Capítulo 4

Métodos no Jerárquicos de Análisis Cluster.

4.1. Introducción.

Los métodos jerárquicos, para un conjunto de m individuos, parten de m clusters de un miembro cada uno hasta construir un solo cluster de m miembros (métodos aglomerativos) o viceversa (métodos disociativos).

Los métodos que se presentan ahora están diseñados para clasificar individuos (no son válidos para variables) en una clasificación de K clusters, donde K se especifica a priori o bien se determina como una parte del proceso.

La idea central de la mayoría de estos procedimientos es elegir alguna partición inicial de individuos y después intercambiar los miembros de estos clusters para obtener una partición *mejor*.

Los diversos algoritmos existentes se diferencian sobre todo en lo que se entiende por *una partición mejor* y en los métodos que deben usarse para conseguir mejoras. La idea general de estos métodos es muy similar a la señalada en los algoritmos descendentes en más de un paso empleados en la optimización sin restricciones en programación no lineal. Tales algoritmos empiezan con un punto inicial y generan una secuencia de movimientos de un punto a otro hasta que se encuentra un óptimo local de la función objetivo.

Los métodos estudiados ahora comienzan con una partición inicial de los individuos en grupos o bien con un conjunto de puntos iniciales sobre los cuales pueden formarse los clusters. En muchos casos, la técnica para establecer una partición inicial es parte del algoritmo cluster, aunque estas técnicas usualmente son proporcionadas por sí solas más que como una parte del algoritmo cluster.

4.2. Elección de puntos semilla.

Supuesto que el número de clusters a formar es k, un conjunto de k puntos semilla no es más que un conjunto de puntos que puede emplearse como núcleo de los clusters sobre los cuales el conjunto de individuos puede agruparse. Los procedimientos, todos subjetivos, que pueden emplearse para tal hecho son:

1. Elegir los primeros k individuos del conjunto de datos, como propone McQueen (1967). Este método es el más simple, siempre y cuando la secuenciación en la que los datos han sido introducidos no influencie el resultado final.

2. Etiquetar los casos de 1 a m y elegir aquellos etiquetados como

$$\left[\frac{m}{k}\right], \left[\frac{2m}{k}\right], \ldots, \left[\frac{(k-1)m}{k}\right] \quad \text{y } m$$

   donde $[x]$ representa la parte entera de x. Con este sistema se pretende compensar la tendencia natural de ordenar los casos en el orden de introducción o alguna otra secuencia no aleatoria.

3. Etiquetar los casos de 1 a m y elegir los casos correspondientes a k números aleatorios diferentes, (McRae, 1971).

4. Tomar una partición de casos en k grupos mutuamente excluyentes y usar sus centroides como semillas, (Forgy, 1965).

5. Emplear el algoritmo de Astrahan (1970) según el cual se eligirían las semillas de tal forma que abarcaran todo el conjunto de datos, o sea, los datos estarán relativamente próximos a un punto semilla, pero las semillas estarán bien separadas unas de otras. Astrahan propuso el siguiente algoritmo para ello:

   - Para cada individuo se calcula la *densidad*, entendiendo por tal el número de casos que distan de él una cierta distancia, digamos d_1.
   - Ordenar los casos por densidades y elegir aquel que tenga la mayor densidad como primer punto semilla.
   - Elegir de forma sucesiva los puntos semilla en orden de densidad decreciente sujeto a que cada nueva semilla tenga al menos una distancia mínima, d_2, con los otros puntos elegidos anteriormente. Continuar eligiendo semillas hasta que todos los casos que faltan tengan densidad cero, o sea, hay al menos una distancia d_1 de cada punto a otro.
   - En el caso de que, por este procedimiento, se produjera un exceso de puntos generados, se agruparán de forma jerárquica hasta que haya exactamente K. Por ejemplo, el método del centroide puede ser elegido para tal cuestión.

Ejemplo: elección de puntos semilla: Algoritmo de Astrahan

Se consideran 7 individuos que toman valores respecto de 2 variables. Para el cálculo de la matriz de distancias utilizamos el cuadrado de la distancia euclídea.

	X_1	X_2
I_1	1	5
I_2	2	10
I_3	5	5
I_4	8	10
I_5	10	1
I_6	12	2
I_7	6	4

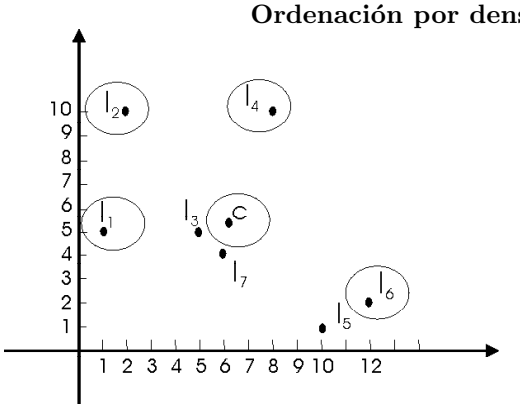
Distancia euclídea $\Longrightarrow$

	I_1	I_2	I_3	I_4	I_5	I_6	I_7
I_1	0						
I_2	26	0					
I_3	16	34	0				
I_4	74	36	34	0			
I_5	97	145	41	85	0		
I_6	130	164	58	80	5	0	
I_7	26	50	2	40	25	40	0

Densidades

•$d_1 = 60$ $\quad d(I_1) = 3 \quad d(I_2) = 2 \quad d(I_3) = 0 \quad d(I_4) = 3 \quad d(I_5) = 3 \quad d(I_6) = 3 \quad d(I_7) = 0$

Ordenación por densidades $(I_1 I_4 I_5 I_6)(I_2)(I_3 I_7)$



1º punto semilla I_1 $d_2 = 40$

2º punto semilla I_4 $d_3 = 30$

3º punto semilla I_5 $d_4 = 15$

4º punto semilla I_2 $d_5 = 10$

5º punto semilla I_3 $d_6 = 5$

6. - Ball y Hall (1967) proponen tomar el vector de medias de los datos como el primer punto semilla; posteriormente se seleccionan los puntos semilla examinando los individuos sucesivamente, aceptando uno de ellos como siguiente punto semilla siempre y cuando esté, por lo menos, a alguna distancia, d, de todos los puntos elegidos anteriormente Se continúa de esta forma hasta completar los k puntos deseados o el conjunto de datos se agota.
   - Notemos que este método es tan simple que permite probar con diversos valores de la distancia d si los anteriormente empleados proporcionaran pocas semillas o examinaran una parte pequeña del conjunto de datos.

Ejemplo: elección de puntos semilla: Algoritmo de Ball y Hall

Consideramos el mismo ejemplo anterior:

	X_1	X_2
I_1	1	5
I_2	2	10
I_3	5	5
I_4	8	10
I_5	10	1
I_6	12	2
I_7	6	4

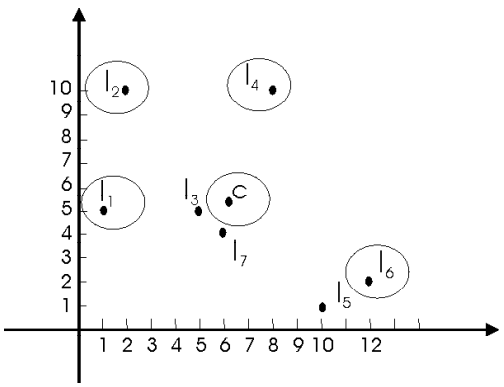
Distancia euclídea $\Longrightarrow$

	I_1	I_2	I_3	I_4	I_5	I_6	I_7
I_1	0						
I_2	26	0					
I_3	16	34	0				
I_4	74	36	34	0			
I_5	97	145	41	85	0		
I_6	130	164	58	80	5	0	
I_7	26	50	2	40	25	40	0

•$C(\bar{x_1}, \bar{x_2}) = (6{,}28, 5{,}28)$
1º punto semilla

$$\begin{cases} d(I_1, C) = 27{,}95 & d(I_2, C) = 40{,}59 & d(I_3, C) = 1{,}71 \\ d(I_4, C) = 23{,}99 & d(I_5, C) = 32{,}15 & d(I_6, C) = 44{,}16 \\ d(I_7, C) = 1{,}71 \end{cases}$$

Ordenación por distancias a C $(I_6 I_2 I_5 I_1)(I_4)(I_3 I_7)$



1º punto semilla C $d_1 = 40$

2º punto semilla I_6

3º punto semilla I_2 $d_2 = 30$

4º punto semilla I_1 $d_3 = 15$

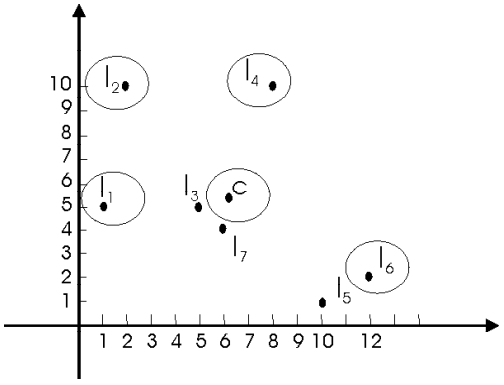
5º punto semilla I_4 $d_4 = 5$

4.3. Elección de particiones iniciales.

En algunos métodos cluster, el énfasis del método recae en generar una partición inicial de los individuos en K clusters exclusivos más que en encontrar un conjunto de puntos semilla. Algunos procedimientos para generar tales particiones son:

1. Para un conjunto de puntos semilla dado, se asigna cada caso al cluster construido sobre el punto semilla más próximo, (Forgy, 1965), permaneciendo los puntos semilla estacionarios durante la asignación. Con ello el conjunto de clusters resultante es independiente de la secuencia en la cual los individuos han sido introducidos.

   Ejemplo de aplicación, si partimos por ejemplo de los puntos semilla obtenidos por el algoritmo de Astrahan.



1º punto semilla I_1 $d_2 = 40$

2º punto semilla I_4 $d_3 = 30$

3º punto semilla I_5 $d_4 = 15$

4º punto semilla I_2 $d_5 = 10$

5º punto semilla I_3 $d_6 = 5$

Procedimiento de Forgy:
Se asigna cada caso al cluster construido sobre el punto semilla más próximo. Los puntos semilla quedan estacionarios durante el proceso.

$$\begin{array}{ccccc} (I_1) & (I_4) & (I_5) & (I_2) & (I_3) \\ & & \uparrow & & \uparrow \\ & & (I_6) & & (I_7) \end{array}$$

2. Dado un conjunto de puntos semilla, sea cada uno de ellos, inicialmente, un cluster unitario. A continuación se asigna cada individuo al cluster con el centroide más próximo. Tras asignarlo, se actualiza el centroide del cluster. Este método tiene una gran semejanza con el método descrito en el tema de métodos jerárquicos. Al igual que en el método del centroide, los clusters pueden irse moviendo, por lo que la distancia entre un individuo y un centroide puede ir variando durante el proceso. Además, el conjunto de clusters resultante es independiente del orden en el que los individuos fueron asignados.

3. Emplear un método jerárquico para producir una partición inicial idónea. Wolfe (1970) emplea el método de Ward para proporcionar un conjunto inicial de clusters para su algoritmo.

4.4. Métodos que fijan el número de clusters.

En primer lugar notemos que, dada una configuración de clusters, se pueden tomar como puntos semilla los centroides de los clusters así como se puede construir un conjunto de clusters asignando cada individuo al cluster con el punto semilla más próximo. El método más simple, iterativo, consiste en alternar estos dos procesos hasta que se converja a una configuración estable. A continuación veremos varios de estos métodos siguiendo el problema básico de ordenar los individuos en un número fijo de clusters de tal forma que cada individuo pertenezca a un solo cluster. Asimismo plantearemos algunas variantes de estos procedimientos.

4.4.1. Método de Forgy y variante de Jancey.

Forgy, en 1965, sugiere un algoritmo simple consistente en la siguiente secuencia de pasos:

1. Comenzar con una configuración inicial. Ir al paso segundo si se comienza con un conjunto de puntos semilla. Ir al paso tercero si se comienza con una partición de los casos.

2. Colocar cada individuo en el cluster con la semilla más próxima. Las semillas permanecen fijas para cada ciclo completo que recorra el conjunto de datos.

3. Calcular los nuevos puntos semilla como los centroides de los clusters.

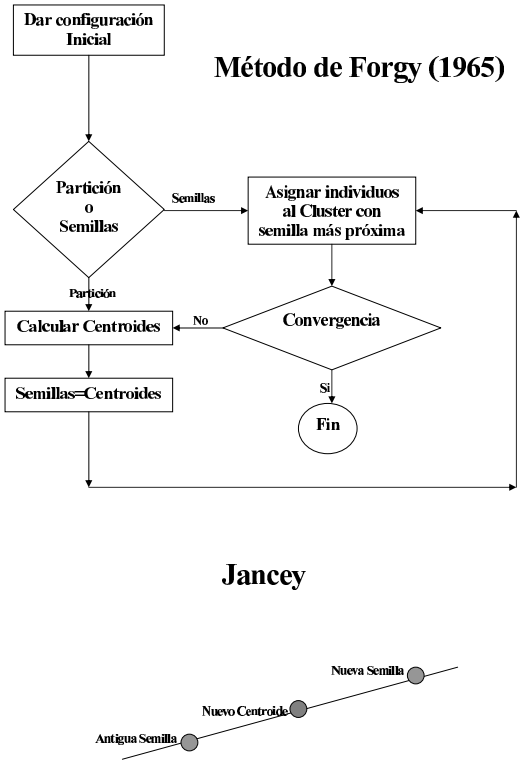
4. Alternar los pasos segundo y tercero hasta que el proceso converja, o sea, continuar hasta que ningún individuo cambie de cluster en el paso segundo.

Notemos que no es posible decir cuantas repeticiones de los pasos segundo y tercero serán precisas para alcanzar la convergencia en un problema concreto; no obstante, evidencias empíricas indican que en la mayoría de los casos no suelen ser necesarios más de 10 ciclos.

En cada ciclo la asignación de K clusters requiere mK cálculos de distancias y $m(K-1)$ comparaciones de distancias. Puesto que K es frecuentemente menor que m y el número de ciclos hasta alcanzar la convergencia es pequeño, el analista puede con frecuencia examinar conjuntos de clusters asociados con varios valores de K, con un coste bastante menor del asociado a un análisis jerárquico completo.

Jancey, en 1966, sugiere el mismo método excepto una modificación en el paso tercero. Así, el primer conjunto de clusters formado por puntos semilla viene dado o bien se calcula como los centroides de los clusters de la partición inicial; en cada etapa, cada nuevo punto semilla se encuentra reflejando el antiguo punto semilla a través del nuevo centroide del cluster, lo cual puede ser visto como una aproximación al gradiente local, o sea, la dirección en la cual el punto semilla debiera moverse para tener un mejor aprovechamiento de la partición, en tanto en cuanto se desplaza en el mismo sentido que lo hace el cluster en su totalidad.

Figura 4.1: Método de Forgy. Variante de Jancey.



4.4.2. Método de las K-Medias de MacQueen.

MacQueen, en 1972, emplea el término *K-Medias* para denotar el proceso de asignar cada individuo al cluster (de los K prefijados) con el centroide más próximo. La clave de este procedimiento radica en que el centroide se calcula a partir de los miembros del cluster tras cada asignación y no al final de cada ciclo, como ocurre en los métodos de Forgy y Jancey.

El algoritmo que propuso es el siguiente:

1. Tomar los K primeros casos como clusters unitarios.
2. Asignar cada uno de los $m - K$ individuos restantes al cluster con el centroide más próximo. Después de cada asignación, recalcular el centroide del cluster obtenido.
3. Tras la asignación de todos los individuos en el paso segundo, tomar los centroides de los clusters existentes como puntos semilla fijos y hacer una pasada más sobre los datos asignando cada dato al punto semilla más cercano.

El último paso es el mismo que el del método de Forgy, excepto que la recolocación se efectúa una vez más sin esperar a que se produzca la convergencia.

Notemos que, usando los K primeros individuos como puntos semilla, este método tiene la virtud de ser el menos *caro* de todos los métodos discutidos. El cómputo total de operaciones desde la configuración inicial hasta la final involucra sólo $K(2m-K)$ cálculos de distancias, $(K-1)(2m-K)$ comparaciones de distancias y $m-K$ cálculos de centroides.

Hay que comentar que el conjunto de clusters construido en el paso segundo del algoritmo depende de la secuencia en la que los individuos han sido procesados. MacQueen (1967) efectuó algunos estudios preliminares en este sentido; su experiencia indicó que la ordenación de los datos tiene sólamente un efecto marginal cuando los clusters están bien separados.

A partir del algoritmo anterior se puede implementar un procedimiento convergente, que llamaremos método de las K-medias convergente:

1. Comenzar con una partición inicial de los individuos en clusters. Si se desea, la partición puede ser construida usando los pasos primero y segundo del método de MacQueen ordinario, si bien cualquier método de los establecidos para particiones iniciales puede emplearse

2. Tomar cada caso sucesivamente y calcular las distancias a todos los centroides de los clusters; si el centroide más próximo no es el del cluster padre de dicho caso, reasignar dicho caso al cluster con centroide más próximo y recalcular los centroides de los clusters afectados en el proceso.

3. Repetir el paso segundo hasta que se obtenga la convergencia, o sea, continuar hasta que un ciclo completo a través de todos los casos no proporcione ningún cambio en los miembros de los clusters.

4.4.3. Algunas cuestiones sobre estos métodos.

Los métodos jerárquicos estudiados en el tema anterior se distinguen completamente unos de otros por las propiedades que imponen en el procedimiento

de clasificación jerárquica que producen, sin embargo, los cuatro métodos iterativos introducidos hasta ahora no se distinguen tan bien unos de otros. Los métodos de Forgy, Jancey y el método de las K-medias convergente emplean variaciones de un proceso central y apenas muestran diferencias en el campo computacional. Estos tres métodos también convergen en el mismo sentido, por lo que el conjunto final de clusters producido por un método puede satisfacer el criterio de convergencia de los otros métodos. Así pues estos tres métodos pueden converger a la misma partición, por lo que el analista debe elegir aquel que sea más conveniente usar o más fácil desarrollar. Si, por el contrario, los métodos tienden a proporcionar diferentes particiones, sería interesante caracterizar esas diferencias para tener una especie de guía de selección del método apropiado a los datos y a los intereses del analista.

Los métodos planteados son simples, económicos y bastante populares. Aún así no queda muy claro si difieren de forma sistemática viendo las propiedades sustanciales de los clusters que producen. Son muchas las preguntas que surgen de forma inevitable sobre estos métodos, preguntas que aparecen ante la subjetividad que subyace en ellos. Algunas posibles cuestiones serían:

1. ¿La configuración inicial tiene algún efecto sobre los clusters finales? Si es así, ¿cuál de los procedimientos es más sensible a esta cuestión? ¿Hay algún método mejor que otro de generar una configuración inicial? ¿Es realmente útil el esfuerzo hecho en la generación de los puntos semilla?

2. El método de Jancey difiere del de Forgy sólo en la generación de los sucesivos conjuntos de puntos semilla. ¿Esta diferencia tiene algún efecto beneficioso en la convergencia posterior? Jancey sugiere que reflejando el punto semilla sobre el centroide se tiende a obtener particiones con un menor error de suma de cuadrados, ¿es cierto este hecho?

3. En el método de las K-medias, el orden en que se han introducido los casos hace que haya una diferencia sustancial en los clusters finales. Las iteraciones extras en el método de las K-medias convergente, ¿hacen disminuir esa sensibilidad procedente de la partición inicial?

4. En las particiones finales hay alguna distinción sistemática que puede ser atribuida a la actualización de los centroides tras la asignación de cada individuo (como ocurre en el método de las K-medias) en contraposición con la actualización sólo después de un ciclo completo a través de los datos. ¿Afecta esta política a la velocidad de convergencia?

5. ¿Es útil continuar recolocando casos hasta que la partición converja? ¿Cuál de los tres métodos convergentes da, en esencia, la misma partición final pero con un costo reducido, parándose cuando un ciclo completo proporciona una reasignación menor que un número pequeño prefijado, por ejemplo el 1 % de los casos?

6. ¿Afecta el número de casos o variables al número de iteraciones necesarias para obtener la convergencia?

Estas cuestiones afectan sólo a algunos de los tópicos que deben ser estudiados. Las respuestas a éstas y otras cuestiones son de gran valor, no sólo en la elección entre estos cuatro métodos particulares sino en identificar criterios pertinentes para la evaluación de métodos clusters en general.

4.4.4. Propiedades de convergencia.

Antes de comenzar con este punto hay que hacer notar que no abordaremos de forma exhaustiva demostraciones relacionadas con la convergencia de los métodos debido a que son muy largas y tediosas, corriéndose el riesgo de oscurecer más que iluminar las conclusiones principales. Más bien haremos comentarios sobre dichas convergencias, comentarios a partir de los cuales se pueden obtener demostraciones rigurosas.

1. Para un cluster dado, la suma de los cuadrados de las desviaciones sobre un punto de referencia es mínima cuando ese punto de referencia es el centroide del cluster. La suma de los cuadrados de las desviaciones sobre el centroide para el K-ésimo cluster viene dada por

$$E_k = \sum_{i=1}^{m_k} \sum_{j=1}^{n} (x_{ijk} - \overline{x}_{jk})^2$$

   Para una partición dada de un conjunto de individuos en h clusters, la suma de los cuadrados de los errores intragrupos es

$$E = \sum_{k=1}^{h} E_k$$

   y E posee un valor característico para dicha partición. Notemos que

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

   es el cuadrado de la distancia euclídea entre el centroide del cluster K y el j-ésimo individuo en dicho cluster.

2. El número de distintas formas en las cuales un conjunto de m casos puede ser particionado en h clusters es un número de Stirling de segunda especie

$$\mathbb{S}_n^{(m)} = \frac{1}{m!} \sum_{k=0}^{m} (-1)^{m-k} \binom{m}{k} k^n$$

   por lo que, para no tener que calcular todas esas particiones, consideraremos métodos en los cuales la partición actual es alterada sólo si el cambio proporciona una nueva partición con un error total E menor. Puesto que cada partición tiene un valor característico E, tales métodos no pueden regenerar una partición que haya sido hecha en una etapa anterior y por lo tanto dichos métodos son convergentes. Así pues, un método es convergente si las sucesivas particiones que genera exhiben una sucesión decreciente de valores para E.

3. El método de las K-medias convergente y el método de Forgy son convergentes en el sentido explicitado anterioremente. Para comprobar esta cuestión consideremos como puntos semilla el más reciente conjunto de centroides calculado. En ambos métodos un individuo es recolocado sólo si es más próximo al punto semilla del cluster que se obtendría que al del que se perdería; si la función distancia elegida es la Euclídea (o una potencia suya), entonces la suma de los cuadrados de las desviaciones sobre el punto semilla decrece más para el cluster perdido que lo que aumenta para el ganado, obteniendo un decrecimiento general en la suma de desviaciones cuadradas para los puntos semilla para la partición como conjunto. Además esa suma de desviaciones cuadráticas disminuye siempre más si es calculada sobre los nuevos centroides que sobre los viejos puntos semilla. Entonces, cada nueva partición tiene un valor menor de E que la partición de la cual los puntos semilla fueron calculados, con lo que estos métodos son convergentes.

Estrictamente hablando, la convergencia ha sido probada sólo para el caso en que el objetivo es encontrar particiones que minimicen E; esos métodos minimizan dicho error cuando se emplea la distancia euclídea y pueden o no ser convergentes dependiendo de la función objetivo.

Si se emplea la distancia L_1 como medida de la divergencia entre un individuo y un punto semilla, entonces el mínimo de

$$E_k = \sum_{i=1}^{m_k} \sum_{j=1}^{n} |x_{ijk} - c_{jk}|$$

se obtiene con c_{jk} igual a la mediana de la j-ésima variable en el k-ésimo cluster. Revisando los métodos de calcular puntos semilla, como las medianas de los clusters, y usando la métrica L_1 entonces se minimiza la suma total entre grupos de los errores absolutos. Este esquema puede ser usado para demostrar la convergencia empleando los mismos argumentos anteriores.

En cuanto al método de Jancey hay que comentar que no se ha encontrado ninguna prueba de su convergencia así como ningún ejemplo que demuestre lo contrario. El argumento empleado en los casos anteriores para los métodos de Forgy y de las k-medias no puede ser empleado en este caso ya que sucesivas particiones pueden tener un valor mayor o menor de E que sus predecesores inmediatos. Con ello puede repetirse la misma partición en etapas sucesivas, con el riesgo de entrar en un ciclo.

Notemos también que el criterio elegido para decidir la convergencia de estos métodos es estabilizar los miembros de los clusters; un criterio alternativo sería estabilizar los puntos semilla. Estos dos criterios son equivalentes para los métodos de Forgy y el de k-medias convergente, ya que los puntos semilla son los centroides de los clusters, que dependen, obviamente, de los miembros de los clusters. Sin embargo dicho criterio no es posible aplicarlo para el método de Jancey.

Supongamos en la figura de la variante de Jancey un cluster que tenga como centroide el punto 2 y que todos los individuos están suficientemente lejos de tal forma que agrupando sobre el punto 1 o sobre el punto 3 se tengan los mismos miembros. Entonces,

tomando el punto 1 como semilla inicial se tiene el 3 como segundo punto semilla con lo cual, tras reflejar de nuevo, se obtiene de nuevo el punto 1 y así sucesivamente, entrando en un ciclo.

4.5. Métodos con el número final de clusters variable.

A continuación vamos a centrarnos en la discusión de métodos más elaborados que emplean métodos heurísticos para ajustar el número de clusters con el fin de atenerse lo más posible a la estructura natural del conjunto de datos.

El gráfico 4.5 nos puede ayudar a motivar alguna de las dificultades con las que nos podemos encontrar:

En esta figura podemos apreciar un conjunto de datos que consta de 5 clusters naturales y un dato anómalo que no pertenece a ninguno.

1. Si forzamos a estos datos a estar configurados en seis o más clusters casi seguramente resultaría un nuevo cluster formado por el dato anómalo. Así, si aumentamos el número de clusters ese dato anómalo podría ser considerado como un cluster unitario y los restantes datos serían clasificados como si dicho dato no estuviera presente.

2. Si sólo ha sido localizado un punto semilla en la proximidad de los clusters 1 y 2, estos dos clusters, probablemente, aparecerían unidos pero con alguna discrepancia entre sus miembros. Si el número de clusters aumentara, este cluster podría ser dividido en dos más compactos.

3. Por otra parte, si hubiera varios puntos semilla en la proximidad del cluster 2 entonces podría haber varios clusters poco diferenciados. Admitiendo que el número de clusters decreciera, estos clusters podrían ser combinados en uno más distintivo de los demás. El hecho de forzar la creación de 5 ó 6 clusters, probablemente, puede llevar a clusters pobremente definidos.

Tengamos presente, no obstante, que esta situaciones son fácilmente reconocibles en dos dimensiones, mientras que en un problema con un número mayor de variables y un gran número de individuos resultaría prácticamente imposible. Los métodos que a continuación veremos están enfocados a evitar estos problemas.

4.5.1. Nueva versión del método de las k-Medias.

Mac Queen propuso una variante de su método de las k-medias que permite variar el número de clusters durante la asignación inicial de los individuos a los clusters. Los pasos del método son

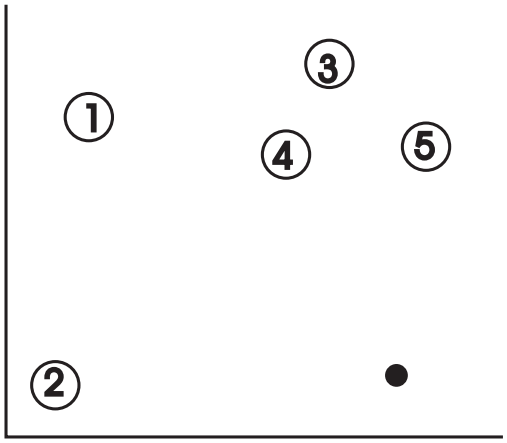
1. Elegir valores para los parámetros K,C y R, los cuales son usados en los siguientes pasos del método.

2. Tomar los K primeros individuos como clusters unitarios.

3. Calcular todas las distancias, dos a dos, entre los K primeros individuos. Si la distancia menor es inferior a C, entonces se unen los dos clusters correspondientes y se recalculan las distancias entre el centroide del nuevo cluster y los clusters restantes. El proceso continúa uniendo los clusters más cercanos hasta que los centroides estén separados por una distancia al menos igual a C.

4. A continuación se asignan los restantes $m - K$ individuos, uno a uno, al cluster con centroide más cercano. Tras cada asignación se recalcula el centroide del cluster obtenido y se calculan las distancias a los centroides de los otros clusters; se une el nuevo cluster con el cluster que tenga el centroide más

Figura 4.2: Necesidad de un número variable de clusters



próximo siempre y cuando la distancia entre los centroides sea menor que C. Se continúa este proceso hasta que todos los centroides disten entre sí en una cantidad mayor que C.

Durante la asignación de individuos, si la distancia al centroide es mayor que R se considera a dicho individuo como un cluster unitario antes que asignarlo a un cluster ya existente.

5. Tras la signación de lso individuos se toman los centroides de los clusters existentes como puntos semillas fijos y se coloca cada individuo atendiendo al punto semilla más próximo.

Al igual que en el método de las k-medias básico, el proceso finaliza después de la primera recolocación y no espera a la convergencia. Consecuentemente este método resulta el menos *costoso* de los que describiremos en este apartado. Nótese que al asignar los clusters con los centroides cercanos para unirlos, el método evita crear distinciones finas que dividan artificialmente los clusters naturales. Creando nuevos clusters cuando un individuo es distante de los demás centroides existentes, el conjunto final de clusters tiende a abarcar el conjunto de los datos y los datos anómalos tienden a resaltar por sí mismos antes que forzarlos a estar en un cluster.

En cuanto a la elección de R y C no hay mucha experiencia que sirva de guía, si bien como es lógico R tiene que ser mayor que C. Tomando R grande y C pequeño este método puede servir para identificar datos anómalos, siendo este un paso preliminar en otros métodos de análisis.

4.5.2. Variante de Wishart del método de las k-Medias.

Este método persigue conseguir una reducción en el número de clusters final utilizando como base el método de las k-medias convergente. Los pasos del algoritmo son los siguientes

1. Elegir valores para los parámetros de control Thresh, Minsiz, Maxit, Minc. En los siguientes pasos del algoritmo se verá el papel que desempeñan estos parámetros.

2. Comenzar con una partición inicial de los datos y calcular los centroides.

3. Sea K el actual número de clusters. Considerar cada individuo de forma sucesiva y calcular las distancias a los K centroides.

   - Si la distancia menor excede a Thresh, entoces se asigna el individuo a un residuo de datos no clasificados y se recalcula el centroide del cluster del que procedía el individuo.
   - Si el centroide más próximo no es el del cluster al que pertenece el individuo y la distancia no excede a Thresh, entonces se reasigna el individuo y se recalculan los centroides del cluster de partida y del de llegada.
   - Si el individuo está actualmente en el residuo y la distancia al centroide más próximo no excede a Thresh, entonces se reasigna el individuo y se recalcula el centroide del cluster al que se asigna el individuo.

4. Tras recolocar todos los individuos según el paso 3, se asignan los clusters con menos de Minsiz miembros al residuo; notemos que tal asignación reduce el número de clusters.

5. Los pasos 3 y 4 se repiten hasta que la partición converja, o sea, hasta que no haya cambios en los miembros en el paso 3 o hasta que el número de iteraciones efectuadas supere a Maxit.

6. Calcular las similaridades dos a dos entre los clusters y unir los dos más similares. Repetir 2 a 5 para obtener otra partición. Continuar de esta forma hasta que el número de clusters se reduzca a Minc.

Este método produce distintas particiones entre la inicial y la última (compuesta por Minc clusters). El método no tiene por qué proporcionar particiones para todos los valores entre el número inicial y Minc ya que el paso 4 puede saltar algunas posibilidades.

Asimismo algunos clusters pueden permanecer en el residuo una vez que se han asignado todos los datos. No obstante, los individuos situados entre clusters, frecuentemente se mueven dentro y fuera del residuo. Esta posibilidad conduce a la necesidad del parámetro Maxit, ya que puede ocurrir un fenómeno de recurrencia que impida la convergencia. El paso 3 del método es parecido al del método de las k-medias convergente, salvo en lo que concierne al movimiento del residuo; tal vez las propiedades de convergencia de este último método puedan recuperarse dejando los datos en el residuo hasta que el número de clusters cambie.

4.5.3. El método Isodata.

El método Isodata es, posiblemente, el más elaborado de los métodos clusters basados en el centroide más cercano. Ball y Hall presentaron en 1965 una completa descripción paso a paso del método original. Este procedimiento ha sido objeto de múltiples estudios existiendo diversas variantes. La versión que presentamos es representativa de la mayor parte de las existentes, si bien difiere en algunos detalles de las otras.

El algoritmo es el siguiente

1. Elegir valores para los siete parámetros de control: Nparts, Nrwdsd, Nclst, Thetan, Thetae, Thetac e Itermax. El papel de cada uno de estos parámetros se verá claro en los pasos siguientes.

2. Si no se han proporcionado los puntos semilla, generarlos mediante el método de Ball y Hall.

3. Asignar cada individuo al cluster con punto semilla más próximo. Los puntos semilla permanecen fijos durante una iteración completa, o sea, mientras se tratan todos los datos en un ciclo. Recalcular los centroides y asignarlos como nuevas semillas. Este procedimiento se repite hasta que se alcance la convergencia o el número de ciclos alcance el valor Nparts. Notemos que este paso es idéntico al del método de Forgy.

4. Descartar los clusters que tengan menos de Thetan unidades. Los individuos asociados a dichos clusters se disgregan y pasan a englobarse con el resto de individuos para continuar el análisis.

5. Desarrollar o bien una unión de clusters o una disgregación de alguno según las siguientes reglas

   - La unión de clusters se produce si el número de clusters es el doble o más del parámetro Nwrdsd.
   - Una disgregación se produce si el número de clusters es la mitad o menos de Nwrdsd.
   - En otro caso, alternar el proceso disgregando en las iteraciones impares y uniendo en las pares.

6. Calcular nuevos puntos semilla como los centroides de los clusters y desarrollar la colocación de los clusters como en el paso 3.

7. Repetir los pasos 4,5 y 6 hasta que el proceso converja o hasta que estos 3 pasos hayan sido repetidos Itermax veces.

En cuanto a las fases de unión y disgregación comentadas anteriormente se tiene

- En la fase de unión se calculan todas las distancias, dos a dos, entre los centroides de los clusters. Si la distancia entre los más próximos es inferior a Thetac, entonces los clusters asociados se unen y se recalculan las distancias a los otros centroides. Este proceso continúa hasta un máximo de Nclst uniones en alguna iteración. Por supuesto no hay uniones si los centroides están separados suficientemente.

- En la fase de disgregación un cluster se elige provisionalmente para disgregarse si, para alguna variable, la desviación típica excede al producto de Thetae y la desviación típica de dicha variable en el conjunto original de datos. Los individuos de ese cluster se asignan a dos nuevos clusters según tengan una media superior o inferior a la media de la variable que separa en el cluster disgregado. Los centroides de estos clusters se calculan y, si la distancia entre ellos es al menos 1.1 veces Thetae, entonces la división del cluster se lleva a cabo.