

Habíamos dicho que obtener tanto el valor de (3) como de (4) pasaba por realizar todas las particiones de K subconjuntos de las n observaciones, cuestión que ya vimos es practicamente imposible. Sin embargo se puede encontrar un algoritmo sub-óptimo en función de la traza de $\mathbf{W}$. Y este algoritmo es como sigue:

1 Partir de una asignación inicial, que eventualmente puede ser arbitraria
2 Comprobar si moviendo algún elemento se reduce la $traza(\mathbf{W})$
3 Si efectivamente se reduce la $traza(\mathbf{W})$, volver a calcular todas las medias de los dos grupos afectados por el cambio (en un grupo la variable que sale y el otro grupo donde recibe a la variable que salió). Luego volver al paso (2). Si no es posible reducir la $traza(\mathbf{W})$ terminar el proceso.

3. Métodos jerárquicos

3.1 Variables cuantitativas binarias: similaridad y distancias

Vamos a suponer que las variables tienen por valor {0, 1}, de tal forma que nuestra matriz de "datos" $\mathbf{X}$ es de la forma

$$\mathbf{X} = (x(i,j))\,;\ x(i,j) \in \{0,1\}; i = 1,...,n;\ j = 1,...,p$$

Nuestro problema es asignar adecuadamente una matriz de similaridad, y en consecuencia una matriz de distancias a las n observaciones muestrales que tenemos.

Recordemos que para cada variable $j = 1,...,p$ un coeficiente de similaridad sobre dos elementos muestrales i, h es una función $s_j(i,h)$, que satisface lo sigiente:

- $s_j(i,i) = 1$
- $0 \leq s_j(i,h) \leq 1$
- $s_j(i,h) = s_j(h,i)$

El problema se trata entonces de encontrar una buena función de similaridad entre las observaciones. En lo que sigue vamos a construir una matriz de similaridad bastante usual.

Definamos las siguientes expresiones

$$\begin{aligned} A(i,h) &= \sum_{j=1}^{p} x(i,j)\, x(h,j) \\ B(i,h) &= \sum_{j=1}^{p} x(i,j)\, (1 - x(h,j)) \\ C(i,h) &= \sum_{j=1}^{p} (1 - x(i,j))\, x(h,j) \\ D(i,j) &= \sum_{j=1}^{p} (1 - x(i,j))(1 - \ x(h,j)) \end{aligned}$$

Observemos que, en términos de nuestras variables binarias, $A(i,h)$ está denotando el número de coincidencias en los elementos i y h para los p atributos cuantitativos, en otras palabras es el número de atributos que están presente en i y en h; de manera análoga $B(i,h)$ es el núnero de atributos que están presentes en el elemento i y no está presente en el elemento h; $C(i,h)$ es el número de atributos presentes en el elemento i y que no están presentes en el elemento h; y finalmente $D(i,j)$ es el total de atributos que no están presentes ni en i ni en h.

A partir de estos valores existen dos criterios para definir función de similitud

1 *Proporción de coincidencias.* Se define la similitud entre dos elementos como el número total de coincidencias sobre el número de atributos totales p, esto es

$$s(i,h) = \frac{A(i,h) + D(i,h)}{p}$$

2 *Proporción de apariciones.* Cuando se quiere dar más preponderancia el hecho de que los atributos aparecen en ambos elementos y considerando que la ausencia de un atributo en común no es relevante, esto es

$$s_1(i,h) = \frac{A(i,h)}{A(i,h) + B(i,h) + C(i,h)}$$

La construcción de las matrices de similitud, en ambos casos, se muestra en el siguiente sencillo ejemplo. Supongamos que a cada uno de tres elementos le hemos medido si posee o no un atributo de un total de 7; y la matriz de información es como sigue:

elementos	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	0	1	1	0	0	0	1
2	1	0	1	1	1	1	0
3	1	0	0	1	1	1	1

Utilizando un sencillo programa[2] realizado en el software DERIVE, tenemos que para el

[2] Puede descargarlo en el Internet
http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/similaridad.dfw

primer caso la matriz de similitud es

$$\mathbf{S} = \begin{pmatrix} 1 & 1/7 & 1/7 \\ 1/7 & 1 & 5/7 \\ 1/7 & 5/7 & 1 \end{pmatrix}$$

Y para el segundo caso es

$$\mathbf{S}_1 = \begin{pmatrix} 1 & 1/7 & 1/7 \\ 1/7 & 1 & 2/3 \\ 1/7 & 2/3 & 1 \end{pmatrix}$$

En cualquier caso una vez obtenida la matriz de similaridad, se define una matriz de distancias asociada a tal similaridad mediante

$$d(i,h) = \sqrt{2(1 - s(i,h)}$$

Con esta matriz $\mathbf{D} = (d(i,h))$ podemos realizar análisis de conglomerado por métodos jerárquicos. La matriz distancia asociada a la similaridad $\mathbf{S}$ anterior es,

$$\mathbf{D} = \begin{pmatrix} 0 & 1.309307341 & 1.309307341 \\ 1.309307341 & 0 & 0.7559289460 \\ 1.309307341 & 0.7559289460 & 0 \end{pmatrix}$$

y la matriz de distancia para la similaridad $\mathbf{S}_1$anterior está dada por

$$\mathbf{D}_1 = \begin{pmatrix} 0 & 1.309307341 & 1.309307341 \\ 1.309307341 & 0 & 0.8164965809 \\ 1.309307341 & 0.8164965809 & 0 \end{pmatrix}$$

3.2 Variables cuantitativas continuas: similaridad y distancias

Si las variables son de tipo continuo, entonces para la $j-$ésima variable ($j-$ésima columna) se define una función de similaridad entre dos elementos de observaciones (filas) como

$$s_j(i,h) = 1 - \frac{|x(i,j) - x(h,j)|}{rango\,(\mathbf{x}_j)}$$

donde

$$rango\,(\mathbf{x}_j) = \max_{1 \le i \le n} \{x(i,j)\} - \min_{1 \le i \le n} \{x(i,j)\}$$

Ahora para definir los coeficientes de similaridad global lo hacemos mediante

$$s(i,h) = \frac{\sum_{j=1}^{p} w_{jih}\, s_j(i,h)}{\sum_{j=1}^{p} w_{jih}}$$

donde w_{jih} son pesos binarios, donde tendrá el valor de 1 si la comparación entre los elementos i y h tiene sentido, y será 0 si no queremos incluir esa variable en la comparación[3].

[3] Esta pondración se aplica en rigor a las variables cualitativas, por ejemplo, si la variable x_1 indica si una persona ha o no ha pedido crédito (1 si lo ha pedido, 0 si no lo ha pedido), y x_2 si la persona ha pagado o no la deuda del crédito, entonces una persona que no ha pedido crédito, esto es $x_1 = 0$ entonces no tiene sentido compararlo con el valor de x_2. De otra forma si para dos elementos muestrales hay al menos un 0 en el valor de x_1 no tiene sentido compararlo con la respuesta de x_2, y en ese caso $w_{jih} = 0$

En particular, si todas las variables son comparables, entonces

$$s(i,h) = \frac{\sum_{j=1}^{p} s_j(i,h)}{n}$$

Una vez obtenida esta función o matriz de similaridad, definimos la función distancia de la manera habitual, esto es

$$d(i,h) = \sqrt{2(1-s(i,h)}$$

Ejemplo. Consideremos la matriz de datos correspondiente a la frecuencia de las letras en 5 libros de escritores iberoamericanos, que vimos en la sección de análisis de textos literarios, y cuya base de datos se puede obtener en el Internet[4]. Vamos a considerar a las 27 variables como la frecuencia relativa de aparición en cada uno de los cinco libros, es decir $p = 27$ y $n = 5$. Puesto que son frecuencias (no tienen unidad) no hay problema de escalamiento, y vamos a considerar los datos sin estandarizar. Para el cálculo de la matriz de similaridad y su matriz de distancia asociada utilizamos un pequeño programita en el software DERIVE llamado **continua.dfw**[5]. La ejecución de este ejemplo la puede ver en el Internet, en la dirección[6]

3.3 Algoritmos jerárquicos

El motor esencial es que los elementos de la muesta son asignados sucesivamente a los grupos, a partir de la matriz de similitud o de distancias. Existen dos tipos de algoritmos. De *aglomeración*, que parten de elementos individuales y se van agregando a los grupos; y de *división*, parten del conjunto de elementos y se van dividiendo sicesivamente hasta llegar a los elementos individuales.

3.4 Métodos aglomerativos

El algoritmo es el siguiente:

1 Comenzar con tantas clases como elementos, es decir al inicio hay n clases (el número de observaciones). La distancia entre los grupos se definen en base a las distancias originales entre los grupos.

2 Se seleccionan los dos elementos más próximos en la matriz de distancia y se forma entre ellos una nueva clase o grupo.

3 Sustituir los dos elementos seleccionados en el paso anterior para definir el nuevo grupo mediante un representante. La distancia entre este nuevo elemento y el resto se define mediante determinados criterios de "distancias de aproximación"

4 Terminado el paso 3 se vuelve al paso 2, y se detiene el proceso hasta que quede una única clase (o el número de clases que a priori se determina)

[4] http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/libritos.txt
[5] http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/continua.dfw
[6] http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/escritores.htm

3.5 Criterios para definir distancias de proximidad.

3.5.1 El vecino más próximo

Supongamos que A y B eran grupos distintos que se han agrupado en una sola clase que denotamos por AB, y sea C otro grupo, entonces la nueva distancia entre AB y C que se propone es la que sigue

$$d(C, AB) = \min\{d(C, A), d(C, B)\}$$

La situación más usual, al inicio del algoritmo, es cuando $\{x_1, ..., x_l\}$ es un grupo ya constituido y se está calculando la distancia con otro elemento x_k, entonces

$$d(x_k, \{x_1, ..., x_l\} = \min_{j=1,...,l}\{d(x_k, x_j\}$$

3.5.2 El vecino más alejado

Con la misma nomenclatura anterior la distancia del vecino más alejado se define como

$$d(C, AB) = \max\{d(C, A), d(C, B)\}$$

3.5.3 Media de grupos

Supongamos que la cardinalidad de A, B son, respectivamente, n_ay n_b; entonces

$$d(C, AB) = \frac{n_a}{n_b + n_b}\, d(C, A) + \frac{n_b}{n_a + n_b}\, d(C, B)$$

Nota: este criterio no es invariante ante transformaciones monótonas, como los dos anteriores.

3.5.4 Media del centroide

Se aplica generalmente a variables continuas. Con la misma nomenclatura anterior se propone la distancia definida en la expresión siguiente

$$d^2(C, AB) = \frac{n_a}{n_b + n_b}\, d^2(C, A) + \frac{n_b}{n_a + n_b}\, d^2(C, B) - \frac{n_a n_b}{(n_a + n_b)^2}\, d^2(A, B)$$

Todos los buenos softwares estadísticos tienen estos (y otros) criterios incorporados. ¿Cuál es el mejor? Evidentemente no existe un test general de bondad, solo dependerá del objetivo de la investigación, de la intencionalidad que se le quiere dar a los cluster, y finalmente de la experiencia del propio investigador.