

Técnicas de clusterización

PFC

Ingeniería Industrial

09/2014

Tutor: Jesús Muñuzuri Sanz

Alumno: Juan de Dios Lara Albín

ÍNDICE

| | Pag. |
|--|------|
| 1 Clasificar | 2 |
| 2 El análisis clúster | 5 |
| 3 Clúster por individuos y variables | 10 |
| 4 Técnicas clúster | 11 |
| 5 Etapas | 16 |
| 5.1 Medidas de asociación | 21 |
| 5.2 Técnicas | 29 |
| 5.3 Validación e interpretación | 46 |
| 6 Metodología seleccionada | 48 |
| 7 Aplicación y resultados | 52 |
| 8 Conclusiones | 87 |
| 9 Referencias | 89 |

1 Clasificar

Clasificar objetos en categorías es una de las actividades más comunes, básicas y primitivas del hombre a lo largo de la historia. En un día nos podemos encontrar con multitud de personas, sucesos u objetos que son demasiado numerosas como para procesarlas mentalmente como elementos aislados.

La identificación o clasificación es el acto o proceso de asignar un objeto en el lugar que corresponda dentro de un conjunto de categorías establecido.

Los atributos básicos de cada categoría son conocidos, aunque haya algunas incertidumbres a la hora de asignar alguna observación dada.

Por ejemplo, para el desarrollo del lenguaje es necesario la clasificación, mediante el lenguaje nos ayudamos a través de las palabras a distinguir y reconocer los diferentes tipos de sucesos, objetos y personas que nos encontramos. Cada palabra es una etiqueta que usamos para describir una clase de objeto que posee ciertas características comunes, con lo cual podemos decir que nombrar es clasificar.

Como vemos, la clasificación es una actividad humana conceptual básica, pero no lo es solo en este ámbito, la clasificación puede ser algo fundamental también en la mayoría de las ramas científicas. Por ejemplo, en Biología, una de las principales preocupaciones desde las primeras investigaciones biológicas era la clasificación de los organismos.

Aristóteles ya construyó un completo sistema de clasificación de especies del reino animal, para comenzar dividió el reino animal en dos grupos principales: los que poseían sangre roja (vertebrados) y los que no poseían sangre roja (invertebrados). A su vez estos dos grupos los subdividió en otros dos grupos dependiendo de la forma en la que estos venían al mundo, ovíparos y vivíparos.

Teófrates, tras Aristóteles, redactó el primer informe básico sobre la clasificación y estructura de estas. Estos libros estaban ampliamente documentados y abarcaban tantos conceptos en sus temas que durante muchos siglos han sido la base de las investigaciones biológicas.

La clasificación no solo ha jugado un papel importante en la rama de la Biología, ha jugado también un papel central para el desarrollo de teorías en otros muchos campos de la ciencia, por ejemplo, Mendelejev en 1860 causó un profundo impacto en la comprensión de la estructura del átomo al clasificar los elementos en la tabla periódica, en Astronomía, la clasificación de las estrellas en enanas y gigantes usando el campo Herstsprung-Russell de temperatura frente a luminosidad, influyó de manera notoria a las teorías de la evolución de las estrellas.

Durante la segunda mitad del siglo XX se ha producido un gran aumento en el número de técnicas numéricas de clasificación disponibles. Este crecimiento ha ido paralelo al desarrollo de los ordenadores, que son necesarios para poder desarrollar el gran número de operaciones aritméticas que son necesarias.

Estas técnicas de clasificación actualmente son usadas en numerosos campos como pueden ser arqueología, psiquiatría, astronomía, investigación de mercados...

Estos métodos de clasificación reciben nombres específicos según el campo al que pertenezcan, en Biología por ejemplo se utiliza en nombre de Taxonomía, en Psicología se le llama Q-análisis, en inteligencia artificial se usa el término Reconocimiento de Patrones, en otras áreas se emplea simplemente agrupación o agrupamiento.

Pero el término genérico de dicha técnica se denomina Análisis Clúster, cuyo problema a resolver es siempre el mismo, dado un conjunto de m objetos individuales que pueden ser plantas, personas, animales... cada uno de los cuales descrito por una serie de n características o variables, hay que deducir una división útil en un número de grupos o clústeres. Tanto el número de grupos o clústeres como sus propiedades deben ser determinadas.

Con lo cual la solución que buscamos es una partición de los m objetos, es decir, un conjunto de grupos en los cuales un objeto pertenezca a un solo grupo y el conjunto de estos grupos contenga a todos los objetos.

2 El análisis clúster

Análisis Clúster es el término genérico usado para una amplia variedad de procedimientos mediante los cuales se pueden crear clasificaciones. Concretamente, un método clúster es un procedimiento estadístico multivariante que comienza con una serie de datos que contienen información sobre una muestra de objetos e intenta reorganizarlas en grupos relativamente homogéneos llamados clusters.

Unas de las características que diferencia al análisis clúster de otros métodos multivariantes es que no se conoce información sobre la estructura de las categorías que finalmente resultaran del análisis, disponemos de una serie de observaciones, siendo nuestro objetivo operacional descubrir las categorías en la que encajan nuestros objetos.

Es decir, el objetivo es ordenar las observaciones en grupos o clusters tales que su grado de asociación natural sea alto entre los miembros del mismo grupo y bajo entre los miembros de grupos diferentes.

Aunque como dijimos a priori no conocemos la estructura de las distintas categorías, podemos tener algunas nociones sobre algunas características deseables o inaceptables a la hora de establecer un determinado esquema de clasificación. De esto es informado el analista previamente de tal forma que pueda distinguir entre estructuras buenas y malas cuando se encuentra con ellas a lo largo del procedimiento.

Técnicas de clusterización

Con esto podemos llegar a la conclusión de que lo más fácil podría ser enumerar todas las posibilidades y elegir la más atractiva. El número de formas en las que podemos clasificar m observaciones en k grupos es un número de Stirling de segunda especie:

$$S_m^{(k)} = \frac{1}{k!} \sum_{i=0}^k (-1)^{k-i} \binom{k}{i} i^m$$

Y este problema se complica aún más ya que el número de grupos que obtendremos al final del procedimiento por lo general es desconocido, por lo que el número de posibilidades que nos encontramos es la suma de los números de Stirling, con lo cual para m observación tendríamos que el número total de posibles clasificaciones sería:

$$\sum_{j=1}^m S_m^{(j)}$$

El número de posibles clasificaciones podría ser excesivamente grandes, por ejemplo para el caso de 25 observaciones tendríamos $4E18$ posibles clasificaciones, número realmente elevado, con lo cual es necesario encontrar una solución aceptable considerando solamente un pequeño número de alternativas.

Técnicas de clusterización

El principal estímulo para el desarrollo de estos métodos fue el libro publicado en 1963 por los biólogos Sokal y Sneath: Principios de Taxonomía Numérica. En este libro sus autores argumentaban que un procedimiento eficiente para la generación de clasificaciones biológicas debe recoger todos los datos posibles sobre un conjunto de organismos de interés, estimar el grado de similitud entre dichos organismos y usar un método clúster para clasificar dichos organismos similares en un mismo grupo

A partir de este momento la literatura sobre Análisis Clúster se desarrolló de forma notoria. Existen dos razones que fueron causante de este rápido crecimiento y desarrollo en este tipo de técnicas.

- El desarrollo de los ordenadores

Antes del desarrollo de los ordenadores, aplicar los métodos clúster a conjunto grandes de datos era una tarea molesta y dificultosa desde el punto de vista computacional. Por ejemplo, clasificar un conjunto de datos con 200 entidades requiere buscar una matriz de similitud con 20.000 valores, tarea que obviamente es costosa en tiempo, con la difusión de los ordenadores esta tarea es mucho más factible.

- La clasificación: un procedimiento científico

Todas las ciencias se basan en clasificaciones que estructuran sus dominios de investigación. En una clasificación nos encontramos los mejores conceptos usados en una ciencia. La clasificación de los elementos, por ejemplo, es la base

Técnicas de clusterización

para comprender la química inorgánica y la teoría atómica de la materia; la clasificación de las enfermedades proporciona la base estructural para la medicina.

Pero a pesar de su popularidad, los métodos clúster están todavía poco comprendidos y desarrollados si los comparamos con otros métodos estadísticos multivariantes existentes como el análisis factorial, análisis discriminante o multidimensional scaling.

La literatura en las ciencias sociales sobre clusters refleja una serie desconcertante y a veces contradictoria de terminologías, métodos y aproximaciones, lo cual ha provocado que sea un método a veces impenetrable.

Debemos tener algunas precauciones sobre los métodos clúster:

- Los métodos clúster son procedimientos que en la mayoría de los casos no soportan un cuerpo de doctrina estadística teórica, es decir, la mayor parte de los métodos son heurísticos. Esto contrasta con otros métodos como el Análisis Factorial, que está basado en una extensa teoría estadística.

- La mayor parte de los métodos clúster han sido creados a partir de ciertas ramas científicas, con lo cual, inevitablemente, están impregnadas de un cierto sesgo procedente de su disciplina. Esta es una cuestión importante ya que cada disciplina tiene sus preferencias como son los datos a emplear en la construcción de la clasificación, con lo cual puede haber métodos válidos en psicología pero que no lo son en biología, o viceversa.

Técnicas de clusterización

- Distintos procedimientos de agrupación pueden generar soluciones distintas sobre el mismo conjunto de datos. Esto es debido al hecho de que estos métodos se han desarrollado a partir de fuentes dispares que dan origen a reglas diferentes de formación de grupos. Con lo cual es necesario la existencia de técnicas que determinen que método nos proporcionará los grupos más homogéneos.

3 Clúster por individuos y variables

En general, para un Análisis Clúster partimos de una matriz X que nos proporciona los valores de las variables para cada uno de los individuos que son objeto de estudio, la matriz X nos queda:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix}$$

Siendo la i -ésima fila de la matriz los valores de cada variable para el i -ésimo individuo y la j -ésima columna los valores de la j -ésima variable a lo largo de todos los individuos de la muestra.

Como ya hemos comentado, el objetivo de clasificar los datos es agrupar individuos u objetos que se encuentran representados por las filas de la matriz X . Podemos obtener también una clasificación de las variables que describen a dichos individuos trasponiendo la matriz X , teniendo ahora los valores de las variables en cada fila.

4 Técnicas Clúster

En la actualidad existen multitud de técnicas clúster diferentes, aunque todas ellas persiguen el mismo objetivo. La clasificación que aquí veremos está referida a algunas de las técnicas existentes, trataremos los métodos más usados en las aplicaciones prácticas, y por ende sobre los que se posee un mayor grado de experiencia, estos suelen ser los implementados normalmente por los paquetes estadísticos existentes, ya que hay que tener en cuenta que para realizar estas técnicas es necesario un potente ordenador y programa informático, de otra manera no sería factible el desarrollo práctico de ninguna técnica clúster.

Podemos distinguir dos categorías principales en el análisis de técnicas clúster: métodos jerárquicos y métodos no jerárquicos.

Los métodos jerárquicos tienen por objetivo agrupar todos los clúster para formar un clúster nuevo o separar alguno existente para crear otros dos clústeres distintos, de manera que se minimice alguna función distancia o se maximice alguna medida de similitud.

Estos métodos a su vez los podemos subdividir en aglomerativos y disociativos. Los métodos aglomerativos comienzan el análisis con tantos grupos como individuos hay en el estudio. A partir de los individuos del estudio, vamos formando grupos de forma ascendente de forma que cuando finalicemos el proceso, todos los individuos se encuentran en un mismo grupo o clúster.

Técnicas de clusterización

Por otro lado los métodos disociativos o divisivos realizan el proceso inverso al anterior, es decir, empiezan con un grupo que contiene todos los individuos de la muestra y a partir de este conglomerado se van formando a través de sucesivas divisiones grupos cada vez más pequeños. Al final del proceso tenemos tantos grupos como individuos hay en la muestra, teniendo en cada grupo un solo individuo.

Independientemente del proceso de agrupamientos, existen distintos criterios para ir formando los conglomerados, estos criterios se basan en una matriz de similitudes o de distancias. De estos métodos podemos destacar:

1. Método del amalgamamiento simple.
2. Método del amalgamamiento completo.
3. Método del promedio entre grupos.
4. Método del centroide.
5. Método de la mediana.
6. Método de Ward.

En cuanto a los métodos no jerárquicos, también conocidos como partitivos o de optimización, su objetivo es realizar una partición de los individuos en K grupos, siendo este número de grupos especificado a priori, es decir, el investigador debe especificar antes de aplicar la técnica clúster el número de grupos que se forman, siendo esta la principal diferencia con respecto a los métodos jerárquicos.

Una vez tenemos el número de grupos a formar, la asignación de los individuos a cada grupo se hace mediante algún procedimiento que optimice el criterio de selección. Otra diferencia importante con respecto a los métodos jerárquicos es que con este método se trabaja con la matriz de datos original y no es necesario su conversión en una matriz de distancias o similitudes.

Dentro de los métodos no jerárquicos podemos distinguir cuatro familias:

1- Reasignación.

Este método permite que un individuo que es asignado a un grupo en un determinado paso del proceso, pueda ser reasignado a otro grupo diferente en un paso posterior si con ellos se optimiza el criterio de selección. Este proceso acaba cuando ya no quedan individuos cuya reasignación pueda optimizar el resultado conseguido. Dentro de estos métodos están:

- a) El método K-Medias.
- b) El Quick-Clúster análisis.
- c) El método de Forgy.
- d) El método de las nubes dinámicas.

2- Búsqueda de la densidad.

Dentro de estos métodos podemos distinguir los que proporcionan una aproximación tipológica y los que proporcionan una aproximación probabilística.

Técnicas de clusterización

En los métodos que nos proporcionan una aproximación tipológica, los grupos se forman buscando las zonas en las que se da una mayor concentración de los individuos, entre estos están:

- a) El análisis modal de Wishart.
- b) El método Taxmap.
- c) El método de Fortin.

Los que nos proporcionan una aproximación probabilística siguen una ley de probabilidad según la cual los parámetros varían de un grupo a otro, tratando en este caso de encontrar los individuos que pertenecen a la misma distribución. Entre los métodos de este tipo podemos destacar el método de las combinaciones de Wolf.

3- Directos

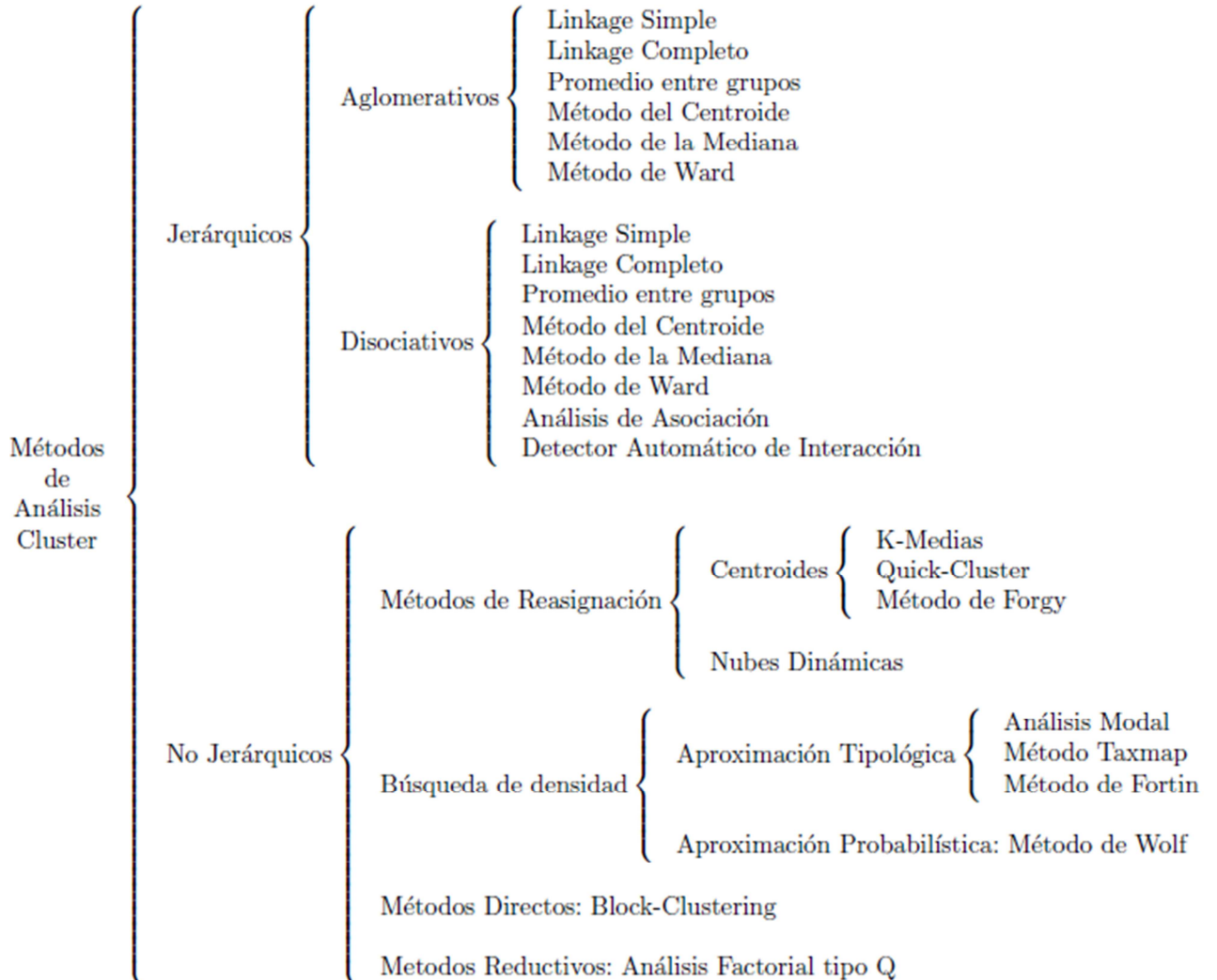
Mediante este método podemos clasificar simultáneamente individuos y variables, entre estos métodos podemos destacar el Block-Clustering.

4- Reducción de dimensiones.

Este método se centra en la búsqueda de unos factores en el espacio de los individuos, correspondiéndose cada factor a un grupo, también se les conoce como Análisis Factorial tipo Q.

Técnicas de clusterización

En resumen este es el esquema de los métodos clúster:



5 Etapas

Podemos resumir en los siguientes puntos las etapas a seguir en cualquier procedimiento que se base en técnicas clúster:

-Elección de variables

Inicialmente debemos elegir el conjunto de características concretas que vamos a usar para describir a cada individuo, las cuales constituirán un marco de referencia para formar los grupos o clusters, dicha elección reflejara nuestra opinión acerca del propósito de la clasificación.

La primera duda que podemos plantearnos es acerca de que variables son realmente relevantes a la hora de hacer la clasificación que deseamos. Es importante tener en cuenta que la elección inicial que hagamos de las variables es en sí misma una categorización de datos, para lo cual solo hay limitadas directrices matemáticas y estadísticas.

Otra cuestión importante es el número de variables que vamos a considerar, ya que es probable que en muchas aplicaciones nos equivoquemos considerando demasiadas medidas, lo cual provoca diversos problemas, estos problemas pueden ser a nivel computacional o porque dichas variables adicionales oscurezcan la estructura de los grupos

Técnicas de clusterización

Puede ocurrir que en algunas aplicaciones las variables que describen nuestros objetos a clasificar no estén medidas todas en las mismas unidades, siendo las variables de tipos completamente diferentes, algunas categóricas, otras ordinales e incluso otras que tengan de una escala de tipo intervalo. De esta manera no sería correcto tratar como equivalente, por ejemplo, la altura medida en metros, el valor de la valentía en una escala de 10 puntos y el peso medido en kilos.

Para variables de tipo intervalo la solución general consiste en tipificar las variables antes del análisis, calculando las desviaciones típicas a partir de todos los individuos, algunos autores como por ejemplo Fleiss y Zubin, consideran que esta técnica puede tener serias desventajas ya que diluye las diferencias entre grupos sobre las variables que más discrimine, como alternativa sugieren emplear la desviación estándar entre grupos para tipificar.

Cuando tenemos todas las variables de tipos diferentes lo que se suele hacer es convertir todas las variables en variables binarias antes de calcular las similitudes. Este procedimiento aunque puede ser muy clarificador tiene la desventaja de que puede sacrificar parte de información.

- Medida de asociación

Para medir la proximidad entre los objetos del estudio se requiere establecer una medida de asociación. Esta medida de asociación suele venir expresada en términos de distancias cuando los objetos de estudio del Análisis Clúster son

individuos, y suele venir expresada en medidas del tipo coeficiente de correlación cuando los objetos de estudio son las variables.

- Técnicas Clúster

Una vez que tenemos la relación entre cada objeto del estudio, podemos pasar a agruparlos atendiendo a dicha relación. Los métodos clúster propuestos y desarrollados en los últimos años son bastante numerosos y diversos en cuanto a su concepción.

Estos métodos podemos clasificarlos en un primer estado, en jerárquicos y no jerárquicos, cuya diferencia fundamental es que mientras que en los jerárquicos las asignaciones de los objetos de estudio a los clústeres que se van creando, van permaneciendo estables durante todo el proceso sin cambiar de grupo, no permitiendo reasignaciones posteriores a clústeres diferentes, en los no jerárquicos si podemos reasignar los objetos de estudio a clústeres distintos en pasos posteriores del proceso.

Otra diferencia importante es que en los métodos jerárquicos, debemos sacar nuestras propias conclusiones a cerca del número final de clúster, mientras que en los no jerárquicos el número general de clústeres esta impuesto de antemano

Con lo cual en algunos problemas que se nos planteen, la elección del método a emplear será relativamente natural, atendiendo a la naturaleza de los datos usados y de los objetivos finales que se persiguen, aunque en otros casos la elección del método a usar no será tan clara.

A la hora de realizar una aplicación práctica, es recomendable no elegir un solo procedimiento, siempre es bueno abarcar un amplio abanico de posibilidades y contrastar los resultados obtenidos con distintos métodos clusters. De esta manera si los resultados obtenidos son similares podemos obtener conclusiones mucho más válidas sobre la estructura de nuestra clasificación, de lo contrario no obtendremos demasiada información, y si existen grandes diferencias entre los distintos métodos podríamos llegar a plantearnos el hecho de que tal vez los datos con los que hemos trabajado no obedezcan a una estructura bien definida.

- Validación e interpretación

La validación de la estructura obtenida es la última etapa en la secuencia del desarrollo del método clúster. Es una etapa muy importante del estudio ya que en ella es donde se van a obtener las conclusiones definitivas del estudio.

Hay diversos métodos en la literatura para validar un procedimiento clúster, cuando trabajamos con métodos jerárquicos nos podemos plantear dos problemas, el primero de ellos es en que medida representan la estructura final obtenida las similitudes o diferencias entre los objetos de estudio, la segunda es la elección del número idóneo de agrupaciones o clusters que mejor representen la estructura natural de los datos.

Para el primero de los problemas, en 1962 Sokal y Rohlf propusieron el uso del coeficiente de correlación cofenético. Este coeficiente mide la correlación entre las distancias iniciales, tomadas a partir de los datos originales, y las distancias

finales con las cuales los individuos se han unido durante el desarrollo del método.

Valores altos del coeficiente cofenético indica que durante el proceso no ha ocurrido una gran perturbación en lo que se refiere a la estructura original de los datos.

Con respecto al segundo problema planteado, son muchas las técnicas existentes, las más importantes las abordaremos más adelante cuando profundizemos en los métodos jerárquicos.

En los métodos no jerárquicos, las cuestiones anteriores van perdiendo sentido ya que el número de agrupaciones viene inicialmente impuesto. Algunos autores han propuesto el empleo de técnicas multivariantes como el análisis multivariante de la varianza, o bien desarrollar múltiples análisis de la varianza sobre cada variable en cada clúster.

Estos procedimientos, evidentemente, plantean serios problemas y no deben ser considerados como definitivos. Una técnica usualmente empleada, de tipo remuestreo, es la de tomar varias submuestras de la muestra original y repetir el análisis sobre cada una.

Si tras repetir el análisis sobre ellas se consiguen soluciones aproximadamente iguales, y parecidas a la obtenida con la muestra principal, se puede intuir que la solución obtenida puede ser válida, si bien esto no sería argumento suficiente para adoptar tal decisión.

No obstante, este método es más útil empleado de forma inversa, en el sentido de que si las soluciones obtenidas en las diversas submuestras no guardan una cierta similitud, entonces parece evidente que se debiera dudar de la estructura obtenida con la totalidad de la muestra.

5.1 Medidas de asociación

Una vez que hemos visto que el objetivo de un análisis Clúster es encontrar agrupaciones naturales del conjunto de objetos de la muestra, tenemos que definir que se entiende por agrupación natural y en base a que criterios podemos decir que dos grupos son más o menos similares.

Para medir la similitud entre objetos de la muestra podemos usar la distancia métrica o la similitud.

La distancia se define como: Siendo U un conjunto finito o infinito de elementos, una función $d: U \times U; \mathbb{R}$ se llama una distancia métrica si todo x e y perteneciente a U cumple:

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \Leftrightarrow x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z) , \forall z \in U$

Técnicas de clusterización

En cuanto a la medida de similaridad se define como: Siendo U un conjunto finito o infinito de elementos. Una función $s: U \times U; \mathbb{R}$ se llama similaridad si todo x e y perteneciente a U cumple:

$$1. s(x, y) \leq s_0$$

$$2. s(x, x) = s_0$$

$$3. s(x, y) = s(y, x)$$

Siendo s_0 un número real finito arbitrario.

Por lo general consideraremos m individuos sobre los cuales se han medido n variables X_1, \dots, X_n . De esta manera tenemos $m \times n$ datos que distribuiremos en una matriz de dimensión $m \times n$:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix}$$

En esta matriz la i -ésima fila de la matriz contiene los valores de cada variable para el i -ésimo individuo, mientras que la j -ésima columna contiene los valores

pertenecientes a la j -ésima variable a lo largo de todos los individuos de la muestra.

Distinguimos entre medidas de asociación para individuos o para variables, aunque como podemos ver, técnicamente hablando pueden ser válidas para ambos, simplemente trasponiendo la matriz y convirtiendo las filas en columnas y las columnas en filas.

Para poder crear agrupaciones de variables debe de existir alguna medida numérica que caracterice las relaciones que existen entre las variables. La base de trabajo de cualquier técnica clúster es que estas medidas numéricas de asociación sean comparables, es decir, si la medida de asociación de dos variables es 0.86 y el de otro par de variables es 0.67, el primer par está más fuertemente asociado que el segundo.

Cada medida refleja una asociación en un sentido particular, y por ello es necesario para el problema concreto elegir una medida apropiada.

Algunas medidas de asociación entre variables son las siguientes:

- Coseno del ángulo de vectores

Consideramos dos variables X_i y X_j , muestreadas sobre m individuos, y siendo x_i y x_j los vectores cuyas k -ésimas componentes indiquen el valor de la variable correspondiente en el k -ésimo individuo:

$$x_i = (x_{1i}, \dots, x_{mi})' \quad ; \quad x_j = (x_{1j}, \dots, x_{mj})'$$

$$\cos(\beta) = \frac{x_i' x_j}{\|x_i\| \|x_j\|} = \frac{\sum_{l=1}^m x_{li} x_{lj}}{\left(\sum_{l=1}^m x_{li}^2 \sum_{l=1}^m x_{lj}^2 \right)^{\frac{1}{2}}}$$

Siendo el coseno del ángulo una medida de similaridad entre x_i y x_j , con valores entre -1 y 1. Esta medida es independiente, salvo signo, de la longitud de los vectores considerados, es invariante ante homotecias, excepto un eventual cambio de signo.

- Coeficiente de correlación.

Considerando las variables anteriores X_i y X_j y centrándolas respecto a sus medias, obtenemos unas nuevas variables cuyos valores para la muestra de los m individuos es:

$$\hat{x}_i = (x_{1i} - \bar{x}_i, \dots, x_{mi} - \bar{x}_i)' \quad ; \quad \hat{x}_j = (x_{1j} - \bar{x}_j, \dots, x_{mj} - \bar{x}_j)'$$

El producto escalar de estas dos variables se llama dispersión. El producto escalar de la primera variable por si misma se llama dispersión de x_i o suma de

los cuadrados de las desviaciones respecto a dicha variable. Dividiendo por m ambas expresiones obtenemos la covarianza y la varianza, respectivamente.

$$\text{Cov}(x_i, x_j) = \frac{\widehat{x}_i' \widehat{x}_j}{m} = \frac{1}{m} \sum_{l=1}^m (x_{li} - \bar{x}_i)(x_{lj} - \bar{x}_j)$$

$$\text{Var}(x_i) = \frac{\widehat{x}_i' \widehat{x}_i}{m} = \frac{1}{m} \sum_{l=1}^m (x_{li} - \bar{x}_i)^2$$

La correlación muestral entre x_i y x_j se define como:

$$r = \frac{\text{Cov}(x_i, x_j)}{(\text{Var}(x_i) \text{Var}(x_j))^{\frac{1}{2}}} = \frac{\sum_{l=1}^m (x_{li} - \bar{x}_i)(x_{lj} - \bar{x}_j)}{\left(\sum_{l=1}^m (x_{li} - \bar{x}_i)^2 \sum_{l=1}^m (x_{lj} - \bar{x}_j)^2 \right)^{\frac{1}{2}}}$$

La diferencia entre este último método y el ángulo del coseno entre variables es que el anterior método del ángulo del coseno se basa en los datos originales y por ende emplea las desviaciones al origen, mientras que el coeficiente de correlación usa los datos centrados y por lo tanto emplea las desviaciones respecto a la media.

Si el origen se encuentra bien establecido y tiene sentido, entonces los datos originales tienen sentido de forma absoluta y el coseno es una medida apropiada

de asociación, si por el contrario el origen es arbitrario o elegido a conveniencia, entonces los datos originales tienen sentido relativo respecto a su media, pero no respecto al origen, en este caso es más apropiado usar el método del coeficiente de correlación.

Consideramos ahora que los objetos de la muestra a clasificar son individuos en vez de variables, algunas medidas de asociación para individuos son las siguientes:

- Distancia euclídea

La distancia euclídea es la más intuitiva y la más utilizada en la práctica. Es la que calcula la distancia en línea recta entre los puntos en el espacio o en el hiperespacio de la nube de puntos original. Esta distancia en realidad es una aplicación del Teorema de Pitágoras.

Consideramos dos individuos tomados de la población, lo cual corresponde a tomar dos filas de la matriz X:

$$x_i = (x_{i1}, \dots, x_{in})'$$

$$x_j = (x_{j1}, \dots, x_{jn})'$$

La distancia euclídea se define como:

$$\sqrt{\sum_{l=1}^n (x_{il} - x_{jl})^2}$$

Cuyos valores son invariantes respecto a las transformaciones ortogonales.

A pesar de su sencillez a la hora de calcular dicha distancia, nos encontramos con dos inconvenientes.

El primero de ellos es que esta distancia es sensible a las unidades de medida de las variables, ya que las diferencias entre valores de variables medidas con valores elevados contribuirán en mayor medida que las diferencias entre los valores de las variables con valores bajos. Esto provoca que un cambio de escala provoca también un cambio en la distancia entre los individuos. Pero esto lo podemos solucionar tipificando previamente las variables, o utilizando la distancia euclídea normalizada.

El segundo problema deriva de la naturaleza de las variables. Si las variables se encuentran correlacionadas, estas variables nos proporcionarán información redundante. Esto trae en consecuencia que la distancia euclídea inflará la disimilaridad o divergencia entre los individuos de la muestra.

En definitiva, dicha distancia, será recomendable cuando las variables sean homogéneas y estén medidas en unidades similares o cuando se desconozca la matriz de varianzas.

- Distancia de Minkowski

La distancia de Minkowski se define como:

$$d_p(x_i, x_j) = \|x_i - x_j\|_p = \left(\sum_{l=1}^n |x_{il} - x_{jl}|^p \right)^{\frac{1}{p}}$$

Siendo esta invariante ante traslaciones.

Algunos casos particulares para valores de p concretos son los siguientes:

- Distancia ciudad o City Block (p=1)

$$d_1(x_i, x_j) = \sum_{l=1}^n |x_{il} - x_{jl}|$$

- Distancia de Chebychev (p=∞)

$$d_\infty(x_i, x_j) = \text{Max}_{l=1, \dots, n} |x_{il} - x_{jl}|$$

5.2 Técnicas Clúster

Una vez seleccionadas las variables y calculada la matriz de similitud, nos queda el proceso de seleccionar el algoritmo para formar las agrupaciones o clusters.

Esta no es una tarea sencilla ya que existen diversos algoritmos y además están en constante desarrollo en la actualidad. Pero el criterio esencial de todos ellos es que intentan maximizar las diferencias entre los conglomerados y minimizar las diferencias entre objetos de un mismo clúster.

Existen dos grandes categorías de algoritmos de obtención de conglomerados: los jerarquizados y los no jerarquizados.

Los métodos jerárquicos tienen por objetivo agrupar clústeres para formar uno nuevo o para separar algún cluster ya existente dando origen a otros dos a partir de este, de manera que si vamos implementando dicho método vamos minimizando alguna medida de distancia o maximizando alguna medida de similitud.

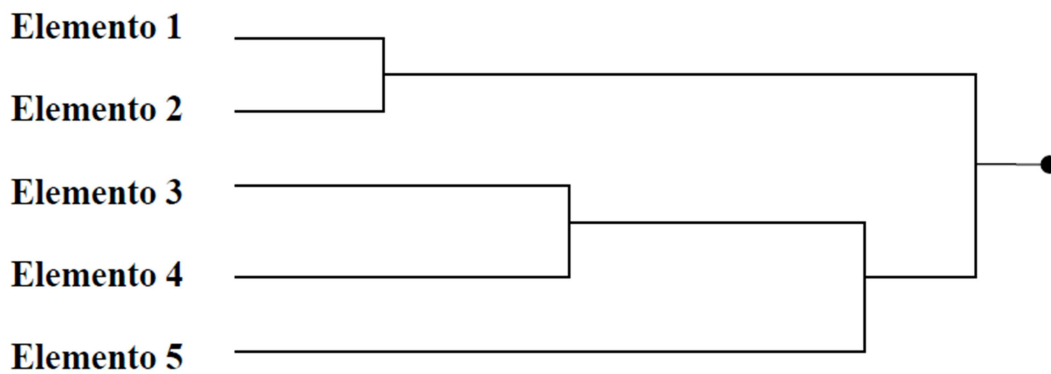
Los métodos jerárquicos a su vez los podemos dividir en aglomerativos y disociativos. Cada una de estas categorías presenta una gran diversidad de variantes.

Los métodos aglomerativos o también llamados métodos ascendentes, comienzan el análisis con tantos grupos como individuos hay en nuestra muestra. A partir de estos grupos iniciales vamos formando los sucesivos grupos de forma

Técnicas de clusterización

ascendente, hasta que tenemos al final del proceso todos los objetos de la muestra en un mismo clúster.

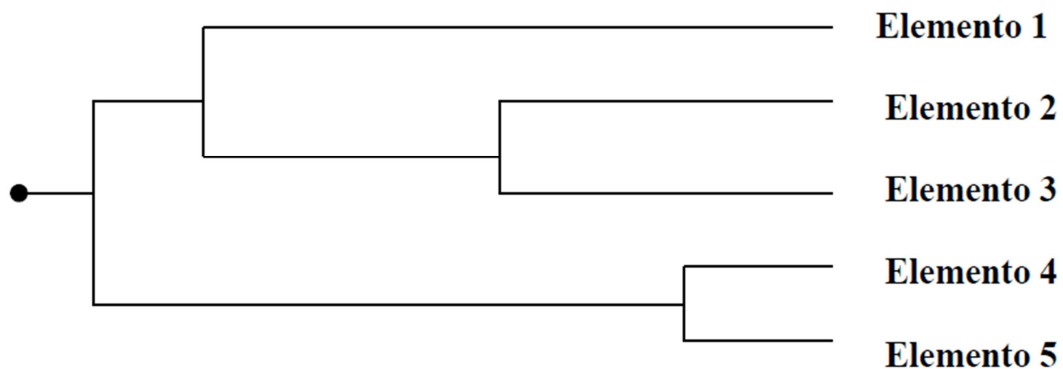
Método jerárquico aglomerativo:



Los métodos disociativos o también llamados métodos descendentes se basan en el mismo principio que los aglomerativos pero de manera inversa, comienzan con un grupo que contiene todos los objetos de la muestra, y a partir de este grupo y a través de sucesivas divisiones se van formando los grupos cada vez más pequeños. Al final del procedimiento obtenemos tantos grupos como objetos hay en la muestra, cada objeto perteneciente a un grupo.

Método jerárquico disociativo:

Técnicas de clusterización



Centrándonos en los métodos aglomerativos, tenemos un conjunto de individuos de la muestra n , en donde en el nivel $K=0$, tenemos n grupos. En el siguiente nivel se agruparan aquellos individuos que tengan la mayor similitud o menor distancia, obteniendo ahora $n-1$ grupos, de forma genérica y siguiendo con la misma estrategia, en los siguientes niveles se agruparan aquellos dos individuos o grupos ya formados con menor distancia o mayor similitud, obteniendo en un nivel genérico L , $n-L$ grupos.

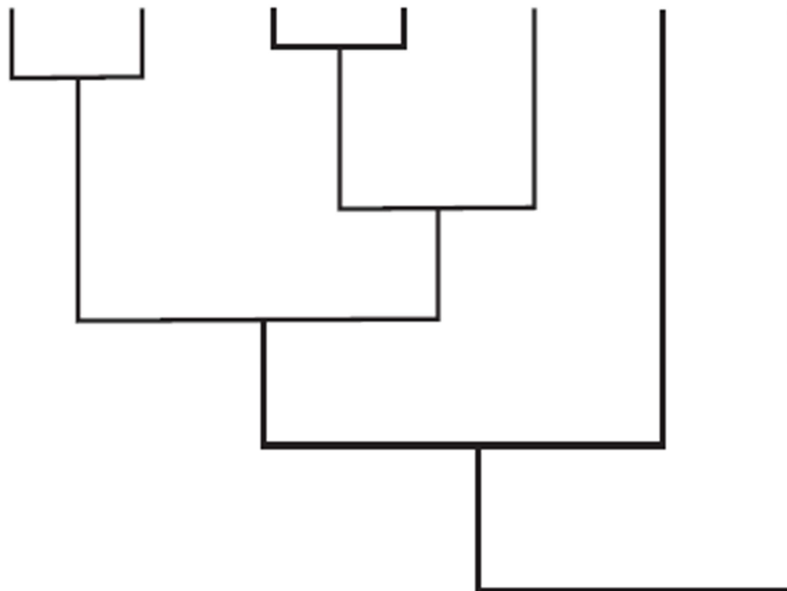
Al final del proceso llegaremos al nivel $n-1$ donde tenemos un clúster que contiene a todos los individuos de la muestra. Aunque el proceso puede terminar antes si lo que queremos obtener es un número de grupos prefijado o se detecta a través de un contraste de significación, que hay razones estadísticas para no continuar agrupando clusters, ya que los más similares no son lo suficientemente homogéneos como para determinar una misma agrupación.

Este proceso tiene la particularidad de que si en un determinado nivel agrupamos dos clusters, estos quedan ya agrupados durante todo el proceso.

Técnicas de clusterización

Los métodos jerárquicos nos permite la construcción de un árbol de clasificación que recibe el nombre de dendograma, mediante el cual podemos observar de forma gráfica el procedimiento seguido para unir los individuos de la muestra, mostrándonos los grupos que se van uniendo y en qué nivel lo hacen, así como el valor de la medida de asociación entre los grupos cuando estos se agrupan

Dendograma:



Ahora vamos a ver algunos de los procedimientos que podemos emplear a la hora de crear las aglomeraciones o clusters en las diversas etapas o niveles de un procedimiento jerárquico. Estos procedimientos no proporcionan una solución óptima para todos los problemas que en la práctica pudieran plantearse, ya que es

posible llegar a resultados diferentes según el método elegido. El conocimiento de nuestro problema, la experiencia y nuestro buen criterio sugerirá el método más adecuado.

Aun así, es recomendable siempre usar varios procedimientos con la finalidad de contrastar los resultados obtenidos y sacar nuestras propias conclusiones, tanto si hubiera coincidencias en los resultados obtenidos como si no los hubiera.

- Distancia mínima o similitud máxima.

Este método está basado en la distancia mínima entre individuos de los distintos conglomerados, también se conoce con el nombre del vecino más cercano. La distancia entre dos clusters es la distancia más corta que exista entre un punto de un conglomerado y otro punto del otro conglomerado.

Es decir, con este método se considera que la distancia o similitud entre dos clusters viene dada por la distancia mínima o similitud máxima entre los individuos que la componen.

Con este método tenemos el problema de que se pueden producir largas cadenas, llegándose a conformar una sola cadena cuando los clusters no están bien definidos.

Sus características más importantes son:

Técnicas de clusterización

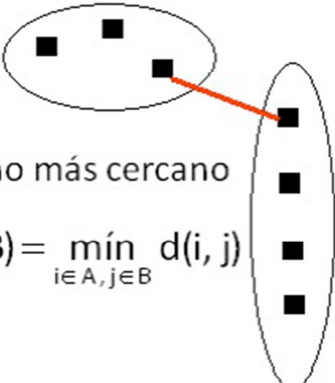
- No es útil para resumir datos.
- Útil para detectar outliers (últimos en unirse a la jerarquía).
- Pueden usarse medidas de similitud o de la distancia.
- Tiende a construir clusters demasiado grandes y sin sentido.
- Invariante bajo transformaciones monótonas de la matriz de distancias.
- Se crean grupos más homogéneos pero permite cadenas de alineamientos entre sujetos muy lejanos

Siendo A y B dos clusters:

conglomerados más alargados

▪ Vecino más cercano:

$$d(A, B) = \min_{i \in A, j \in B} d(i, j)$$



vecino más cercano

$$d(A, B) = \min_{i \in A, j \in B} d(i, j)$$

- Distancia máxima o similitud mínima.

Técnicas de clusterización

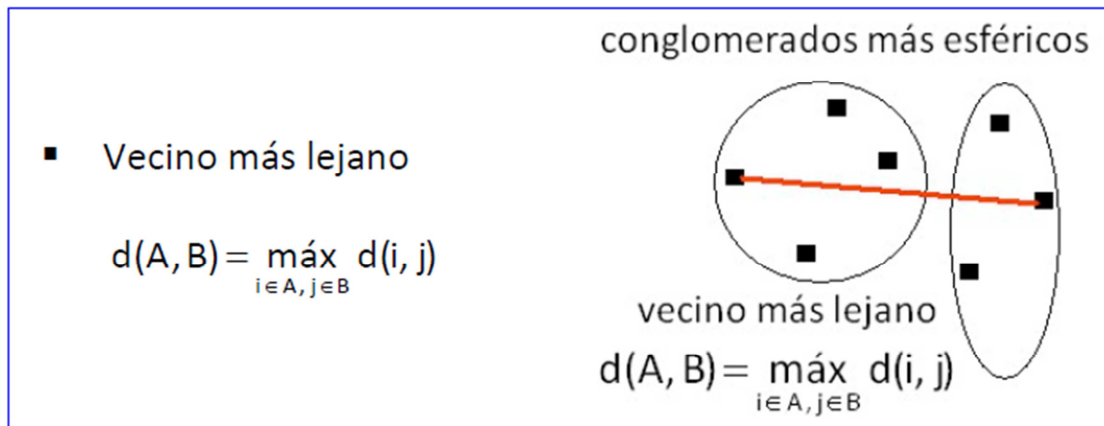
Este método es similar al anterior pero con la diferencia de que las distancias entre dos clusters las tomamos como la máxima distancia existente entre dos puntos de los clusters. Esta técnica elimina el problema de las largas cadenas del método anterior. También es conocido como el método del vecino más lejano.

Es decir, consideramos que la distancia o similitud entre dos clusters hay que medirla atendiendo a sus componentes más dispares, es decir, la distancia o similitud entre los dos clusters viene dada respectivamente por la máxima distancia o la mínima similitud entre los componentes de estos clusters.

Sus características son:

- Útil a la hora de detectar outliers.
- Pueden usarse medidas de la similitud o de la distancia.
- Tienden a construirse clusters pequeños y compactos.
- Invariante bajo transformaciones monótonas de la matriz de distancias.
- Solventa el problema del método anterior pero los grupos son más heterogéneos.

Siendo A y B dos clusters:



- Promedio no ponderado

Este método se basa en el criterio de aglomeración de la distancia media de todos los individuos de un conglomerado con los de otro. Esta técnica ya no depende de los valores extremos, como en los casos anteriores, y la partición se basa en todos los miembros de los conglomerados en lugar de un par único de miembros extremos.

Este enfoque tiende a agrupar los conglomerados con variaciones reducidas dentro del conglomerado, aunque tiende a estar sesgado hacia la producción de conglomerados con aproximadamente la misma varianza.

Es decir mediante este método, la distancia o similitud de dos clusters se obtiene como la media aritmética entre la distancia o la similitud de los componentes de los mismos.

Técnicas de clusterización

Mediante este método no se tiene en cuenta el tamaño de los clusters, lo que significa que le da igual importancia a la distancia entre dos clusters, independientemente del tamaño de estos.

Sus características son:

- Proporciona clusters de tamaño intermedio.
 - Pueden utilizarse medidas de la similitud o de la distancia.
 - No es invariante por transformaciones monótonas de las distancias.
 - Tiende a fusionar clusters con varianzas pequeñas y con la misma varianza.
 - Buena representación gráfica de los resultados.
- Promedio ponderado.

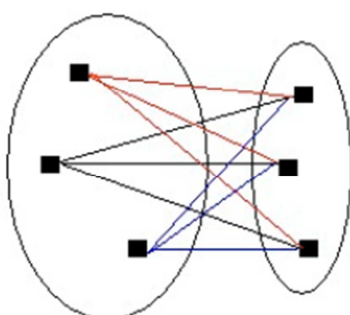
Ahora consideramos que la distancia o similitud entre dos clusters viene definida por el promedio ponderado de las distancias o similitudes de los componentes de un clusters respecto a los de otro.

Ahora si, al contrario del método anterior, tenemos en cuenta el tamaño de los clusters.

Siendo A y B dos clusters:

conglomerados más robustos

- Promedio de grupo

$$d(A, B) = \frac{1}{n_A \cdot n_B} \sum_{i \in A, j \in B} d(i, j)$$


promedio de grupo

$$d(A, B) = \frac{1}{n_A \cdot n_B} \sum_{i \in A, j \in B} d(i, j)$$

- Método del centróide

Este método se basa en que la distancia entre dos clusters es la distancia (normalmente euclidiana) entre sus centróides. Cada vez que agrupamos dos clusters, calculamos de nuevo su centróide, es decir los centróides de un grupo cambian a medida que se fusionan clusters.

Al ser un método ponderado, tenemos en cuenta el tamaño de los clusters a la hora de calcular la distancia o similitud entre estos.

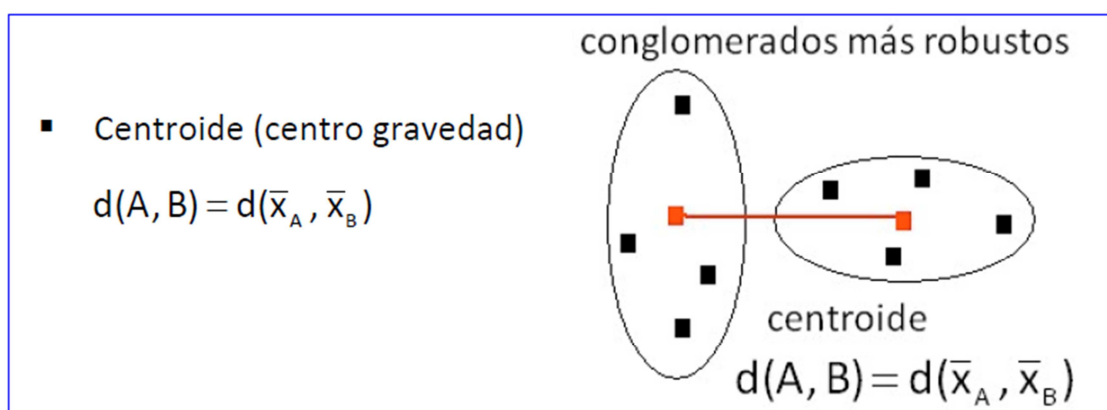
Técnicas de clusterización

Una desventaja de esto, es que si dos clusters son muy diferentes en tamaño, se corre el peligro de que el centroide del clúster resultante este excesivamente influenciado por el componente de tamaño superior y se pierdan las cualidades del grupo pequeño o no se tenga en cuenta prácticamente.

Este problema lo solucionamos con el método de la mediana, el cual es un caso particular del método del centroide en el que consideramos igual el tamaño de todos los clusters, con lo cual no consideramos a la hora de calcular distancia o similitudes entre clusters el tamaño de los mismos.

Una característica importante de estos métodos es que la distancia asociada con los clusters enlazados puede aumentar o disminuir de una etapa a otra, ya que la distancia entre centroides puede ser menor que la de otro par de centroides unidos en una etapa anterior, esto puede ocurrir porque los centroides, en cada etapa, pueden cambiar de lugar. Este problema puede llevar a que el dendograma sea difícil de interpretar.

Siendo A y B dos clusters:



- Método de Ward.

Cuando unimos dos conglomerados, independientemente del método que utilicemos, aumenta la varianza. El método de Ward se basa en la búsqueda de la minimización de la varianza dentro de cada grupo, y uniendo clústeres cuando dicha varianza es mínima.

Para este proceso calculamos en primer lugar la media de todas las varianzas en cada clúster, a continuación, calculamos la distancia entre cada caso y la media del clúster, sumando después la distancia entre todos los casos.

Ahora vemos cuales son los clusters que generan menos aumentos en la suma de las distancias dentro de cada clúster y los vamos agrupando, con lo que vamos creando grupos homogéneos y con tamaños similares, algo característico de este método.

Ward llegó a la conclusión de que la pérdida de información que se produce al unir los distintos individuos en cada clúster se puede medir a través de la suma total de los cuadrados de las desviaciones entre cada individuo y la media del clúster en el que se va a unir.

Para que este procedimiento sea óptimo, en cada paso del proceso consideró la posibilidad de la unión de cada par de grupos y optar por la fusión de aquellos dos grupos que menos incrementen la suma de los cuadrados de las desviaciones al unirse.

Técnicas de clusterización

Este método es uno de los más utilizados en la práctica ya que posee casi todas las ventajas del método de las K-medias (hablaremos más adelante, cuando hablemos de los métodos no jerárquicos) y suele ser más discriminativo en la determinación de los niveles de agrupación. Una investigación realizada por Kuiper y Fisher demostró que este método era capaz de acercarse más a la clasificación óptima que otros métodos.

Sus características son:

- Método muy eficiente.
- Creación de clusters pequeños.
- Podemos usar tanto matriz de distancias como tablas de contingencia.
- Invariante bajo transformaciones monótonas de la matriz de distancias.
- Puede ser sensible a outliers.
- Genera clusters de tamaños similares.

- Lance y Williams

Este método busca agrupar todos los métodos vistos anteriormente bajo una misma fórmula.

Técnicas de clusterización

Siendo dos clusters P y Q que se han agrupado, la distancia de estos con otro clúster R puede calcularse como una función de las distancias entre los 3 clusters de la forma siguiente:

$$d(R, P+Q) = \delta_1 d(R, P) + \delta_2 d(R, Q) + \delta_3 d(P, Q) + \delta_4 |d(R, P) - d(R, Q)|$$

Esta fórmula se referirá a cada uno de los métodos vistos hasta ahora dependiendo del valor que tomen sus constantes de ponderación:

| Método | δ_1 | δ_2 | δ_3 | δ_4 |
|-----------------|-------------------------------------|-------------------------------------|----------------------------------|----------------|
| Salto mínimo | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $-\frac{1}{2}$ |
| Salto máximo | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{2}$ |
| Media | $\frac{n_P}{n_P + n_Q}$ | $\frac{n_Q}{n_P + n_Q}$ | 0 | 0 |
| Centroide | $\frac{n_P}{n_P + n_Q}$ | $\frac{n_Q}{n_P + n_Q}$ | $-\frac{n_P n_Q}{(n_P + n_Q)^2}$ | 0 |
| Mediana | $\frac{1}{2}$ | $\frac{1}{2}$ | $-\frac{1}{4}$ | 0 |
| Ward | $\frac{n_R + n_P}{n_R + n_P + n_Q}$ | $\frac{n_R + n_Q}{n_R + n_P + n_Q}$ | $-\frac{n_R}{n_R + n_P + n_Q}$ | 0 |
| Método Flexible | $\frac{1-\beta}{2}$ | $\frac{1-\beta}{2}$ | β | 0 |

Donde n_R , n_P , n_Q se refieren al número de objetos de cada uno de los clusters y β es un valor arbitrario entre 0 y 1.

Técnicas de clusterización

Todos estos métodos vistos hasta ahora son métodos jerárquicos asociativos, como dijimos anteriormente, dentro de los métodos jerárquicos también se encuentran los disociativos.

Estos métodos constituyen el proceso inverso a los aglomerativos, es decir, comienzan con un conglomerado que engloba todos los objetos de la muestra y a partir de este grupo inicial se van dividiendo en grupos hasta que al final tenemos tantos grupos como objetos hay en la muestra.

La filosofía de estos métodos es la misma que para los métodos aglomerativos, con la salvedad de que ahora se seguirá la estrategia de maximizar las distancias o minimizar las similitudes, ya que ahora buscamos los objetos que sean menos similares para separarlos del resto del conglomerado.

Hasta ahora hemos visto procedimientos jerárquicos mediante los cuales partimos de m objetos hasta llegar a un solo clúster (aglomerativos) o partimos de un solo clúster que vamos dividiendo hasta formar m clusters. Ahora vamos a presentar los métodos no jerárquicos, los cuales están diseñados para clasificar individuos.

Los métodos no jerárquicos están diseñados para clasificar individuos en K clusters, donde K debemos de especificarlo a priori o se determina como parte del proceso. Este tipo de método es conveniente utilizarlo cuando los datos a utilizar son elevados o para afinar una clasificación obtenida previamente mediante algún método jerárquico.

Técnicas de clusterización

En estos procedimientos al contrario que en los jerárquicos, no construimos árboles. En lugar de esto, asignamos los individuos a conglomerados una vez que tenemos el número de conglomerados especificado. Esto es, ahora cuando obtenemos 4 conglomerados, no procede de una combinación de dos conglomerados a partir de una solución de 5 conglomerados, sino que se basa en la búsqueda de la mejor solución de esos 4 conglomerados.

Estos métodos se basan en la idea de elegir una partición inicial de los individuos para posteriormente intercambiar los miembros de estos conglomerados para obtener una partición mejor.

Los diversos algoritmos existentes se diferencian en los métodos que utilizaremos para conseguir esta partición mejor. Estos métodos comienzan con una partición inicial de individuos en K grupos o con un conjunto de puntos iniciales sobre los que se formarán los clusters llamados puntos semillas.

Existen multitud de métodos para asignar los individuos a los diferentes clusters.

- Umbral secuencial: mediante este método seleccionamos un número de puntos semillas y vamos formando conglomerados con los puntos que estén a una distancia 'd' de estos puntos semillas.

Cuando tenemos todos los objetos dentro de las distancias, seleccionamos una segunda semilla e incluimos de nuevo los individuos dentro de la distancia especificada. Cuando incluimos un individuo en un conglomerado, no consideramos este individuo para posteriores semillas.

Técnicas de clusterización

- **Umbral paralelo:** este método selecciona varias semillas de conglomerado simultáneamente al principio y asigna individuos dentro de la distancia umbral hasta la semilla más cercana. En algunas variantes de este método, los individuos permanecen fuera de los conglomerados si están fuera de la distancia previamente especificada desde cualquiera de las semillas de conglomerado.
- **Optimización:** Este método se diferencia de los dos anteriores en que permite la reubicación de los individuos, es decir, si en el curso de asignación de los individuos, uno de ellos se acerca más a otro clúster que no es el que tenía asignado en este momento, el procedimiento de optimización cambiaría al individuo hacia el conglomerado más cercano.

A continuación se muestra un cuadro comparativo de las dos familias principales vistas:

| JERÁRQUICO | NO JERÁRQUICO |
|---|---|
| <ul style="list-style-type: none">▪ <i>No exigen una definición previa del número de conglomerados.</i> | <ul style="list-style-type: none">▪ <i>Exigen definir previamente el número de clusters.</i> |
| <ul style="list-style-type: none">▪ <i>Llevan a cabo un proceso iterativo, de abajo hacia arriba con (n-1) pasos, partiendo de n grupos para terminar en 1 (aglomerativos).</i> | <ul style="list-style-type: none">▪ <i>Poseen algunos índices que indican el número óptimo de conglomerados.</i> |
| <ul style="list-style-type: none">▪ <i>Permite obtener distintos tipos de resultados gráficos y numéricos que facilitan la interpretación de los resultados.</i> | <ul style="list-style-type: none">▪ <i>Proporcionan los valores de los centroides de los grupos, lo que facilita la interpretación.</i> |
| <ul style="list-style-type: none">▪ <i>Precisan una gran cantidad de cálculos, que en ocasiones limita la posibilidad de aplicación con muestras muy grandes.</i> | <ul style="list-style-type: none">▪ <i>Ofrecen resultados adicionales que permiten seleccionar las variables para la interpretación de los conglomerados.</i> |
| <ul style="list-style-type: none">▪ <i>Pueden aplicarse sobre los casos y sobre las variables.</i> | <ul style="list-style-type: none">▪ <i>Sólo pueden aplicarse sobre casos. Dan soluciones de tipo óptimo.</i> |

5.3 Validación e interpretación

La utilización de métodos jerárquicos provoca una estructura sobre nuestros datos que a veces es necesario considerar si dicha estructura es o no aceptable o si se introducen distorsiones inaceptables en las relaciones originales.

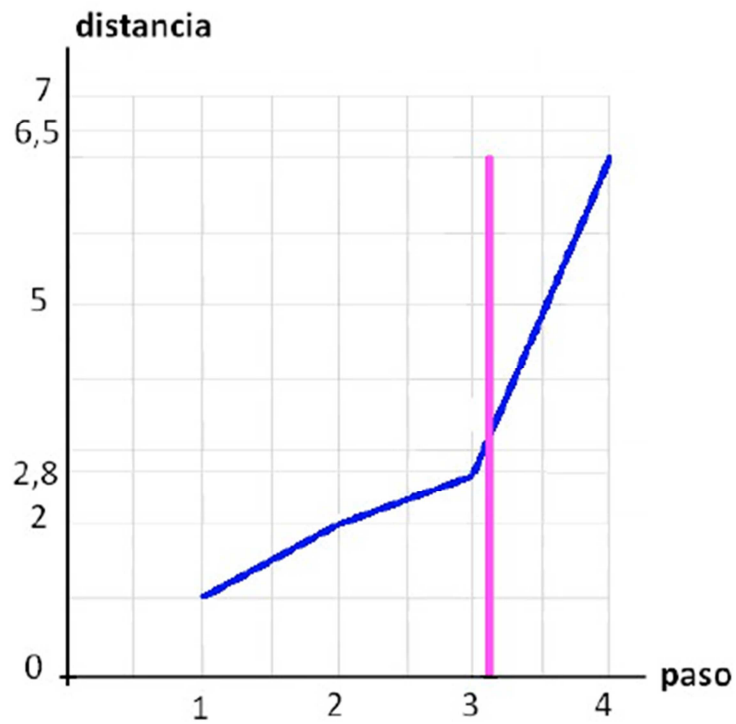
Para esto el método más utilizado es el coeficiente de correlación cofenético, mediante el cual observamos la relación entre el dendograma y la matriz de proximidades original.

Este coeficiente es simplemente la correlación entre los $\frac{n(n-1)}{2}$ elementos de la parte superior de la matriz de proximidades observada y los correspondientes en la llamada matriz cofenética cuyos elementos se definen como aquellos que determina la proximidad entre los elementos i y j cuando estos se unen en un mismo clúster. El coeficiente cofenético interesa que se lo mas elevado posible, siendo siempre menor o igual que 1.

En cuanto al número de clusters a utilizar, para tomar esta decisión se suele representar los distintos pasos del algoritmo y la distancia a la que se produce la fusión.

En los primeros pasos los saltos en las distancias será pequeño, mientras que en los últimos pasos los saltos serán cada vez mayores.

El punto de corte será aquel en el que comiencen a producirse saltos bruscos.,



Por ejemplo, en esta representación de 4 pasos, vemos como se produce un salto brusco o cambio brusco de pendiente del paso 3 al 4, con lo cual el óptimo se produce en el paso 3, ahora solo faltaría ver el número de clusters que tenemos en el tercer paso y este sería el número óptimo.

6 Metodología seleccionada

Nuestro caso en concreto es el estudio de las densidades poblacionales en Andalucía a partir de datos discretos de conteos relacionados con el uso de las redes sociales. Con lo cual crearemos agrupaciones en función de dicha densidad identificando automáticamente cuáles son las áreas de mayor densidad poblacional.

El primer paso dentro del diseño de investigación es decidir sobre la medida de similitud entre los objetos (en nuestro caso las Provincias de Andalucía). Entre las distintas alternativas, he optado por la distancia euclídea dado que el conjunto de variables es de carácter métrico.

Previamente se han explicado los distintos métodos para medir la similitud entre objetos, pero ahora detallaremos analíticamente este método en concreto que será el utilizado en nuestro proyecto.

Vamos a considerar dos individuos (Provincias Andaluzas en nuestro caso concreto) tomados de la población, o lo que es lo mismo dos filas de nuestra matriz X:

$$x_i = (x_{i1}, \dots, x_{in})'$$

$$x_j = (x_{j1}, \dots, x_{jn})'$$

La métrica más conocida que corresponde a la generalización de más de dos dimensiones de la distancia entre dos puntos en el plano, es la derivada de la norma de un vector:

$$\|x_i\|_2 = \sqrt{x_i' x_i} = \sqrt{\sum_{l=1}^n x_{il}^2}$$

Y a partir de ella obtenemos la distancia euclídea:

$$d_2(x_i, x_j) = \|x_i - x_j\|_2 = \sqrt{(x_i - x_j)' (x_i - x_j)} = \sqrt{\sum_{l=1}^n (x_{il} - x_{jl})^2}$$

El segundo paso en el diseño de investigación es la elección del método de aglomeración.

Para este proyecto vamos a proceder a implementar 3 métodos diferentes (distancia mínima, distancia máxima y promedio), ya que como vamos diciendo a lo largo del proyecto, no hay un método que siempre sea el óptimo.

Estos métodos son métodos jerárquicos aglomerativos, si bien ya han sido explicados anteriormente, ahora se detallarán de forma más práctica por ser estos los que implementemos en nuestro proyecto:

- Distancia mínima

Siendo n el número de elementos, tras la etapa K -ésima tenemos formados $n-K$ clusters, y la distancia entre el nuevo cluster formado C_i y el resto C_j será:

$$d(C_i, C_j) = \underset{\substack{x_l \in C_i \\ x_m \in C_j}}{\text{Min}} \{d(x_l, x_m)\} \quad l = 1, \dots, n_i ; m = 1, \dots, n_j$$

- Distancia máxima

Siendo n el número de elementos, tras la etapa K -ésima tenemos formados $n-K$ clusters, y la distancia entre el nuevo cluster formado C_i y el resto C_j será:

$$d(C_i, C_j) = \underset{\substack{x_l \in C_i \\ x_m \in C_j}}{\text{Max}} \{d(x_l, x_m)\} \quad l = 1, \dots, n_i ; m = 1, \dots, n_j$$

- Promedio

Llamamos al cluster nuevo formado C_i , este a su vez se formó a partir de 2 clusters que llamaremos C_{i_1} y C_{i_2} . La distancia del cluster C_i con el resto C_j será:

$$d(C_i, C_j) = \frac{d(C_{i_1}, C_j) + d(C_{i_2}, C_j)}{2}$$

Técnicas de clusterización

Con esto obtenemos las distancias entre los clusters que se van formando, que como hemos visto depende del método que estemos implementando, y tendremos distancias distintas en función del método implementado.

Para decidir que Clusters se unen, el procedimiento es común a todos ellos al tratarse de métodos aglomerativos, con lo cual vamos uniendo los clusters cuya distancia entre ellos sea la menor.

Una vez hemos implementado los 3 métodos, calculamos el coeficiente de correlación cofenético ya explicado anteriormente, y nos quedamos con la clusterización del método cuyo coeficiente sea mayor.

Para finalizar solo nos queda decidir el número de Clusters, para tomar esta decisión representamos los distintos pasos del algoritmo y la distancia a la que se produce la fusión. Cuando observemos que existe un salto brusco (pendiente elevada), paramos el proceso y nos quedamos con el número de clusters formados hasta ese paso.

7 Aplicación y Resultados

En este proyecto vamos a clasificar las Provincias de Andalucía según el número de usuarios de redes sociales. Para ello nuestros datos serán el número de usuarios activos de Facebook y Twitter.

| | Facebook | Twitter |
|---------|----------|---------|
| Almería | 63.200 | 31.200 |
| Cádiz | 110.500 | 60.100 |
| Córdoba | 79.300 | 43.200 |
| Granada | 105.200 | 57.900 |
| Huelva | 49.100 | 24.700 |
| Jaén | 42.300 | 19.100 |
| Málaga | 185.200 | 93.200 |
| Sevilla | 153.500 | 82.900 |

La matriz X que contiene los valores de las provincias nos queda:

$$X = \begin{pmatrix} 63.200 & 31.200 \\ 110.500 & 60.100 \\ 79.300 & 43.200 \\ 105.200 & 57.900 \\ 49.100 & 24.700 \\ 42.300 & 19.100 \\ 185.200 & 93.200 \\ 153.500 & 82.900 \end{pmatrix}$$

Técnicas de clusterización

Siendo las filas cada una de las provincias y las columnas correspondientes al número de usuarios de Facebook y Twitter respectivamente.

Ahora pasamos a calcular la matriz de similitud mediante la distancia euclídea, por ejemplo para el caso Almería-Cádiz tenemos:

$$\sqrt{(63.200 - 110.500)^2 + (31.200 - 60.100)^2} = 55.430,14$$

Para el resto:

Almería-Córdoba:

$$\sqrt{(63.200 - 79.300)^2 + (31.200 - 43.200)^2} = 20.080,09$$

Almería-Granada:

$$\sqrt{(63.200 - 105.200)^2 + (31.200 - 57.900)^2} = 49.768,36$$

Almería-Huelva:

$$\sqrt{(63.200 - 49.100)^2 + (31.200 - 24.700)^2} = 15.526,11$$

Almería-Jaén:

$$\sqrt{(63.200 - 42.300)^2 + (31.200 - 19.100)^2} = 24.149,95$$

Almería-Málaga:

$$\sqrt{(63.200 - 185.200)^2 + (31.200 - 93.200)^2} = 136.850,28$$

Almería-Sevilla:

$$\sqrt{(63.200 - 153.500)^2 + (31.200 - 82.900)^2} = 104.052,78$$

Cádiz-Córdoba:

$$\sqrt{(110.500 - 79.300)^2 + (60.100 - 43.200)^2} = 35.483,09$$

Cádiz-Granada:

$$\sqrt{(110.500 - 105.200)^2 + (60.100 - 57.900)^2} = 5.738,47$$

Cádiz-Huelva:

$$\sqrt{(110.500 - 49.100)^2 + (60.100 - 24.700)^2} = 70.873,97$$

Cádiz-Jaén:

$$\sqrt{(110.500 - 42.300)^2 + (60.100 - 19.100)^2} = 79.575,37$$

Cádiz-Málaga:

$$\sqrt{(110.500 - 185.200)^2 + (60.100 - 93.200)^2} = 81.704,96$$

Cádiz-Sevilla:

$$\sqrt{(110.500 - 153.500)^2 + (60.100 - 82.900)^2} = 48.670,73$$

Córdoba-Granada:

$$\sqrt{(79.300 - 105.200)^2 + (43.200 - 57.900)^2} = 29.780,87$$

Córdoba-Huelva:

$$\sqrt{(79.300 - 49.100)^2 + (43.200 - 24.700)^2} = 35.415,96$$

Córdoba-Jaén:

$$\sqrt{(79.300 - 42.300)^2 + (43.200 - 19.100)^2} = 44.156,65$$

Córdoba-Málaga:

$$\sqrt{(79.300 - 185.200)^2 + (43.200 - 93.200)^2} = 117.110,25$$

Córdoba-Sevilla:

$$\sqrt{(79.300 - 153.500)^2 + (43.200 - 82.900)^2} = 84.153,02$$

Granada-Huelva:

$$\sqrt{(105.200 - 49.100)^2 + (57.900 - 24.700)^2} = 65.187,81$$

Granada-Jaén:

$$\sqrt{(105.200 - 42.300)^2 + (57.900 - 19.100)^2} = 73.904,33$$

Granada-Málaga:

$$\sqrt{(105.200 - 185.200)^2 + (57.900 - 93.200)^2} = 87.441,92$$

Granada-Sevilla:

$$\sqrt{(105.200 - 153.500)^2 + (57.900 - 82.900)^2} = 54.386,49$$

Huelva-Jaén:

$$\sqrt{(49.100 - 42.300)^2 + (24.700 - 19.100)^2} = 8.809,09$$

Huelva-Málaga:

$$\sqrt{(49.100 - 185.200)^2 + (24.700 - 93.200)^2} = 152.366,20$$

Huelva-Sevilla:

$$\sqrt{(49.100 - 153.500)^2 + (24.700 - 82.900)^2} = 119.526,57$$

Jaén-Málaga:

$$\sqrt{(42.300 - 185.200)^2 + (19.100 - 93.200)^2} = 160.969,62$$

Jaén-Sevilla:

$$\sqrt{(42.300 - 153.500)^2 + (19.100 - 82.900)^2} = 128.202,50$$

Málaga-Sevilla:

$$\sqrt{(185.200 - 153.500)^2 + (93.200 - 82.900)^2} = 33.331,37$$

Técnicas de clusterización

Nos queda:

| | Almería | Cádiz | Córdoba | Granada | Huelva | Jaén | Málaga | Sevilla |
|---------|------------|-----------|------------|-----------|------------|------------|-----------|---------|
| Almería | 0,00 | | | | | | | |
| Cádiz | 55.430,14 | 0,00 | | | | | | |
| Córdoba | 20.080,09 | 35.483,09 | 0,00 | | | | | |
| Granada | 49.768,36 | 5.738,47 | 29.780,87 | 0,00 | | | | |
| Huelva | 15.526,11 | 70.873,97 | 35.415,96 | 65.187,81 | 0,00 | | | |
| Jaén | 24.149,95 | 79.575,37 | 44.156,65 | 73.904,33 | 8.809,09 | 0,00 | | |
| Málaga | 136.850,28 | 81.704,96 | 117.110,25 | 87.441,92 | 152.366,20 | 160.969,62 | 0,00 | |
| Sevilla | 104.052,78 | 48.670,73 | 84.153,02 | 54.386,49 | 119.526,57 | 128.202,50 | 33.331,37 | 0,00 |

Esta matriz de distancia inicial es común a los 3 métodos que a continuación vamos a implementar.

- Método de la distancia mínima

Para simplificar etiquetaremos a cada Provincia con sus dos primeras letras:

Almería: al

Cádiz: ca

Técnicas de clusterización

Córdoba: co

Granada: gr

Huelva: hu

Jaén: ja

Málaga: ma

Sevilla: se

En la matriz de distancia inicial vemos que la menor distancia corresponde a la pareja Cádiz-Granada (5.738,47).

El primer clúster que se forma es Cádiz-Granada, para calcular la nueva matriz de distancias obtenemos las distancias a este nuevo clúster.

Las distancias del nuevo clúster (Cádiz-Granada) con el resto de provincias se obtiene mediante la mínima distancia de este con los demás, así por ejemplo para la distancia del clúster (Cádiz-Granada) con el clúster Córdoba, tenemos:

$$D [(ca-gr)-co] = \min (35.483,09; 29.780,87) = 29.780,87$$

Para el resto:

$$D [(ca-gr)-hu] = \min (70.873,97; 65.187,81) = 65.187,81$$

$$D [(ca-gr)-ja] = \min (79.575,37; 73.904,33) = 73.904,33$$

$$D [(ca-gr)-ma] = \min (81.704,96; 87.441,92) = 81.704,33$$

Técnicas de clusterización

$$D [(ca-gr)-se] = \min (48.670,73; 54.386,49) = 48.704,73$$

$$D [(ca-gr)-al] = \min (55.430,14; 49.768,36) = 49.768,36$$

Nos queda la matriz de distancias:

| | al | (ca-gr) | co | hu | ja | ma | se |
|---------|------------|-----------|------------|------------|------------|-----------|------|
| al | 0,00 | | | | | | |
| (ca-gr) | 49.768,36 | 0,00 | | | | | |
| co | 20.080,09 | 29.780,87 | 0,00 | | | | |
| hu | 15.526,11 | 65.187,81 | 35.415,96 | 0,00 | | | |
| ja | 24.149,95 | 73.904,33 | 44.156,65 | 8.809,09 | 0,00 | | |
| ma | 136.850,28 | 81.704,96 | 117.110,25 | 152.366,20 | 160.969,62 | 0,00 | |
| se | 104.052,78 | 48.670,73 | 84.153,02 | 119.526,57 | 128.202,50 | 33.331,37 | 0,00 |

Ahora para esta nueva matriz de distancias, la distancia mínima corresponde a Huelva-Jaén (8.809,09).

Formamos este nuevo clúster y calculamos la matriz de distancias:

$$D [(hu-ja)-al] = \min (15.526,11; 24.149,95) = 15.526,11$$

Técnicas de clusterización

$$D [(hu-ja)-(ca-gr)] = \min (65.187,81; 73.904,33) = 65.187,81$$

$$D [(hu-ja)-co] = \min (35.415,96; 44.156,65) = 35.415,96$$

$$D [(hu-ja)-ma] = \min (152.366,20; 160.969,62) = 152.366,20$$

$$D [(hu-ja)-se] = \min (119.526,57; 128.205,50) = 119.526,57$$

Nos queda:

| | al | (ca-gr) | co | (hu-ja) | ma | se |
|---------|------------|-----------|------------|------------|-----------|------|
| al | 0,00 | | | | | |
| (ca-gr) | 49.768,36 | 0,00 | | | | |
| co | 20.080,09 | 29.780,87 | 0,00 | | | |
| (hu-ja) | 15.526,11 | 65.187,81 | 35.415,96 | 0,00 | | |
| ma | 136.850,28 | 81.704,96 | 117.110,25 | 152.366,20 | 0,00 | |
| se | 104.052,78 | 48.670,73 | 84.153,02 | 119.526,57 | 33.331,37 | 0,00 |

Se forma el clúster Almería-Huelva-Jaén (15.526,11).

Para este nuevo clúster calculamos la matriz de distancias:

Técnicas de clusterización

$$D [(al-(hu-ja))-(ca-gr)] = \min (65.187,81; 49.768,36) = 49.768,36$$

$$D [(al-(hu-ja))-co] = \min (35.415,96; 20.080,09) = 20.080,09$$

$$D [(al-(hu-ja))-ma] = \min (136.850,28; 152.366,20) = 136.850,28$$

$$D [(al-(hu-ja))-se] = \min (104.052,78; 119.526,57) = 104.052,78$$

Matriz de distancias:

| | (al-(hu-ja)) | (ca-gr) | co | ma | se |
|--------------|--------------|-----------|------------|-----------|------|
| (al-(hu-ja)) | 0,00 | | | | |
| (ca-gr) | 49.768,36 | 0,00 | | | |
| co | 20.080,09 | 29.780,87 | 0,00 | | |
| ma | 136.850,28 | 81.704,96 | 117.110,25 | 0,00 | |
| se | 104.052,78 | 48.670,73 | 84.153,02 | 33.331,37 | 0,00 |

Se forma el clúster Almería-Huelva-Jaén-Córdoba (20.080,09).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [((al-(hu-ja))-co)-(ca-gr)] = \min (49.768,36; 29.780,87) = 29.780,87$$

Técnicas de clusterización

$$D [((\text{al}-(\text{hu-ja}))-\text{co})-\text{ma}] = \min (136.850,28; 117.110,25) = 117.110,25$$

$$D [((\text{al}-(\text{hu-ja}))-\text{co})-\text{se}] = \min (104.052,78; 84.153,02) = 84.153,02$$

Matriz de distancias:

| | ((al-(hu-ja))-co) | (ca-gr) | ma | se |
|-------------------|-------------------|-----------|-----------|------|
| ((al-(hu-ja))-co) | 0,00 | | | |
| (ca-gr) | 29.780,87 | 0,00 | | |
| ma | 117.110,25 | 81.704,96 | 0,00 | |
| se | 84.153,02 | 48.670,73 | 33.331,37 | 0,00 |

Se forma el clúster Almería-Huelva-Jaén-Córdoba-Cádiz-Granada (29.780,87).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [(((\text{al}-(\text{hu-ja}))-\text{co})-(\text{ca-gr}))-\text{ma}] = \min (117.110,25; 81.704,96) = 81.704,96$$

$$D [(((\text{al}-(\text{hu-ja}))-\text{co})-(\text{ca-gr}))-\text{se}] = \min (84.153,02; 48.670,73) = 48.670,73$$

Matriz de distancias:

Técnicas de clusterización

| | (((al-(hu-ja))-co)-(ca-gr)) | ma | se |
|-----------------------------|-----------------------------|-----------|------|
| (((al-(hu-ja))-co)-(ca-gr)) | 0,00 | | |
| ma | 81.704,96 | 0,00 | |
| se | 48.670,73 | 33.331,37 | 0,00 |

Se forma el clúster Málaga-Sevilla (33.331,37).

Para este nuevo clúster calculamos la matriz de distancias:

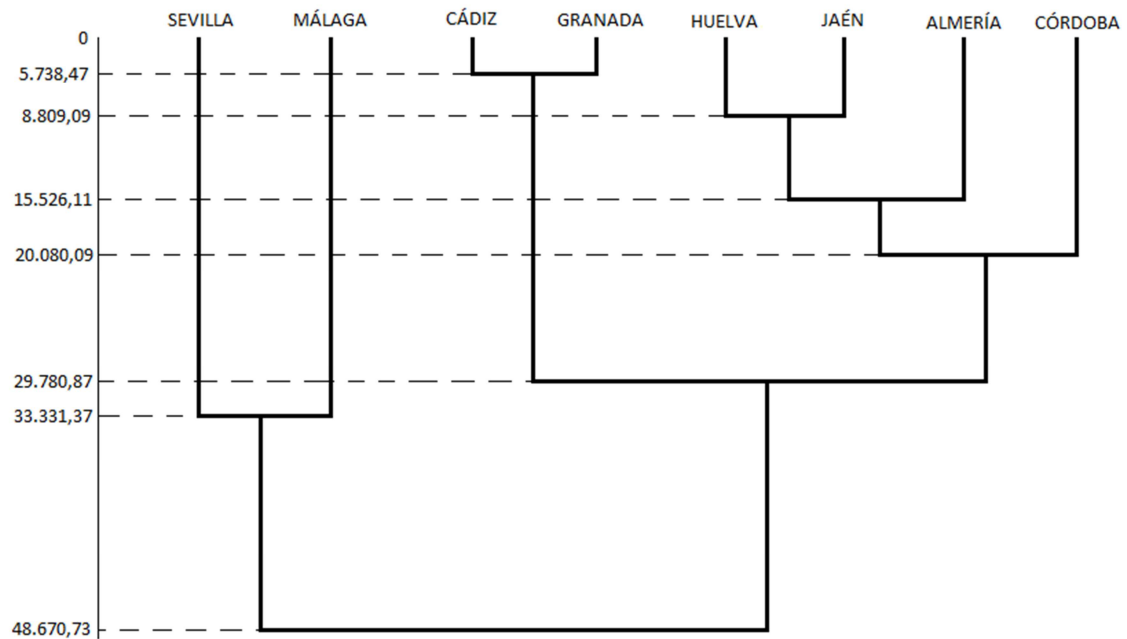
$$D [(ma-se)-(((al-(hu-ja))-co)-(ca-gr))] = \min (81.704,96; 48.670,73) = 48.670,73$$

Matriz de distancias:

| | (((al-(hu-ja))-co)-(ca-gr)) | (ma-se) |
|-----------------------------|-----------------------------|---------|
| (((al-(hu-ja))-co)-(ca-gr)) | 0,00 | |
| (ma-se) | 48.670,73 | 0,00 |

Y finalmente se unen todas las provincias en un único clúster a distancia 48.670,73.

Representación gráfica del proceso (Dendograma):



- Método de la distancia máxima

Ahora para calcular las distancias del cluster formado con el resto, aplicamos el criterio de la distancia máxima.

De la matriz de distancias inicial vemos que se unen Cádiz-Granada (5.738,47).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [(ca-gr)-co] = \max (35.483,09; 29.780,87) = 35.483,09$$

Técnicas de clusterización

$$D [(ca-gr)-hu] = \max (70.873,97; 65.187,81) = 70.873,97$$

$$D [(ca-gr)-ja] = \max (79.575,37; 73.904,33) = 79.575,37$$

$$D [(ca-gr)-ma] = \max (81.704,96; 87.441,92) = 87.441,92$$

$$D [(ca-gr)-se] = \min (48.670,73; 54.386,49) = 54.386,49$$

$$D [(ca-gr)-al] = \max (55.430,14; 49.768,36) = 55.430,14$$

Matriz de distancias:

| | al | (ca-gr) | co | hu | ja | ma | se |
|---------|------------|-----------|------------|------------|------------|-----------|------|
| al | 0,00 | | | | | | |
| (ca-gr) | 55.430,14 | 0,00 | | | | | |
| co | 20.080,09 | 35.483,09 | 0,00 | | | | |
| hu | 15.526,11 | 70.873,97 | 35.415,96 | 0,00 | | | |
| ja | 24.149,95 | 79.575,37 | 44.156,65 | 8.809,09 | 0,00 | | |
| ma | 136.850,28 | 87.441,92 | 117.110,25 | 152.366,20 | 160.969,62 | 0,00 | |
| se | 104.052,78 | 54.386,49 | 84.153,02 | 119.526,57 | 128.202,50 | 33.331,37 | 0,00 |

Técnicas de clusterización

Se forma el clúster Huelva-Jaén (8.809,09).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [(hu-ja)-al] = \max (15.526,11; 24.149,95) = 24.149,95$$

$$D [(hu-ja)-(ca-gr)] = \max (70.575,37; 79.575,37) = 79.575,37$$

$$D [(hu-ja)-co] = \max (35.415,96; 44.156,65) = 44.156,65$$

$$D [(hu-ja)-ma] = \max (152.366,20; 160.969,62) = 160.969,62$$

$$D [(hu-ja)-se] = \max (119.526,57; 128.202,50) = 128.202,50$$

Matriz de distancias:

| | al | (ca-gr) | co | (hu-ja) | ma | se |
|---------|------------|-----------|------------|------------|-----------|------|
| al | 0,00 | | | | | |
| (ca-gr) | 55.430,14 | 0,00 | | | | |
| co | 20.080,09 | 35.483,09 | 0,00 | | | |
| (hu-ja) | 24.149,95 | 79.575,37 | 44.156,65 | 0,00 | | |
| ma | 136.850,28 | 87.441,92 | 117.110,25 | 160.969,62 | 0,00 | |
| se | 104.052,78 | 54.386,49 | 84.153,02 | 128.202,50 | 33.331,37 | 0,00 |

Técnicas de clusterización

Se forma el clúster Almería-Córdoba (20.080,09).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [(al-co)-(ca-gr)] = \max (55.430,14; 35.483,09) = 55.430,14$$

$$D [(al-co)-(hu-ja)] = \max (24.149,95; 44.156,65) = 44.156,65$$

$$D [(al-co)-ma] = \max (136.850,28; 117.110,25) = 136.850,28$$

$$D [(al-co)-se] = \max (104.052,78; 84.153,02) = 104.052,78$$

Matriz de distancias:

| | (al-co) | (ca-gr) | (hu-ja) | ma | se |
|---------|------------|-----------|------------|-----------|------|
| (al-co) | 0,00 | | | | |
| (ca-gr) | 55.430,14 | 0,00 | | | |
| (hu-ja) | 44.156,65 | 79.575,37 | 0,00 | | |
| ma | 136.850,28 | 87.441,92 | 160.969,62 | 0,00 | |
| se | 104.052,78 | 54.386,49 | 128.202,50 | 33.331,37 | 0,00 |

Técnicas de clusterización

Se forma el clúster Málaga-Sevilla (33.331,37).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [(ma-se)-(al-co)] = \max (136.850,28; 104.052,78) = 136.850,28$$

$$D [(ma-se)-(ca-gr)] = \max (87.441,92; 54.386,49) = 87.441,92$$

$$D [(ma-se)-(hu-ja)] = \max (160.969,62; 128.202,50) = 160.969,62$$

Matriz de distancias:

| | (al-co) | (ca-gr) | (hu-ja) | (ma-se) |
|---------|------------|-----------|------------|---------|
| (al-co) | 0,00 | | | |
| (ca-gr) | 55.430,14 | 0,00 | | |
| (hu-ja) | 44.156,65 | 79.575,37 | 0,00 | |
| (ma-se) | 136.850,28 | 87.441,92 | 160.969,62 | 0,00 |

Se forma el clúster Almería-Córdoba-Huelva-Jaén (44.156,65).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [((al-co)-(hu-ja))-(ca-gr)] = \max (55.430,14; 79.575,37) = 79.575,37$$

Técnicas de clusterización

$$D [((al-co)-(hu-ja))-(ma-se)] = \max (136.850,28; 160.969,62) = 160.969,62$$

Matriz de distancias:

| | ((al-co)-(hu-ja)) | (ca-gr) | (ma-se) |
|-------------------|-------------------|-----------|---------|
| ((al-co)-(hu-ja)) | 0,00 | | |
| (ca-gr) | 79.575,37 | 0,00 | |
| (ma-se) | 160.969,62 | 87.441,92 | 0,00 |

Se forma el clúster Almería-Córdoba-Huelva-Jaén-Cádiz-Granada (79.575,37).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [(((al-co)-(hu-ja))-(ca-gr))-(ma-se)] = \max (160.969,62; 87.441,92) = 160.969,62$$

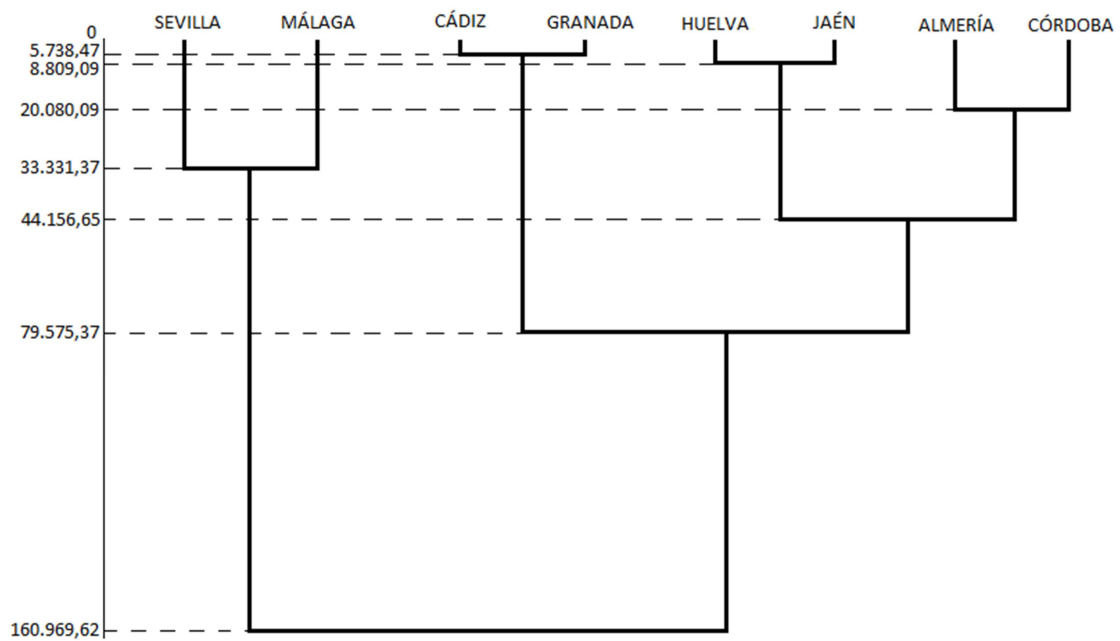
Matriz de distancias:

| | (((al-co)-(hu-ja))-(ca-gr)) | (ma-se) |
|-----------------------------|-----------------------------|---------|
| (((al-co)-(hu-ja))-(ca-gr)) | 0,00 | |
| (ma-se) | 160.969,62 | 0,00 |

Técnicas de clusterización

Y finalmente se unen todas las provincias en un único clúster a distancia 160.969,62.

Representación gráfica del proceso (Dendograma):



- Método del Promedio

Ahora para calcular las distancias del cluster formado con el resto, aplicamos el criterio del promedio.

De la matriz de distancias inicial vemos que se unen Cádiz-Granada (5.738,47).

Para este nuevo clúster calculamos la matriz de distancias, ahora estas distancias serán el promedio:

$$D [(ca-gr)-al] = \frac{55.430,14 + 49.768,36}{2} = 52.599,25$$

$$D [(ca-gr)-co] = \frac{35.483,09 + 29.780,87}{2} = 32.631,98$$

$$D [(ca-gr)-hu] = \frac{70.873,97 + 65.187,81}{2} = 68.030,98$$

$$D [(ca-gr)-ja] = \frac{79.575,37 + 73.904,33}{2} = 76.739,85$$

$$D [(ca-gr)-ma] = \frac{81.704,96 + 87.441,92}{2} = 84.573,44$$

$$D [(ca-gr)-se] = \frac{48.670,73 + 54.386,49}{2} = 51.528,61$$

Matriz de distancias:

Técnicas de clusterización

| | al | (ca-gr) | co | hu | ja | ma | se |
|---------|------------|-----------|------------|------------|------------|-----------|------|
| al | 0,00 | | | | | | |
| (ca-gr) | 52.599,25 | 0,00 | | | | | |
| co | 20.080,09 | 32.631,98 | 0,00 | | | | |
| hu | 15.526,11 | 68.030,89 | 35.415,96 | 0,00 | | | |
| ja | 24.149,95 | 76.739,85 | 44.156,65 | 8.809,09 | 0,00 | | |
| ma | 136.850,28 | 84.573,44 | 117.110,25 | 152.366,20 | 160.969,62 | 0,00 | |
| se | 104.052,78 | 51.528,61 | 84.153,02 | 119.526,57 | 128.202,50 | 33.331,37 | 0,00 |

Se forma el clúster Huelva-Jaén (8.809,09).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [(hu-ja)-al] = \frac{15.526,11 + 24.149,95}{2} = 19.838,03$$

$$D [(hu-ja)-(ca-gr)] = \frac{68.030,89 + 76.739,85}{2} = 72.385,37$$

$$D [(hu-ja)-co] = \frac{35.415,96 + 44.156,65}{2} = 39.786,30$$

$$D [(hu-ja)-ma] = \frac{152.366,20 + 160.969,62}{2} = 156.667,91$$

Técnicas de clusterización

$$D [(hu-ja)-se] = \frac{119.526,57 + 128.202,50}{2} = 123.864,53$$

Matriz de distancias:

| | al | (ca-gr) | co | (hu-ja) | ma | se |
|---------|------------|-----------|------------|------------|-----------|------|
| al | 0,00 | | | | | |
| (ca-gr) | 52.599,25 | 0,00 | | | | |
| co | 20.080,09 | 32.631,98 | 0,00 | | | |
| (hu-ja) | 19.838,03 | 72.385,37 | 39.786,30 | 0,00 | | |
| ma | 136.850,28 | 84.573,44 | 117.110,25 | 156.667,91 | 0,00 | |
| se | 104.052,78 | 51.528,61 | 84.153,02 | 123.864,53 | 33.331,37 | 0,00 |

Se forma el clúster Almería-Huelva-Jaén (19.838,03).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [(al-(hu-ja))-(ca-gr)] = \frac{52.599,25 + 72.385,37}{2} = 62.492,31$$

$$D [(al-(hu-ja))-co] = \frac{39.786,30 + 20.080,09}{2} = 29.933,20$$

$$D [(al-(hu-ja))-ma] = \frac{136.850,28 + 156.667,91}{2} = 146.759,10$$

Técnicas de clusterización

$$D [(al-(hu-ja))-se] = \frac{104.052,78 + 123.864,53}{2} = 113.958,65$$

Matriz de distancias:

| | (al-(hu-ja)) | (ca-gr) | co | ma | se |
|--------------|--------------|-----------|------------|-----------|------|
| (al-(hu-ja)) | 0,00 | | | | |
| (ca-gr) | 62.492,31 | 0,00 | | | |
| co | 29.933,20 | 32.631,98 | 0,00 | | |
| ma | 146.759,10 | 84.573,44 | 117.110,25 | 0,00 | |
| se | 113.958,65 | 51.528,61 | 84.153,02 | 33.331,37 | 0,00 |

Se forma el clúster Córdoba-Almería-Huelva-Jaén (29.933,20).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [(co-(al-(hu-ja)))-(ca-gr)] = \frac{32.631,98 + 62.492,31}{2} = 47.562,15$$

$$D [(co-(al-(hu-ja)))-ma] = \frac{146.759,10 + 117.110,25}{2} = 131.934,67$$

$$D [(co-(al-(hu-ja)))-se] = \frac{113.958,65 + 84.153,02}{2} = 99.055,83$$

Técnicas de clusterización

Matriz de distancias:

| | (co-(al-(hu-ja))) | (ca-gr) | ma | se |
|-------------------|-------------------|-----------|-----------|------|
| (co-(al-(hu-ja))) | 0,00 | | | |
| (ca-gr) | 47.562,15 | 0,00 | | |
| ma | 131.934,67 | 84.573,44 | 0,00 | |
| se | 99.055,83 | 51.528,61 | 33.331,37 | 0,00 |

Se forma el clúster Málaga-Sevilla (33.331,37).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [(ma-se)- (co-(al-(hu-ja)))] = \frac{131.934,67 + 99.055,83}{2} = 115.495,25$$

$$D [(ma-se)- (ca-gr)] = \frac{84.573,44 + 51.528,61}{2} = 68.051,02$$

Matriz de distancias:

| | (co-(al-(hu-ja))) | (ca-gr) | (ma-se) |
|-------------------|-------------------|-----------|---------|
| (co-(al-(hu-ja))) | 0,00 | | |
| (ca-gr) | 47.562,15 | 0,00 | |
| (ma-se) | 115.495,25 | 68.051,02 | 0,00 |

Técnicas de clusterización

Se forma el clúster Cádiz-Granada- Córdoba-Almería-Huelva-Jaén (47.562,15).

Para este nuevo clúster calculamos la matriz de distancias:

$$D [((ca-gr)-(co-(al-(hu-ja))))-(ma-se)] = \frac{115.495,25 + 68.051,02}{2} = 91.773,14$$

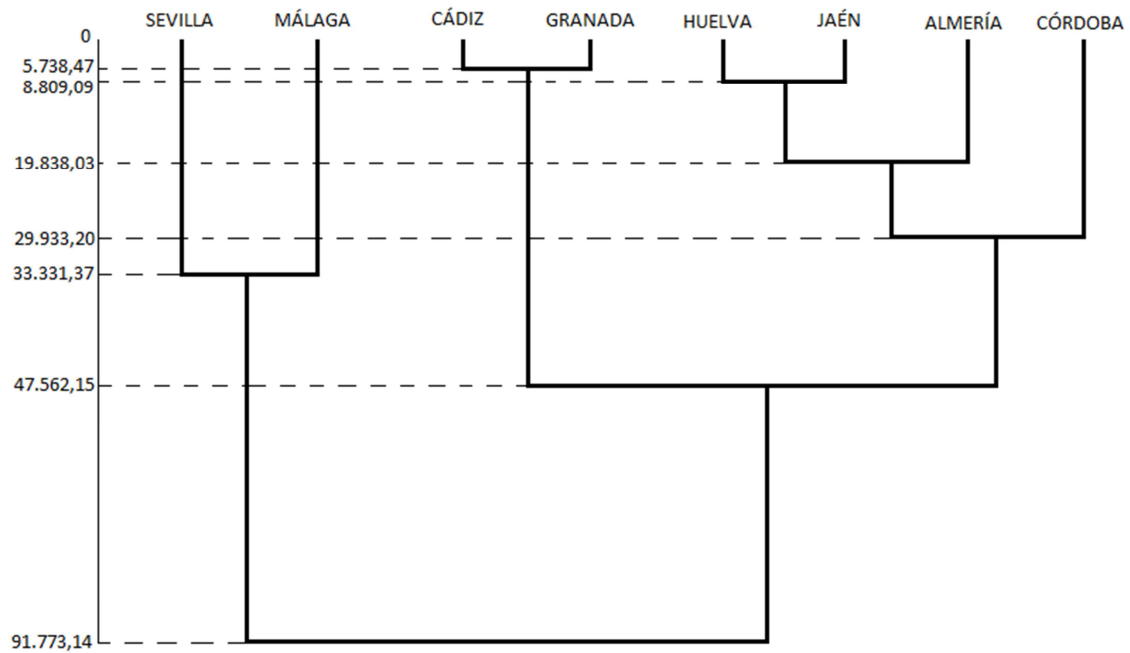
Matriz de distancias:

| | ((ca-gr)-(co-(al-(hu-ja)))) | (ma-se) |
|-----------------------------|-----------------------------|---------|
| ((ca-gr)-(co-(al-(hu-ja)))) | 0,00 | |
| (ma-se) | 91.773,14 | 0,00 |

Y finalmente se unen todas las provincias en un único clúster a distancia 91.773,14.

Técnicas de clusterización

Representación gráfica del proceso (Dendograma):



Observamos que durante el proceso, se han ido agrupando las Provincias de forma diferente según el método utilizado.

Para escoger uno de los 3 métodos usaremos el coeficiente copenético, ya explicado anteriormente.

Valores altos del coeficiente copenético indica que durante el proceso no ha ocurrido una gran perturbación en lo que se refiere a la estructura original de los datos.

Técnicas de clusterización

Mediante este coeficiente observamos la relación entre el dendograma (matriz cofenética) y la matriz de distancias inicial.

Este coeficiente es simplemente la correlación entre los $\frac{n(n-1)}{2}$ elementos de la matriz de distancias inicial y la matriz cofenética cuyos elementos se definen como aquellos que determina la proximidad entre los elementos i y j cuando estos se unen en un mismo clúster.

Con lo cual, tenemos que el coeficiente de correlación es:

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

Siendo:

$$x = X - X^*$$

$$y = Y - Y^*$$

X = valores de la matriz de distancias iniciales

X^* = valor medio de los elementos de la matriz de distancias iniciales

Y = valores de la matriz cofenética

Y^* = valor medio de los elementos de la matriz cofenética

Técnicas de clusterización

El coeficiente cofenético interesa que sea lo más elevado posible, siendo siempre menor o igual que 1.

La matriz de distancias inicial, vista anteriormente, es común a los 3 métodos:

| | Almería | Cádiz | Córdoba | Granada | Huelva | Jaén | Málaga | Sevilla |
|---------|------------|-----------|------------|-----------|------------|------------|-----------|---------|
| Almería | 0,00 | | | | | | | |
| Cádiz | 55.430,14 | 0,00 | | | | | | |
| Córdoba | 20.080,09 | 35.483,09 | 0,00 | | | | | |
| Granada | 49.768,36 | 5.738,47 | 29.780,87 | 0,00 | | | | |
| Huelva | 15.526,11 | 70.873,97 | 35.415,96 | 65.187,81 | 0,00 | | | |
| Jaén | 24.149,95 | 79.575,37 | 44.156,65 | 73.904,33 | 8.809,09 | 0,00 | | |
| Málaga | 136.850,28 | 81.704,96 | 117.110,25 | 87.441,92 | 152.366,20 | 160.969,62 | 0,00 | |
| Sevilla | 104.052,78 | 48.670,73 | 84.153,02 | 54.386,49 | 119.526,57 | 128.202,50 | 33.331,37 | 0,00 |

La matriz cofenética si difiere según el método implementado, para calcularla nos ayudamos del dendograma de cada uno de los métodos. Los valores de dicha matriz cofenética son las distancias entre los elementos i y j cuando estos se unen en un mismo clúster.

Técnicas de clusterización

- Distancia mínima

| | Almería | Cádiz | Córdoba | Granada | Huelva | Jaén | Málaga | Sevilla |
|---------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---------|
| Almería | 0,00 | | | | | | | |
| Cádiz | 29.780,87 | 0,00 | | | | | | |
| Córdoba | 20.080,09 | 29.780,87 | 0,00 | | | | | |
| Granada | 29.780,87 | 5.738,47 | 29.780,87 | 0,00 | | | | |
| Huelva | 15.526,11 | 29.780,87 | 20.080,09 | 29.780,87 | 0,00 | | | |
| Jaén | 15.526,11 | 29.780,87 | 20.080,09 | 29.780,87 | 8.809,09 | 0,00 | | |
| Málaga | 48.670,73 | 48.670,73 | 48.670,73 | 48.670,73 | 48.670,73 | 48.670,73 | 0,00 | |
| Sevilla | 48.670,73 | 48.670,73 | 48.670,73 | 48.670,73 | 48.670,73 | 48.670,73 | 33.331,37 | 0,00 |

$$r_{min} = 0,807890649$$

Técnicas de clusterización

- Distancia máxima

| | Almería | Cádiz | Córdoba | Granada | Huelva | Jaén | Málaga | Sevilla |
|---------|------------|------------|------------|------------|------------|------------|-----------|---------|
| Almería | 0,00 | | | | | | | |
| Cádiz | 79.575,37 | 0,00 | | | | | | |
| Córdoba | 20.080,09 | 79.575,37 | 0,00 | | | | | |
| Granada | 79.575,37 | 5.738,47 | 79.575,37 | 0,00 | | | | |
| Huelva | 44.156,65 | 79.575,37 | 44.156,65 | 79.575,37 | 0,00 | | | |
| Jaén | 44.156,65 | 79.575,37 | 44.156,65 | 79.575,37 | 8.809,09 | 0,00 | | |
| Málaga | 160.969,62 | 160.969,62 | 160.969,62 | 160.969,62 | 160.969,62 | 160.969,62 | 0,00 | |
| Sevilla | 160.969,62 | 160.969,62 | 160.969,62 | 160.969,62 | 160.969,62 | 160.969,62 | 33.331,37 | 0,00 |

$$r_{max} = 0,818763797$$

Técnicas de clusterización

- Promedio

| | Almería | Cádiz | Córdoba | Granada | Huelva | Jaén | Málaga | Sevilla |
|---------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---------|
| Almería | 0,00 | | | | | | | |
| Cádiz | 47.562,15 | 0,00 | | | | | | |
| Córdoba | 29.933,20 | 47.562,15 | 0,00 | | | | | |
| Granada | 47.562,15 | 5.738,47 | 47.562,15 | 0,00 | | | | |
| Huelva | 19.838,03 | 47.562,15 | 29.933,20 | 47.562,15 | 0,00 | | | |
| Jaén | 19.838,03 | 47.562,15 | 29.933,20 | 47.562,15 | 8.809,09 | 0,00 | | |
| Málaga | 91.773,14 | 91.773,14 | 91.773,14 | 91.773,14 | 91.773,14 | 91.773,14 | 0,00 | |
| Sevilla | 91.773,14 | 91.773,14 | 91.773,14 | 91.773,14 | 91.773,14 | 91.773,14 | 33.331,37 | 0,00 |

$$r_{prom} = 0,82098597$$

Observamos que el mayor valor del coeficiente cofenético, y por tanto el proceso que menos distorsiona la matriz de distancias inicial, es el método del promedio.

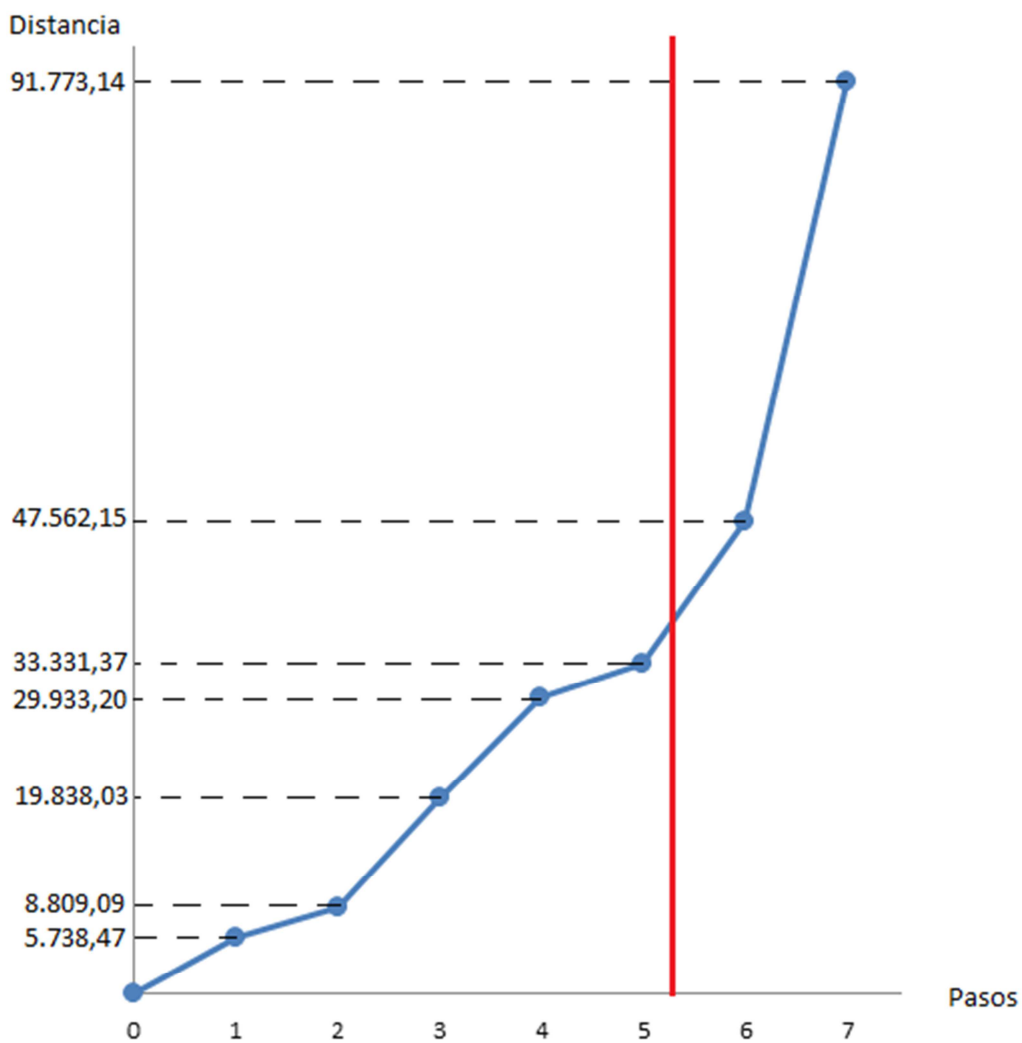
Nos quedamos pues con la clasificación propuesta por este método.

Por último nos queda la elección del número de clusters. Para esta elección representaremos mediante un gráfico los pasos seguidos en la agrupación de clusters y la distancia a la que se unen.

Técnicas de clusterización

Como ya explicamos anteriormente, cuando observemos que en uno de los pasos existe un salto brusco (pendiente elevada), paramos el proceso y nos quedamos con el número de clusters obtenidos hasta dicho paso.

Representación Gráfica:



Técnicas de clusterización

Observamos que la pendiente comienza a ser demasiado elevada del paso 5 al 6, con lo cual paramos el proceso en el paso 5 (distancia 33.331,37), en el que teníamos formados 3 clusters:

- Sevilla-Málaga
- Granada-Cádiz
- Almería-Huelva-Jaén-Córdoba

Podemos clasificar al primer grupo como “alto número de usuarios”, al segundo “moderado número de usuarios”, y al tercero “bajo número de usuarios”.



8 Conclusiones

El análisis Cluster o también llamado análisis de conglomerados, es una técnica estadística multivariante cuya finalidad es dividir un conjunto de objetos en grupos de forma que los perfiles de los objetos de un mismo grupo sean muy similares entre si y los de los objetos de clusters diferentes sean distintos.

Para llevar a cabo dicho análisis hemos seguido los siguientes pasos:

1. Plantear el problema a resolver.
2. Establecer medidas de semejanza y de distancia entre los objetos a clasificar en función del tipo de datos que estamos analizando.
3. Analizar algunos de los métodos de clasificación propuestos.
4. Interpretar los datos obtenidos.
5. Analizar la validez de la clasificación obtenida.

Esta técnica, es una técnica exploratoria cuya finalidad es sugerir ideas al analista a la hora de elaborar hipótesis y modelos que expliquen el comportamiento de las variables analizadas.

Técnicas de clusterización

Los resultados del análisis deberían tomarse como punto de partida en la elaboración de teorías que expliquen dicho comportamiento.

Los distintos métodos vistos en el caso práctico nos confirman que el método de la distancia mínima conduce a clusters encadenados, el de la distancia máxima a clusters compactos, siendo este menos sensible a valores atípicos que el de distancia mínima, y el método del promedio es el menos sensible a valores atípicos y tiende a formar clusters más compactos y de igual tamaño.

La agrupación exacta de un cluster no es una tarea sencilla y es difícil hacer recomendaciones generales. Siempre es aconsejable intentar más de un método. Si varios métodos dan resultados semejantes, entonces se puede suponer que en realidad existen agrupaciones naturales.

9 Referencias

“Análisis clúster” [en línea] (n.d.), Estadística aplicada a las ciencias económicas y sociales, Universidad de Valencia,

<http://www.uv.es/ceaces/multivari/cluster/CLUSTER.htm>

Fernández, S. (2011). “Análisis conglomerados” [en línea], Universidad autónoma de Madrid,

<http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/SEGMENTACION/CONGLOMERADOS/conglomerados.pdf>

Gallardo, M.A. (2013). “Ampliación de Análisis de datos multivariantes” [en línea], Universidad de Granada, <http://www.ugr.es/~gallardo/>

Llopis, J. (2013). “Análisis clúster” [blog],

<http://estadisticaorquestainstrumento.wordpress.com/2013/01/02/tema-19-analisis-cluster/>

Mahía, R. (n.d.). “Análisis clúster” [en línea], Universidad Autónoma de Madrid,

http://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF

Técnicas de clusterización

Perea, J. (n.d.). *“Curso de Postgrado en Herramientas Estadísticas Avanzadas”* [en línea], Universidad de Córdoba,
http://www.uco.es/zootecniaygestion/img/pictorex/09_13_25_sesion_8.pdf

Prieto, R. (2006). *“Técnicas estadísticas de clasificación”* [en línea], Estado de Hidalgo,
<http://www.uaeh.edu.mx/docencia/Tesis/icbi/licenciatura/documentos/Tecnica%20estadisticas%20de%20clasificacion.pdf>

Salvador, M. (2001.). *“Análisis de conglomerados o clúster “* [en línea], Scampus.org, Estadística, <http://ciberconta.unizar.es/leccion/cluster/inicio.html>

Terrádez, M. (n.d.). *“Análisis de conglomerados”* [en línea], Universidad Oberta de Cataluña, <http://www.uoc.edu/in3/e-math/docs/Cluster.pdf>

Villardón, J.L. (n.d.). *“Introducción al Análisis Clúster”* [en línea], Departamento de Estadística, Universidad de Salamanca,
<http://benjamindespensa.tripod.com/spss/AC.pdf>