4 Análisis de grupos (*clustering*)

Clasificar casos u objetos en diferentes grupos es una necesidad en el mundo en el que vivimos donde hace falta poner orden y agrupar desde libros, organismos, personas, eventos, autos o, como en el presente trabajo, texto. El análisis de grupos o *clustering*, es una colección de métodos estadísticos que permiten agrupar casos sobre los cuales se miden diferentes variables que representan rasgos o características de los textos [8]. Así, casos que presenten características muy similares quedarán agrupados en conjuntos llamados grupos o *clusters*, de manera que se diferencien respecto de los casos agrupados en otros grupos. Estos grupos deben ser hallados sin información a priori y serán sugeridos únicamente por la propia esencia de los datos.

Uno de los problemas fundamentales del análisis de grupos es que no existe una definición precisa de grupo. Ello ha dado lugar al desarrollo de una gran cantidad de métodos; así, podemos hablar de dos grandes bloques de métodos de agrupamiento: los jerárquicos y los no jerárquicos o particionales. En los métodos jerárquicos, la pertenencia a un grupo en un nivel o jerarquía condiciona la pertenencia a grupos de nivel superior. Además, se dividen en aglomerativos y divisivos o disociativos, según la jerarquía, ya sea construida agrupando casos o bien dividiéndolos secuencialmente. Los métodos particionales obtienen una única partición de los datos mediante la optimización de alguna función adecuada. Estos métodos son también conocidos como métodos de optimización.

Los métodos particionales utilizan una matriz de datos mientras que los jerárquicos parten de una matriz de distancias o similitudes.

4.1 Distancias y similitudes

Supóngase un conjunto de datos X representado por una matriz m x n donde se conjuntan datos correspondientes a m casos con n variables. Cada renglón (caso) es un vector $X_i = (x_1, x_2, x_3, \ldots, x_n)$ de observaciones de n variables. Ahora, teniendo en cuenta que nuestro objetivo principal es hallar grupos que contengan casos similares, va a ser necesario medir las similitudes o bien las distancias que hay entre los casos.

Concepto de distancia: Dados dos puntos X_i y X_j pertenecientes a $\mathbb{R}^n$, se dice que se ha establecido una distancia, o métrica, entre ellos, si se ha definido una función d con las siguientes propiedades [18] [19] [8].

1. $d(X_i, X_j) \geq 0, \ \forall X_i, X_j \in X$
2. $d(X_i, X_i) = 0, \ \forall i \in X$

3. $d(X_i, X_j) = d(X_j, X_i),\ \forall X_i, X_j \in X$

La primera de las propiedades dice que todas las distancias deben ser no negativas. La segunda dice que cada caso tiene una distancia de cero respecto a sí mismo y la última de las propiedades establece la simetría, es decir, la distancia que puede haber de un caso X_i a otro caso X_j es la misma que hay del caso X_j al X_i. Obviamente, cuanto mayor sea la distancia $d(X_i, X_j)$, más alejados estarán los casos X_j y X_i.

Además de lo anterior, la distancia debe verificar que si tenemos tres puntos, la suma de las longitudes de dos lados cualesquiera del triángulo formado por los tres puntos debe siempre ser mayor que el tercer lado. Esta propiedad se conoce como la propiedad triangular. $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$, $\forall X_i, X_j, X_k \in X$

Como el número de casos es finito, es posible ordenar las interdistancias en una matriz simétrica de $m\ x\ m$, conocida como matriz de distancias sobre X:

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & & & \vdots \\ d_{m1} & d_{m2} & & d_{mm} \end{pmatrix}$$

Una matriz simétrica donde su diagonal principal son ceros.

Concepto de similitud: Dados dos puntos X_i y X_j pertenecientes a $\mathbb{R}^n$, se dice que se ha establecido una medida de similitud entre ellos si se ha definido una función s con las siguientes propiedades.

1. $0 \leq s(X_i, X_j) \leq 1,\ \forall X_i, X_j \in X$
2. $1 = s(X_i, X_i) \geq s(X_i, X_j),\ \forall X_i \in X$
3. $s(X_i, X_j) = s(X_j, X_i),\ \forall X_i, X_j \in X$

La primera propiedad nos dice que la similitud debe ser no negativa y establece una escala entre cero y uno. La segunda, que cada caso se parece a sí mismo más que a cualquier otro caso y la última establece la simetría. En cuanto a la interpretación se puede decir que cuanto mayor sea la similitud $s(X_i, X_j)$ más parecidos serán los casos X_i y X_j. La matriz de similitud se construye de la misma forma que la matriz de distancias, pero con la característica de que la diagonal principal está compuesta de unos.

$$s = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & & & \vdots \\ s_{m1} & s_{m2} & & s_{mm} \end{pmatrix}$$

La idea de similitud en esta matriz está muy asociada a la distancia y en ocasiones puede ser más fácil de calcular la distancia. Existen varias formas de pasar de una matriz de similitud a una matriz de distancias sobre *X* y viceversa.

Trasformación de una matriz de distancias a matriz de similitud: Dados dos casos X_i y X_j, la transformada de Gower [9] asegura que se cumple la siguiente relación:

$$d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$$

En general, sobre cada caso de X se habrán medido *n* variables y por lo tanto cada caso puede ser representado como un punto $p = (p_1, p_2, \dots, p_n)$ de $\mathbb{R}^n$, de manera que cada p_i es el valor que toma la *i*-ésima variable medida sobre el caso *p*. Dependiendo de la naturaleza de las variables que se hayan considerado (variables continuas, binarias o mixtas) utilizan diferentes tipos de distancias o similitudes. Así pues, hay una enorme variedad de distancias y similitudes. En esta tesis sólo se pretende dar una introducción al tema y exponer algunos de los métodos más utilizados, específicamente distancias para variables continuas y similitudes para variables binarias.

4.1.1 Distancias para variables continuas

Supongamos que las *n* variables consideradas sean continuas. A continuación se presentan algunos ejemplos de distancias estadísticas entre dos casos de $\boldsymbol{X}$ representados por los puntos $p = (p_1, p_2, \dots, p_m)$ y $q = (q_1, q_2, \dots, q_m)$.

1. Distancia euclidiana [18] [19] [21],

$$d_E(p,q) = [(p-q)(p-q)'] = \left[\sum_{i=1}^{m}(p_i - q_i)^2\right]^{1/2}$$

2. Distancia de Minkowsky [8][19] ($q \geq 1$),

$$d_M(p,q) = \left(\sum_{i=1}^{m}[p_i - q_i]^q\right)^{1/q}$$

Cuando *q*=2 ésta se reduce a la distancia euclidiana. Cuando *q*=1, se obtiene la distancia también conocida como métrica de Manhattan o distancia de Hamming.

3. Distancia valor absoluto[8],

$$d_{ABS}(p,q) = \sqrt{\sum_{i=1}^{m}[p_i - q_i]}$$

4. Distancia de Clark

$$d_C(p,q) = \sqrt[r]{\sum_{i=1}^{m}\left(\frac{1}{([p_i + q_i]^r)}([p_i - q_i]^r)\right)}$$

Cuando *r*=2, esta distancia corresponde a una distancia euclidiana normalizada. Cuando *r*=1 (corresponde a una de Hamming normalizada) se obtienen buenos resultados, pero implica mucha carga de procesamiento. Hay que notar que es muy parecida a la distancia de Minkowsky. La diferencia entre ambas es la ponderación que realiza la distancia de Clark a partir del valor de los mismos atributos que se están evaluando. Este factor de ponderación o normalización permite que comparativas numéricas entre atributos muy distantes se equilibren y por lo tanto mejoren los resultados [6].

4.1.2 Similitudes para variables binarias

Cuando todas las variables $x_1, x_2,\ldots,x_n$ de un caso X_i medidas sobre los casos son binarias, es decir, solamente toman valores 0 o 1, es más fácil calcular las similitudes para luego transformar estas en distancias. Habitualmente, el valor 0 indica que la característica en estudio no está presente, mientras que el valor 1 indica la presencia de la característica. Consideremos los casos X_i y X_j de algún conjunto de casos *X* representados como $p = (p_1, p_2, \ldots, p_n)$ y $q = (q_1, q_2, \ldots, q_n)$.
Para calcular la similitud entre ellos se usa la tabla que sigue (ver tabla 4). En ella se resume el recuento de las coincidencias de los valores que han tomado las *n* variables en los dos casos [8] [19]. Es decir:

- Los dos casos han tomado el valor de 1 simultáneamente en *a* variables,
- Los dos casos han tomado el valor de 0 simultáneamente en *d* variables,
- El caso i ha tomado el valor de 0 mientras que el caso j ha tomado el valor de 1 en b variables,
- El caso i ha tomado el valor de 1 mientras que el caso j ha tomado el valor de 0 en c variables.

		Caso i		
		1	0	
Caso j	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	n

Tabla 4 – Recuento de las coincidencias de *n* variables binarias definidas para dos casos *i* y *j* con *n* = *a*+*b*+*c*+*d*

Hay muchas maneras de definir similitudes con base en las cantidades a, b, c y d, pero sólo se presentarán las tres más habituales y una nueva medida conocida como energía textual [10] la cual no depende de la tabla 4.

1. Similitud de Sokal-Michener [19] [8],

$$S_{SM}(i,j) = \frac{a+d}{n}$$

2. Similitud de Rogers-Tanimoto [19],

$$S_{RT}(i,j) = \frac{a+d}{(a+d)+2(b+c)}$$

3. Similitud de Jaccard [8],

$$S_J(i,j) = \frac{a}{a+b+c}$$

4. Energía textual

Para el análisis de la energía textual es necesario tomar todos los casos en su conjunto. Así pues, consideramos una matriz *A* donde cada vector renglón es un elemento sobre *X*, y donde los valores A_{ij} representan la presencia o ausencia de la característica.

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & & & \vdots \\ A_{m1} & A_{m2} & & A_{mm} \end{pmatrix}$$

Para calcular la energía textual (esto es, la matriz de similitud) se utiliza la siguiente fórmula:

$$E_{textual} = -\frac{1}{2}(AxA')^2$$

Sin pérdida de generalidad, es posible considerar las magnitudes de la matriz de energía, pues todos los elementos obtenidos de la formula anterior son negativos. Esto es:

$$E = -E_{textual}$$

Cuya estructura cumple con la definición de similitud, donde cada vector de *X* se parece más a sí mismo y menos con los otros vectores de *X*.

4.2 Métodos jerárquicos

Los llamados métodos jerárquicos tienen por objetivo reunir grupos para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si se va efectuando este proceso de aglomeración o división sucesivamente, se minimice alguna distancia o bien se maximice alguna medida de similitud.

Los métodos jerárquicos se subdividen en aglomerativos y disociativos. Cada una de estas categorías presenta una gran diversidad de variantes.

1. Los métodos aglomerativos, también conocidos como ascendentes, comienzan el análisis con tantos grupos como casos haya. A partir de estas unidades iniciales se van formando grupos, de forma ascendente, hasta que al final del proceso todos los casos tratados quedan englobados en un mismo conglomerado.

2. Los métodos divisivos o disociativos, también llamados descendentes, constituyen el proceso inverso al anterior. Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez más pequeños. Al final del proceso se tienen tantas agrupaciones como casos han sido tratados.

4.2.1 Dendograma y jerarquía indexada

Las clasificaciones jerárquicas se pueden representar con un diagrama bidimensional conocido como dendograma en el cual se puede seguir de forma gráfica el procedimiento de unión de los grupos, mos-

trando qué grupos se van uniendo, en qué nivel concreto lo hacen, así como el valor de la medida de asociación entre ellos (índice de fusión $\propto$) cuando éstos se agrupan.

Para poder entender este proceso necesitamos la definición de jerarquía indexada.

Definición: Una jerarquía indexada $(\boldsymbol{C}, \propto)$ sobre el conjunto de casos $\boldsymbol{X}$ la forman una colección de grupos $\boldsymbol{C} \subset P(X)$ y un índice de fusión $\propto$: $\boldsymbol{C} \rightarrow \mathbb{R}^+$ de manera que $\boldsymbol{C}$ cumple los siguientes axiomas.

1. Intersección: Si $C, C' \in \boldsymbol{C}$, entonces $C \cap C' \in \{C, C', \emptyset\}$,
2. Unión: Si $C \in \boldsymbol{C}$, entonces $C = \cup \{C' | C' \in \boldsymbol{C}\ y\ C' \subset C\}$,
3. La unión de todos los grupos recoge todos los casos: $X = \cup \{C | C \in \boldsymbol{C}\}$

Y el índice de fusión $\propto$ cumple:

- $\propto (\mathrm{i}) = 0, \forall \mathrm{i} \in \mathrm{X}$
- $\propto (\mathrm{C}) \leq \propto (\mathrm{C}')$ si $\mathrm{C} \subset C'$

El primer axioma nos dice que, dados dos grupos de la jerarquía, o bien uno está contenido en el otro, o bien son disjuntos. Esto nos garantiza que un mismo caso no pueda estar en dos grupos diferentes. El segundo axioma dice que cada grupo es la unión de los grupos más pequeños que están contenidos en él. El índice de fusión $\propto$ mide la heterogeneidad de cada grupo. Es decir, cuanto mayor sea el índice de un grupo, más heterogéneo será éste.

Una vez establecida una jerarquía indexada sobre el conjunto de casos *X*, ésta puede resumirse en un dendograma y viceversa. Para ello representamos en el eje horizontal los elementos y simplemente se dibuja el proceso de creación de los grupos teniendo en cuenta el índice de fusión $\propto$.

4.2.2 Método del mínimo

Este método también se conoce como *Single Linkage* o vecino más próximo.
Supóngase que se ha construido una matriz de distancias D sobre el conjunto de casos *X*.

$$D = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & & & \vdots \\ d_{n1} & d_{n2} & & 0 \end{pmatrix}$$

1. Inicialmente cada caso forma un grupo. Es decir, la partición inicial es:

$$P_0 = \{\{1\}, \{2\}, \dots \{n\}\}$$

2. Supóngase la distancia más pequeña en la matriz de distancias; esto es, i y j ($d(i,j) = min_{k,l}\{d(k,l)\}$). Entonces la unión de estos casos formará un único nuevo grupo $\{i\} \cup \{j\} = \{i,j\}$

3. Se actualiza la matriz de distancias con la distancia entre dos grupos, definida como el mínimo de las distancias entre los casos de cada grupo.

$$d'(k,\{i,j\}) = \min\{d(k,i), d(k,j)\}$$

   La matriz de distancias actualizada D' de dimensión (n-1) x (n-1) está bien definida porque d es métrica.

4. Considerando la partición obtenida en el paso anterior, $P_1 = \{\{1\}, \{2\}, \dots, \{i,j\}, \dots \{m\}\}$
   se repiten los pasos 2 y 3 del algoritmo hasta que todos los casos de X formen un único grupo.

Finalmente, se define el índice de fusión $\propto$ de la siguiente manera:

$$\propto (\{\mathrm{i}\}) = 0, \mathrm{i} = 1, \dots, \mathrm{n};$$
$$\propto (\mathrm{C_i} \cup \mathrm{C_j}) = \mathrm{d}(\mathrm{C_i}, \mathrm{C_j})$$

Donde $\mathrm{C_i}$ y $\mathrm{C_j}$ son grupos que se han creado en el proceso de aglomeración. El resultado de este proceso, $(C, \propto)$, es una jerarquía indexada.

Ejemplo: Supóngase que se tiene la siguiente matriz de distancias D definida sobre $X = \{1,2,3,4,5\}$, se calculará paso a paso la jerarquía indexada que da el método del mínimo.

$$D = \begin{pmatrix} 0 & 1 & 3 & 4 & 7 \\ & 0 & 4 & 4 & 8 \\ & & 0 & 2 & 8 \\ & & & 0 & 7 \\ & & & & 0 \end{pmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix}$$

Como dice el paso 1, en la partición inicial $\mathrm{P_0}$ cada caso forma un grupo; así,

$$\mathrm{P_0} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$$

Se buscan los casos i, j más cercanos. En este caso, $min_{k,l}(k,l) = d(1,2) = 1$. Por lo tanto, son unidos para formar el primer grupo $\{1,2\}$. Posteriormente, se definen las distancias de cualquier caso al nuevo grupo $\{1,2\}$; así, por ejemplo,

$$\begin{aligned} d(3,\{1,2\}) &= \min\{d(3,1), d(3,2)\} \\ &= \min\{3,4\} \\ &= 3 \\ d(5,\{1,2\}) &= \min\{d(5,1), d(5,2)\} \\ &= \min\{7,8\} \\ &= 7 \end{aligned}$$

Con lo que queda la nueva matriz de distancias:

$$\begin{pmatrix} 0 & 3 & 4 & 7 \\ & 0 & 2 & 8 \\ & & 0 & 7 \\ & & & 0 \end{pmatrix} \begin{matrix} \{1,2\} \\ 3 \\ 4 \\ 5 \end{matrix}$$

Como se indica en el algoritmo se deben repetir los pasos 2 y 3. Es decir, se buscan los dos casos que tengan la distancia mínima. En este caso esta distancia mínima es $d(3,4) = 2$ y posteriormente actualizar la matriz de distancias, teniendo en cuenta el nuevo grupo $\{3,4\}$, así por ejemplo,

$$\begin{aligned} d(\{1,2\},\{3,4\}) &= \min\{d(\{1,2\},3), d(\{1,2\},4)\} \\ &= \min\{3,4\} \\ &= 3 \end{aligned}$$

Y la nueva matriz queda:

$$\begin{pmatrix} 0 & 3 & 7 \\ & 0 & 7 \\ & & 0 \end{pmatrix} \begin{matrix} \{1,2\} \\ \{3,4\} \\ 5 \end{matrix}$$

Al examinar esta nueva matriz se observa que la distancia mínima es $3 = d(\{1,2\},\{3,4\})$, por lo tanto forma un nuevo grupo $\{1,2,3,4\}$ y actualizando la matriz de distancias.

$$\begin{pmatrix} 0 & 7 \\ & 0 \end{pmatrix} \begin{matrix} \{1,2,3,4\} \\ 5 \end{matrix}$$

Este proceso ha generado la siguiente jerarquía aglomerativa indexada:

$$P_0 = \{\{1\},\{2\},\{3\},\{4\},\{5\}\}$$
$$P_1 = \{\{1,2\},\{3\},\{4\},\{5\}\}$$
$$P_2 = \{\{1,2\},\{3,4\},\{5\}\}$$
$$P_3 = \{\{1,2,3,4\},\{5\}\}$$
$$P_4 = \{\{1,2,3,4,5\}\} = \boldsymbol{X}$$

Finalmente se puede ver que el índice de fusión $\propto$ muestra la distancia a la que se han unido los grupos.

El dendograma correspondiente a esta jerarquía se puede ver en la siguiente figura.

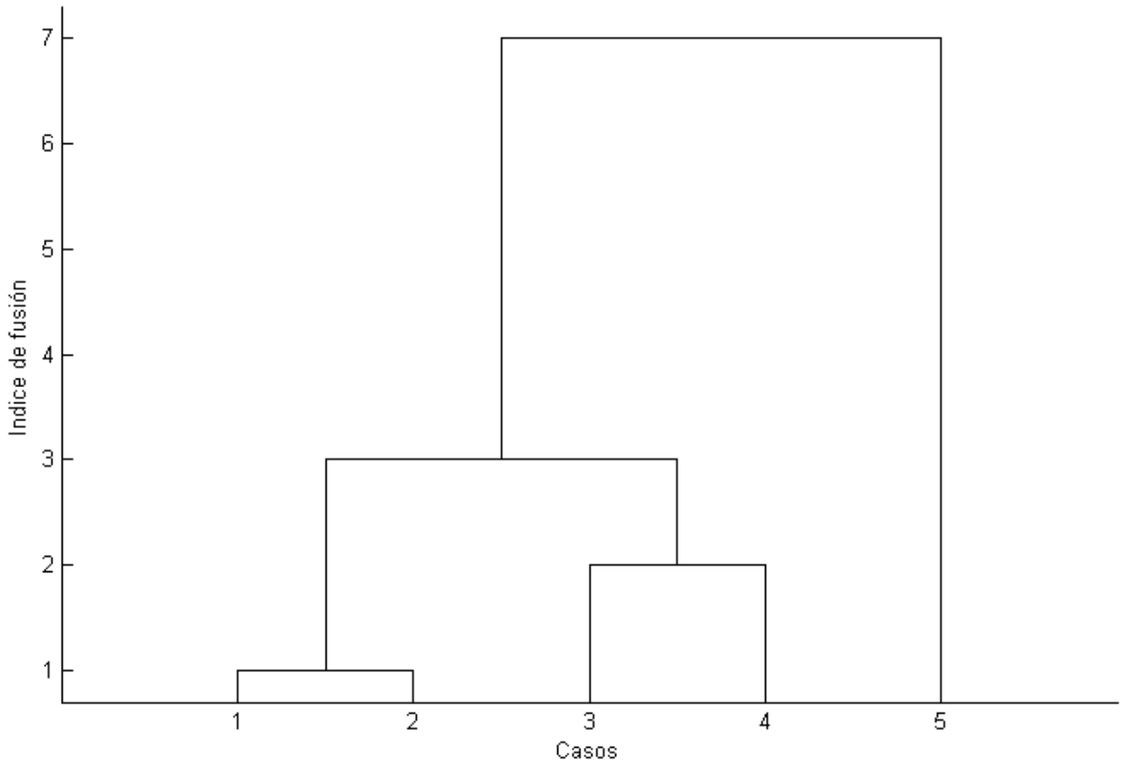

Figura 2 – Dendograma correspondiente al ejemplo del vecino más próximo