Índice general

1 INTRODUCCIÓN A LAS TÉCNICAS DE AGRUPAMIENTO **1**

1.1 Introducción . . . 1
1.2 Conceptos Básicos . . . 2
1.2.1 Definiciones iniciales . . . 2
1.3 Los datos de partida . . . 4
1.3.1 La matriz de patrones . . . 5
1.3.2 Indices de proximidad: distancias y semejanzas . . . 6
1.4 Técnicas de agrupamiento jerárquico . . . 8
1.4.1 Ideas básicas . . . 8
1.4.2 Algoritmos para el agrupamiento jerárquico . . . 8
1.5 Técnicas de agrupamiento particional . . . 11
1.5.1 Ideas iniciales . . . 11
1.5.2 El método de las k-medias . . . 13
1.5.3 DBSCAN: un método basado en el análisis de densidad . . . 15
1.6 Nuevos resultado y extensiones . . . 18
1.6.1 Resultados recientes sobre agrupamiento jerárquico . . . 18
1.6.2 Resultados recientes sobre métodos particionales prototípicos: los métodos de k-medoides . . . 20
1.7 Técnicas de agrupamiento difuso . . . 21
1.7.1 Agrupamientos difusos particionales . . . 22
1.7.2 Agrupamientos jerarquicos y conjuntos difusos . . . 25

Índice de figuras

1.1 Arbol de tipos de clasificaciones 3
1.2 Ejemplo de agrupamiento por enlace completo 9
1.3 Ejemplos de distintos métodos de agrupamiento 11
1.4 Ejemplo del método de las K-medias con 3 grupos 12
1.5 Ejemplo del método de las K-medias con 5 grupos 14
1.6 Ejemplo del método de las K-medias con 3 grupos, datos de partida elegidos 15
1.7 Ejemplo del método de las K-medias con 5 grupos, datos de partida elegidos 16
1.8 Conceptos básicos referentes a DBSCAN 17

Índice de Tablas

1.1 Distintas funciones de distancia 6
1.2 Distintos índices de semejanza 7
1.3 Ejemplo de matriz de proximidad, basada en distancias 8
1.4 Coeficientes para el agrupamiento jerárquico 10
1.5 Ejemplos de medida SSE en distintas aplicaciones del método de las k-medias 15

Capítulo 1

INTRODUCCIÓN A LAS TÉCNICAS DE AGRUPAMIENTO

1.1 Introducción

Una definición clásica y sucinta de agrupamiento es:

> Clasificación no supervisada de *patrones* (observaciones, datos o vectores de características) en grupos (clusters).

Dado que la clasificación es una de las abstracciones básicas en la inducción de conocimiento, este problema ha sido tratado en muchos contextos y por investigadores de muchas disiciplinas (Biología, Psicología, Análisis Económico, Sociología etc..), de forma que el término agrupamiento (*clustering*) se usa en numerosas comunidades de investigación para describir el proceso de clasificar en grupos un conjunto de items sin tener una información previa acerca de su estructura. Estas comunidades tienen diferente terminología para describir los elementos del proceso y diversas metodologías para resolver los problemas que el agrupamiento presenta.

Además, el agrupamiento han sido ampliamente estudiado desde hace más de cuarenta años, desde el final de los 60 se desarrollan técnicas de agrupamiento dentro el ámbito del Anaĺisis de Datos y de la Taxonomía Numérica, posteriormente el agrupamiento se ha incluido dentro del campo de la Inteligencia Artificial encuadrándose dentro del Aprendizaje no Supervisado, por último la Minería de Datos recoge el Agrupamiento como una clase de problemas a tratar y recupera las técnicas y metodologías previamente desarrolladas extendiendolas y adaptándolas al volumen de datos que se procesan en este campo.

Nos encontramos pues con un tema muy desarrollado, tanto por el largo tiempo que lleva estudiándose como por la amplia variedad de los enfoques teóricos empleados y de los campos donde se ha aplicado. Por ello no es posible dar dar una visión muy detallada del tema en un libro de tipo generalista como el que nos ocupa. Así pués, en este capítulo nos centraremos más en presentar las ideas básicas acerca del agrupamiento que en estudiar en profundidad los matices y variaciones de diversos algoritmos. Para ello remitimos al lector interesado a textos específicos, algunos de los cuales se recogen en la bibliografía [15, 14, 2].

El capítulo está organizado de la siguiente manera, comenzaremos con los conceptos básicos, definiendo y encuadrando problemas y técnicas, estudiaremos después los problemas de representación datos y las medidas de similaridad más habituales que son el punto de partida para el agrupamiento. Posteriormente analizaremos las técnicas más clasicas de agrupamiento jerarquico

y particional, comentando después técnicas más actuales, haciendo especial mención de las que se han adaptado para resolver problemas de Minería de Datos. Terminaremos con los métodos que contemplan grupos no exclusivos basados en Lógica Difusa .

1.2 Conceptos Básicos

1.2.1 Definiciones iniciales

Como se ha dicho, el proceso de agrupamiento (*cluster analysis*) no es más que la organización de una colección de patrones, o items en un conjunto de grupos homogéneos. Habitualmente estos items están representados por un vector de valores de atributos, es decir son puntos de algún espacio multidimensional, estos valores también se suelen denominar *factores*, *componentes* o simplemente variables. Intuitivamente, dos items pertencientes a un agrupamiento válido deben ser más parecidos entre sí que aquellos que estén en grupos distintos y partiendo de esta idea se desarrollan las técnicas de agrupamiento. Obviamente estas técnicas dependen de cómo sean los datos de partida, de qué medidas de parecido (semejanza) se estén utilizando y de qúe clase de problemas se estén resolviendo. A título de ejemplo: no es lo mismo clasificar bacterias según una taxonomía jerárquica que agrupar los píxels de una imagen por colores.

También es importante distinguir entre clasificación no supervisada (dentro de la cual se incluye el análisis cluster) y clasificación supervisada (dentro de la cual se encuentra el análisis discriminante). En el primer caso no tenemos ninguna información acerca de organización los items en grupos o clases y el objetivo es encontrar dicha organización en base a la proximidad entre items. No existe apenas información previa acerca de esta estructura y la interpretación de las clases o grupos obtenidos es una labor a realizar "a posteriori" por parte del analista. En el segundo se posee la información de a qué clase pertenece cada item y lo que se desea es determinar cuales son los factores que itervienen en la definición de las clases y que valores de los mismos determinan estas. Un ejemplo clásico de agrupamiento sería la búsqueda de grupos de clientes de una entidad bancaria utilizando para ello datos de la cuenta corriente: edad, dirección, nivel de renta, .. etc. Un ejemplo de clasificación sería encontrar los elementos que determinan la aparición de cancer de pulmón analizando datos de, edad, calidad de vida, nivel económico,..etc. tanto de personas enfermas como sanas.

Los problemas y metodologias para el agrupamiento pueden clasificarse según distintos criterios. La figura 1.1, tomada de [14]y [18] presenta un árbol de categorias de técnicas de agrupamiento. Comentaremos brevemente estas divisiones, y algunas que se relaciones y superponen a ellas ya que serán el hilo conductor de este capítulo

Clasificación supervisada y no supervisada Como hemos comentado la clasificación es el proceso mediante el cual se agrupa un conjunto de items en función de una representación de los mismos en un espacio n-dimensional. En el caso de que sea supervisada se tiene información acerca de los grupos de manera que sabemos a que grupo pertence cada clase, entonces lo que se desea es encontrar un conjunto de "criterios", probablemente reglas, que nos permitan, dado un nuevo item, situarlo en un grupo. Este problema se trata en un capítulo diferente de este libro. En el caso de la clasificación no supervisada no se tiene mucha información acerca de los grupos, a veces no se sabe siquiera cuantos grupos hay, se trata entonces es de encontrar un agrupamiento que reuna en un mismo grupo los items más "parecidos" y coloque en grupos diferentes a los items "dispares".

Como vemos hay una gran diferencia de enfoque entre ambos problemas, de hecho, en muchos

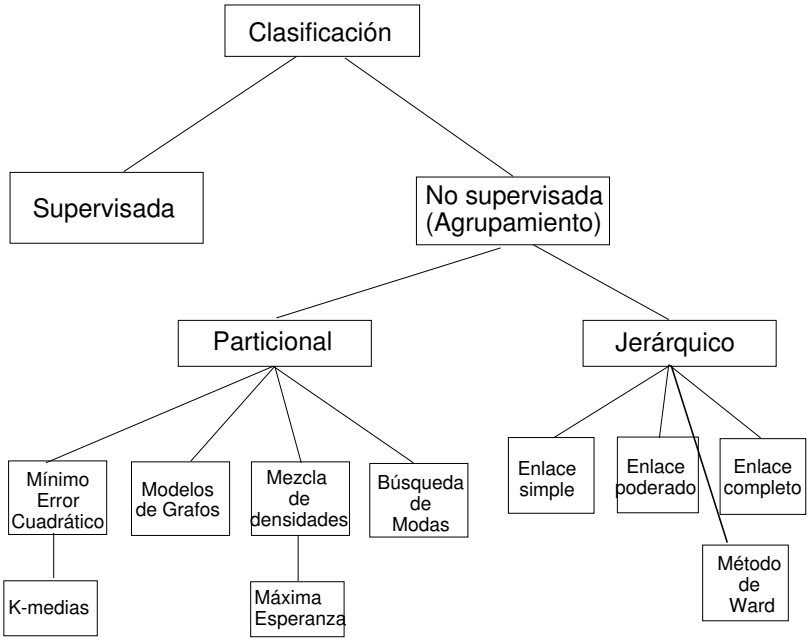


Figura 1.1: Arbol de tipos de clasificaciones

casos el agrupamiento puede verse como el primer paso para la clasificación supervisada.

Agrupamiento particional y agrupamiento jerárquico El objetivo final del proceso de agrupmamiento es obtener un conjunto de clases o grupos. Cuando estos grupos son disjuntos y cubren todo el conjunto de items se dice que el agrupamiento es *particional*. En algunos casos lo que se desea no es exactamente un agrupamiento particional sino jerarquía de agrupamientos particionales "anidados ", de tal manera que cada grupo de un nivel se divide en varios en el nivel siguiente. Esta estructura se denomina *agrupamiento jerárquico* y tiene una representación gráfica muy intuitiva denominada *dendrograma*.

Las técnicas de agrupamiento jerárquico son muy populares en ciencias biológicas, sociales y de la conducta donde se hace necesario construir taxonomías. Las técnicas particionales se usan fundamentalmente en aplicaciones de ingeniería donde se necesitan particiones simples. El agrupamiento particional es particularmente útil en el caso de que se trabaje con grandes bases de datos, es decir en aplicaciones de Minería de Datos, ya que los dendrogramas son poco prácticos cuando manejan más de unos pocos cientos de items.

Dependiendo de cómo se se planteen los objetivos del agrupamiento aparecen distintos enfoques, tanto en el caso particional como en el jerárquico, estos enfoques se representan en las distintas hojas del árbol anterior y realmente se identifican con la forma en que se contemplen o interpreten tanto los items como los grupos.

- Si se considera que los datos están representados mediante un grafo donde los vértices son los items o patrones y las aristas son conexiones entre ellos definidas a travás de semejanza, entonces los grupos aparecen como componentes conectadas del grafo. Este enfoque es el que hemos denominado *Modelos de Grafos*, y aparece tanto en caso de agrupamiento particional como en el jerárquico, ya que los agrupamientos jerárquicos de enlace, simple, completo

y medio se encuadran también dentro de este enfoque. Una filosofía similar siguen el bien conocido método del *vecino más cercano*

- Cuando se considera que los grupos deben ser "cohesionados" de manera que los items de un mismo grupo estén más cercanos a entre sí y la distancia entre grupos sea la mayor posible, aparece una amplia familia de modelos de agrupamiento particional. Uno de los más extendido es el de mínimos cuadrados, donde el criterio de cohesión se obtiene como la suma total de la distancia de cada item al punto medio (*centroide*) del grupo al que pertence, este valor obviamente debe ser mínimo. El bien conocido *método de las k-medias* se encuadra en esta categoria.

- Cuando se considera que un grupo es una región del espacio n-dimensional donde la densidad de items es muy alta, rodeada de una zona de baja densidad aparecen los denominados métodos basados en análisis de densidad. Dentro de este enfoque se encuadran los métodos de *estimación de densidad* y de *búsqueda de modas*, cuya idea de base es emplear estimaciones estadísticas de la densidad de probabilidad de cada grupo, suponiendo que cada uno de ellos es una muestra representativa de una población. La misma filosofia sigue el método hemos denominado de *mezcla de densidades*

La clasificación de métodos de agrupamiento que presenta la figura 1.1 es bastante exahustiva pero no es la única división posible, y desde el punto de vista de la relación entre los diferentes agrupamientos o de la forma de obtenerlos podemos hablar de:

Agrupamientos exclusivos, agrupamientos no exclusivos y agrupamientos difusos Todos los enfoques comentados en los párrafos anteriores parten de la hipótesis de no-solapamiento, es decir que deseamos un agrupamiento donde cada punto pertenezca sólo a un grupo. Cuando se relaja esta hipótesis aparecen métodos de agrupamiento que admiten solapamiento o no-exclusivos. Los métodos de agrupamiento no-exclusivos que han tenido más éxito son los que suponen que los grupos son *conjuntos difusos* de forma que un item puede pertenecer a diversos grupos con un grado de pertenencia a cada uno.

Agrupamientos aglomerativos y divisivos Este aspecto se refiere fundamentalmente a la estructura del algoritmo que desarrolla el método. Si se parte de un agrupamiento en el que cada item es un grupo y se van construyendo nuevas soluciones uniendo grupos en otros más amplios, se tiene un algoritmo de tipo *aglomerativo*, si el proceso es el contrario, dividiendo el espacio total en grupos que sucesivamente se van haciendo menores y más compactos tendremos un algoritmo *divisivo*.

Otras posibles clasificaciones se pueden encontrar en [15]. Obviamente una técnica concreta reúne un conjunto de características de acuerdo con las divsiones antes citadas. Por ejemplo la mayoría de las técnicas de agrupamiento jerárquico se han recogido bajo el acrónimo de técnicas SHAN (Secuencial,Aglomerativo,Hierarchical, Nonoverlaping).

1.3 Los datos de partida

Los algoritmos de agrupamiento reunen los items basándose en índices de proximidad (o distancia) entre ellos; pero la información de partida puede estar representado de dos formas:

- Por medio de una *matriz de patrones*, donde cada item de un conjunto que contiene a n de ellos, está representado por un conjunto de m medidas (atributos, puntuaciones..), cada item

está entonces reprsentado por un patrón o vector n-dimensional, y el conjunto en si mismo se ve como una matriz $\mathcal{X} = [x_{ij}]$ de dimensiones $n \times m$, cada fila $\mathbf{x_i}$ de esta matriz es un patrón y cada columna $\mathbf{y_j}$ se denomina *factor* o medida.

- Por medio de una *matriz de proximidad*, se trata de una matriz $n \times n$ donde el valor de la casilla *ik* representa una medida de la proximidad (distancia) entre el patrón i y el k. Habitualmente la matriz de proximidad se calcula a partir de la matriz de patrones; pero en ciertas aplicaciones psicométricas y sociológicas es posble que los datos se recojan directamente en términos de concordancias y se tenga de partida una matriz de proximidad.

En los siguientes párrafos comentamos los problemas de obtención de ambos matrices

1.3.1 La matriz de patrones

El problema de elegir una representación adecuada de los datos de partida para obtener un agrupamiento no es en absoluto sencillo. De hecho la preparación de datos es una de las tareas más importantes en todo proceso de extracción de conocimiento y puede dar origen a muchos errores. En [20] se puede encontrar un capítulo completo dedicado a la preparación de datos.

Existen distintos tipos de factores para representar un patrón. No hay que olvidar que un item pueder ser un objeto físico (persona, pixel,etc.)o un ente abstracto (estilos de escritura, opiniones políticas etc.) y que cualquier medida de estos puede ser usada para agruparlos. Tenemos medidas que pueden ser cuantificadas (factores cuantitativos) y medidas que no representan forma alguna de cantidad (factores cualitativos) y ambas divisiones presentan a su vez distintos tipos según la clase de valores que presentan:

Los factores cuantitativos se dividen en:

Factores con valores continuos , por ejemplo el peso de una persona, o el nivel de sodio en un suelo.

Factores con valores discretos , por ejemplo el número de ordenadores de un centro. Un caso importante particular de estos son *factores binarios* que sólo toman los valores 0 o 1 y que representa la presencia o ausencia de una determinada característica.

Factores con valores intervalares , por ejemplo la duración de un suceso.

Los factores cualitativos se dividen en:

Factores con valores nominales o no ordenados , por ejemplo el color de un suelo, o el diagnóstico de un enfermo.

Factores con valores ordinales , por ejemplo el rango de un militar o el nivel de gravedad de un enfermo.

No siempre es posible elegir los datos que intervienen en nuestro problema y reconocer su tipo es esencial para diseñar una medida de proximidad adecuada e interpretar los resultados del agrupamiento. Es necesario mencionar que existen muchos trabajos acerca de la selección de los factores para procesos de clasificación supervisada, pero que para agrupamiento se hace necesario un proceso de ensayo-error ya que no se tiene ninguna información a priori para llevar a cabo un filtrado de dichos factores.

NOMBRE	EXPRESION
Euclídea o norma-l_2	$d_2(i,k) = [\sum_{j=1}^{m}(x_{ij} - x_{kj})^2]^{1/2} = [(\mathbf{x_i} - \mathbf{x_k})^{\mathbf{T}}(\mathbf{x_i} - \mathbf{x_k})]$
Manhattan o norma-l_1	$d_1(i,k) = \sum_{j=1}^{m} \|(x_{ij} - x_{kj}\|$
norma del supremo	$d_\infty(i,k) = sup_{j\in\{1,2..m\}}\|(x_{ij} - x_{kj}\|$
Minkowski o norma-l_p	$d_p(i,k) = \sum_{j=1}^{m}[\|x_{ij} - x_{kj}\|^p]^{1/p}$
Distancia de Mahalanobis	$d_M = [(\mathbf{x_i} - \mathbf{x_k})^{\mathbf{T}}\mathbf{\Sigma^{-1}}(\mathbf{x_i} - \mathbf{x_k})]$ Σ es la covarianza muestral o una matriz de covarianza intra-grupos

Tabla 1.1: Distintas funciones de distancia

1.3.2 Indices de proximidad: distancias y semejanzas

En este apartado presentamos un resumen de las formas más conocidas de construir una matriz de proximidad a partir de una matriz de patrones. Para más detalles acerca de este tema se recomienda consultar [15, 8].

Definición 1 ***Indice de proximidad*** *Consideremos un conjunto de n patrones que notaremos por* $i, l, k.. \in I = \{1, 2, ..., n\}$ *,decimos que* $d : I \times I \longrightarrow R$ *es un índice de proximidad si y sólo si verifica:*

1. *(a)* $\forall i \in I\ d(i,i) = 0$ *(medidas de disimilaridad o distancia)*
   (b) $: \forall i \in I\ d(i,i) \geq max_{k\in I} d(i,k)$ *(medidas de similaridad)*
2. $d(i,k) = d(k,i)\ \forall i, k \in I$
3. $d(i,k) \geq 0 \forall i, k \in I$

Los índices de proximidad generalizan conceptos conocidos, algunos de los cuales pasamos a discutir.

Definición 2 ***Distancia*** *Se dice que un índice de proximidad d es una distancia si y sólo si verifica:*

1. *Las propiedades 1.a, 2 y 3 de la definición 1*
2. $d(i,k) \leq d(i,l) + d(l,k)\ \forall i, l, k \in I$

Las distancias son especialmente utilizadas en el caso de factores cuantitativos continuos, la tabla 1.3.2 muestra las funciones de distancia más habituales.

La distancia Euclídea es la más intuitiva y trabaja muy bien cuando se tienen grupos "compactos" y "aislados". La distancia de Mahalanobis generaliza la distancia euclídea, mientras que la distancia de Minkowski generaliza a todas las demás.

El principal inconveniente de las métricas de Minkowski es que, en general, todas ellas dan un gran peso a los factores con valores muy grandes, problema que se soluciona con una normalización. Otro problema que presentan los factores continuos es la posible existencia de correlación entre ellos, lo que se puede paliar utilizando la distancia de Mahalanobis o reduciendo previamente el espacio por medio de un análisis factorial que nos devuelva un nuevo conjunto de factores independientes. Para más detalles ver [15].

Existen también funciones de distancia basadas en la distancia de dos distribuciones de probabilidad y funciones de distancia basadas en el coeficiente de correlación que se aplican al espacio de factores, no al de patrones.

NOMBRE	EXPRESION
Indice de Jaccard	$\frac{n_{IK}}{n_{IK}+n_{iK}+n_{Ik}}$
Indice de acoplamiento simple	$\frac{n_{IK}+n_{ik}}{m}$
Indice de Russell	$\frac{n_{IK}}{m}$
Indice de Dice	$\frac{2n_{IK}}{2n_{IK}+n_{iK}+n_{Ik}}$
	$\frac{2(n_{IK}+n_{ij})}{m+n_{iK}+n_{Ik}}$
	$\frac{n_{IK}}{n_{IK}+2(n_{iK}+n_{Ik})}$
	$\frac{(n_{IK}+nik)}{m+n_{iK}+n_{Ik}}$

Tabla 1.2: Distintos índices de semejanza

Definición 3 ***Indice de semejanza*** *Se dice que un índice de proximidad s es una función de semejanza si y sólo si verifica:*

1. $\forall i \in I\ d(i,i) = 1$
2. *Las propiedades 2 y 3 de la definición 1*

Obviamente se puede obtener un índice de semejanza a partir de una distancia, tomando:

$$\forall i,k \in I\ s(i,k) = 1 - (d(i,k)/D) \text{ siendo } D = max_{i,k}d(i,k)$$

La mayoría de los índices de semejanza, no basados en distancia, se han definido para patrones cuyos factores son binarios, es decir aquellos que sólo toman valores 0 o 1 y que sirven para reflejar la presencia o ausencia de una determinada característica. La tabla 1.3.2 nos muestra los más utilizados con la siguiente notación (para $\mathbf{x_i}$ y $\mathbf{x_k}$, patrones formados por m variables binarias):

- n_{IK} número de factores que toman el valor 1 en $\mathbf{x_i}$ y $\mathbf{x_k}$
- n_{ik} número de factores que toman el valor 0 en $\mathbf{x_i}$ y $\mathbf{x_k}$
- n_{iK} número de factores que toman el valor 0 en $\mathbf{x_i}$ y 1 en $\mathbf{x_k}$
- n_{ik} número de factores que toman el valor 1 en $\mathbf{x_i}$ y 0 en $\mathbf{x_k}$

Un índice de semejanza bastante utilizado cuando se trabaja con documentos es la denominada medida del coseno. Se parte de la representación de cada documento como un vector de frecuencias de aparición de términos, de forma que si $t_1 = (t_{11}...t_{1d})$ y $t_2 = (t_{21}...t_{2d})$ son dos vectores de documentos en un espacio d-dimensional, entonces la medida de semejanza entre ellos se calcula como :

$$cos(t_1,t_2) = (t_1 \odot t_2)/|t_1||t2|$$

donde $\odot$ representa el producto escalar y $|.|$ el módulo:

$$cos(t_1,t_2) = \frac{\sum_{j=1}^{d} t_{1j}t_{2j}}{\sqrt{\sum_{j=1}^{d} t_{1j}^2}\sqrt{\sum_{j=1}^{d} t_{2j}^2}} \tag{1.1}$$

	1	2	3	4	5	6	7	8	9	10
1	,000	1,490	5,440	2,440	8,290	14,690	22,690	21,640	38,090	36,040
2	1,490	,000	1,250	4,250	8,000	9,000	25,000	21,250	41,000	36,250
3	5,440	1,250	,000	9,000	11,250	7,250	31,250	25,000	48,250	41,000
4	2,440	4,250	9,000	,000	2,250	10,250	10,250	10,000	21,250	20,000
5	8,290	8,000	11,250	2,250	,000	5,000	5,000	3,250	13,000	10,250
6	14,690	9,000	7,250	10,250	5,000	,000	16,000	9,250	26,000	18,250
7	22,690	25,000	31,250	10,250	5,000	16,000	,000	1,250	2,000	2,250
8	21,640	21,250	25,000	10,000	3,250	9,250	1,250	,000	4,250	2,000
9	38,090	41,000	48,250	21,250	13,000	26,000	2,000	4,250	,000	1,250
10	36,040	36,250	41,000	20,000	10,250	18,250	2,250	2,000	1,250	,000

Tabla 1.3: Ejemplo de matriz de proximidad, basada en distancias

Tanto las distancias como las semejanzas se utilizan para obtener la matriz de proximidad de un conjunto de patrones que es el punto de partida para un proceso de agrupamiento. Ahora bien, cada uno de los enfoques corresponde a un tipo de factor, claramente las distancias se utilizarán en presencia de factores continuos, y pueden usarse con valores enteros e incluso ordinales asimilables a enteros. Las semejanzas son adecuadas cuando se trabaje con factores binarios y pueden utilizarse con factores nominales no ordinales tranformándoles en un conjunto de factores binarios.

Es importante tener en cuenta que puede ser problemático mezclar ambos enfoques directamente, cuando se tienen los dos tipos de factores. No obstante se han propuesto técnicas mixtas tal como la que se presenta en [22].

A partir de ahora supondremos que tenemos construida la matriz de proximidad. sSe puede encontrar un análisis acerca de tipos de datos, cambios de escala, etc. en [15, 20]

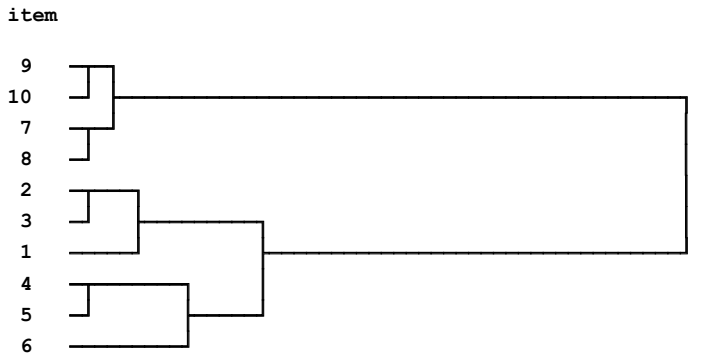
1.4 Técnicas de agrupamiento jerárquico

1.4.1 Ideas básicas

Como hemos comentado anteriormente, un agrupamiento jerárquico es una sucesión de particiones "anidadas" donde cada grupo de patrones perteneciente a una determinada partición está totalmente incluido en algún grupo de la partición siguiente, esta estructura tiene una representación gráfica muy intuitiva que se denomina *Dendrograma*. En la figura 1.2 se muestra el agrupamiento jerárquico de 10 patrones o items, cuya matriz de proximidad se refleja en la tabla 1.3, obtenido por el método de enlace completo que describiremos posterioremente. El agrupamiento se describe por medio del dendrograma donde se presenta cómo se van uniendo los distintos patrones en grupos: partiendo de una primera partición donde se consideran todos los items aislados, en una primera etapa se agrupan los items 2 y 3, 7 y 8 , y 9 y 10, en la segunda etapa se agrupan los items 4 y 5, en la tercera se unen los grupos {7,8} y{8,10} y así sucesivamente. Obviamente el criterio de unión se obtiene a partir de la matriz de distancia, mediante procesos algoritmicos que pasamos a describir.

1.4.2 Algoritmos para el agrupamiento jerárquico

El propio concepto de agrupamiento jerárquico nos indica que los algoritmos para obtenerlo serán de tipo aglomerativo, de forma que partiendo de una partición en la que cada item es un



1. Nivel 0
   Partición={{1},{2},{3},{4},{5},{6},{7},{8},{9},{10}}
2. Nivel 1.25
   Partición={{1},{2,3},{4},{5},{6},{7,8},{9,10}}
3. Nivel 2.25
   Partición={{1},{2,3},{4,5},{6},{7,8},{9,10}}
4. Nivel 4.25
   Partición={1,{2,3},{4,5},6,{7,8,9,10}}
5. Nivel 5.40
   Partición={{1,2,3},{4,5},6,{7,8,9,10}}
6. Nivel 10.25
   Partición={{1,2,3},{4,5,6},{7,8,9,10}}
7. Nivel 14.69
   Partición={{1,2,3,4,5,6},{7,8,9,10}}
8. Nivel 48.25
   Partición={1,2,3,4,5,6,7,8,9,10}

Figura 1.2: Ejemplo de agrupamiento por enlace completo

grupo se van obteniendo nuevas particiones uniendo grupos entre sí.

Un primer enfoque para el desarrollo de estos algoritmos es el basado en grafos donde se considera que cada item es un vértice de un grafo y se van generando particiones, conectando los vertices de menor distancia. Aparecen entonces dos formas de agrupamiento (ver [15]):

Agrupamiento de enlace simple *(Single-link clustering)* Los grupos se obtienen buscando las componentes conexas de grafo y se termina cuando todos los vértices están conectados.

Agrupamiento de enlace completo *(Complete-link clustering)* Los grupos se obtienen buscando los subgrafos completamente conectados (cliques), es decir, tendremos un grupo de dos vertices si hemos colocado una arista, tendremos un grupo de tres vértices si hemos colocado las tres aristas que los unen etc. El algoritmo termina cuando hemos incluido todos los vértices en un grupo.

Un enfoque diferente y más general para obtener agrupamientos jerárquicos es debido a Jhonson (1967) y se basa en sucesivas transformaciones de la matriz de proximidad, que para este algoritmo siempre es de distancia, reduciendo la dimension de la misma siempre que se forme un nuevo grupo. La idea es que se trabaje con una *matriz de distancia entre grupos*, y que esta se vaya calculando iterativamente a partir de la matriz de la etapa anterior. La generalidad de este enfoque radica en el hecho de que la distancia entre grupos se puede calcular de distintas formas y que dependiento de cómo se calcule dicha distancia aparecen diferentes formas de agrupamiento. Lance y William

METODO	a(R)	a(S)	b	c
Enlace Simple	1/2	1/2	0	$-1/2$
Enlace completo	1/2	1/2	0	1/2
Media de grupos	n_R/n_K	n_S/n_K	0	0
Centroide	$n_R/(n_R+n_T)$	$n_S/(n_s+n_T)$	$-(n_R n_S)/n_K^2$	0
Método de Ward	$(n_s+n_T)/(n_K+n_T)$	$(n_R+n_T)/(n_K+n_T)$	$-(n_T)/(n_K+n_T)$	0

Tabla 1.4: Coeficientes para el agrupamiento jerárquico

[18], han establecido la forma general de dicho cálculo y los parámetros que dan lugar a los diversos enfoques

La descripción formal del algoritmo es la siguiente:

1. Sean m=0, $D_m = D$ la matriz de distancia de partida, $\mathcal{C}_m = \{\{1\}, .., \{n\}\}$ el agrupamiento inicia y $L(m) = 0$ el nivel al cual se hace este agrupamiento.

2. Sean R y S aquellos grupos de $\mathcal{C}_m$ que van a fusionarse en esta etapa, ya que tienen distancia mínima:

   - $L(m+1) = D_m(R, S)$
   - Hacer $\mathcal{C}_{m+1} = \mathcal{C}_m \cup (R \cup S) - R - S$, es decir sustituir R y S por K en el agrupamiento. formar un nuevo grupo $K = R \cup S$ de y transformar la matriz D_m de la siguiente manera.
     - Eliminar la fila y columna correspondiente a uno de los grupos anteriores y asignar la fila y columna del otro al nuevo grupo K.
     - Para todo grupo T perteciente a $\mathcal{C}_m$ disinto de $K = R \cup S$, hacer:

$$D_{m+1}(K,T) = a(R)D_m(R,T) + a(S)D_m(S,T) + bD_m(R,S) + c|D_m(R,T) - D_m(S,T) \quad (1.2)$$

3. Hacer $m = m + 1$

4. Si se han unido todos los items parar, en caso contrario ir a 2

La tabla 1.4.2 nos muestra los coeficientes de la expresión 1.2 para los algoritmos aglomerativos de agrupamiento jerárquico más conocidos (n_X nota el numero de elementos que tiene el grupo X:

La elección de un método u otro depende de las propiedades que se busquen en los grupos y el resultado es muy dependiente del método, como se muestra en la figura 1.3 donde se recogen agrupamientos según los métods de enlace simple, del centroide y de Ward de mismo ejemplo que se presentón en el apartado anterior, cuya matriz de distancia es la tabla 1.3 y cuyo agrupamiento por medio de enlace completo se muestra en la figura 1.2.

En [15] se pueden encontrar un estudio teórico muy completo acerca de las propiedades del agrupamiento jerárquico, [20] recoge un interesante análisis acerca de las ventajas e inconvenientes de su uso. En el apartado 1.6 presentaremos algunas de las nuevas tendencias en agrupamiento jerárquico que permiten utilizar gran volumen de datos.

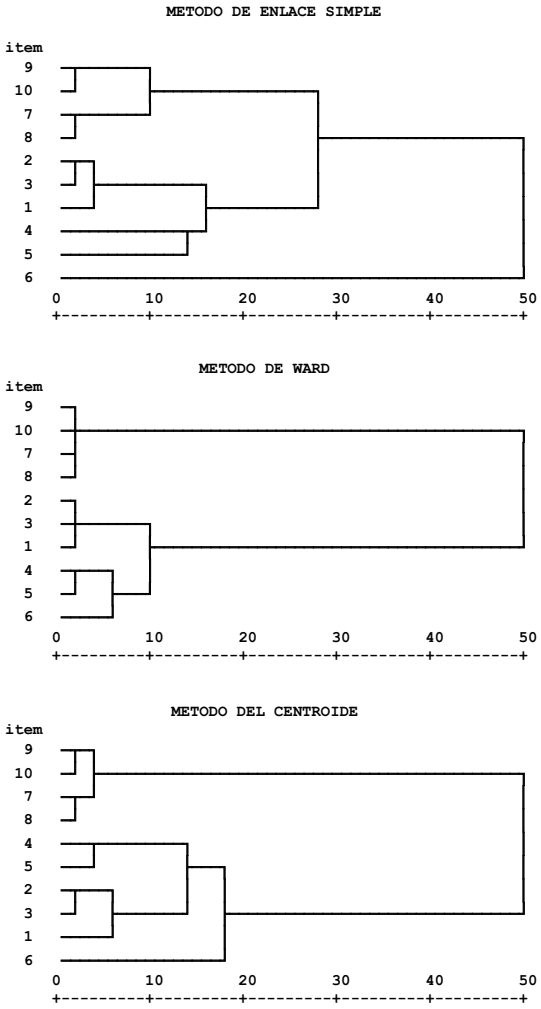


Figura 1.3: Ejemplos de distintos métodos de agrupamiento

1.5 Técnicas de agrupamiento particional

1.5.1 Ideas iniciales

El problema de agrupamiento particional puede formalizarse como sigue:

> Dados n patrones representados en un espacio d-dimensional en elque hay definida una distancia, determinar una partición de los mismos en K subconjuntos o grupos tales que los patrones incluidos en un grupo se parezcan más entre ellos de lo que se parecen a los clasificados en otros grupos.

El número K de grupos a generar, puede estar definido previamente o no, aunque la práctica generalidad de los algoritmos actualmente en uso suponen que es un parámetro que debe estar fijado, a partir de ahora nos referiremos exclusivamente a estos algoritmos. Algoritmos con número de clases no fijado aparecen descritos en [15], que contiene también un conjunto de citas al respecto. En cualquier caso lo que siempre hay que establecer es un criterio para medir tanto la "similaridad" o coherencia de un grupo como la de un conjunto de grupos (agrupamiento).

- Existen métodos basados en "criterios globales" que suponen que cada grupo está representado por un prototipo de modo que cada patrón se asigna al grupo cuyo prototipo esté más cercano. Se usan en este enfoque medidas de coherencia basadas en la distancia de cada patrón a su prototipo y dependiendo de la distancia que se considere aparecerán distintas medidas.

  - Para datos con atributos continuos, el prototipo de un grupo es habitualmente la media de los patrones que lo integran (*centroide*). El método de las k-medias que estudiaremos en el apartado 1.5.2 pertenece a esta categoria.
  - En el caso de factores categóricos se suele utilizar el patrón más representativo del grupo (*medoide*). Veremos algunas ideas acerca de estos métodos en el apartado 1.6

- Por el contrario, los métodos basados en "criterios locales" forman los grupos utilizando utilizando la estructura local de los datos, ejemplos de esta forma de trabajar son los métodos basados en la identificación de regiones de alta densidad de puntos o aquellos que asignan al mismo grupo un patrón y sus k-vecinos más cercanos. Uno de los métodos más conocidos dentro de este enfoque es DBSCAN [9], que veremos con más detalle en el apartado 1.5.3

En [14] puede encontrar un estudio bastante amplio de los "métodos de vecino más cercano" y en [20] se pueden encontrar técnicas basadas en identificación de regiones. Un resumen de los enfoques más actuales de las técnicas de k-medoides se encuentra en [2], que también incluye una amplia bibliografía acerca del tema. Estos temas se tratarán en el apartado 1.6

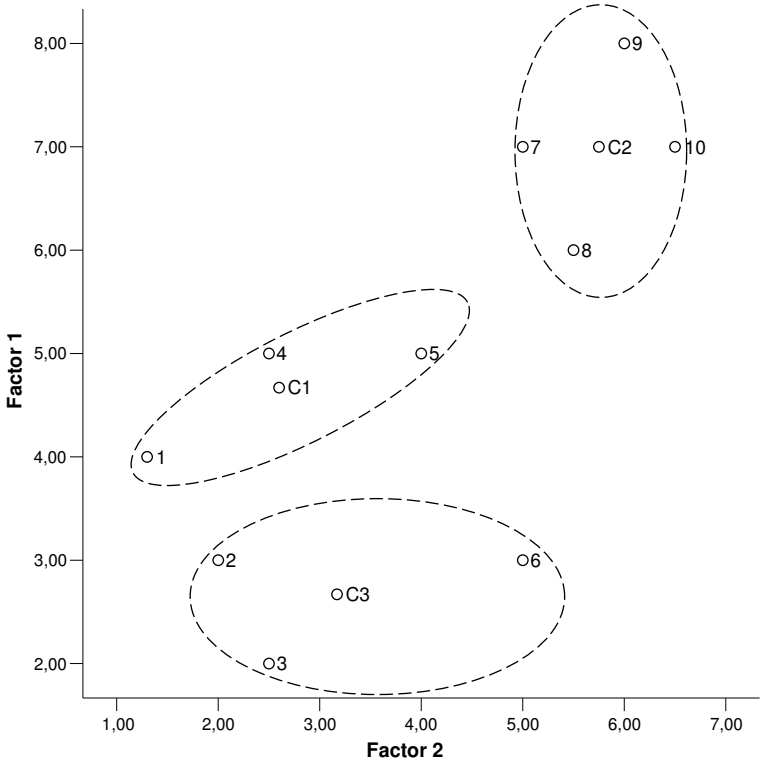


Figura 1.4: Ejemplo del método de las K-medias con 3 grupos

1.5.2 El método de las k-medias

Tal como ya hemos comentado el método de las K-medias es muy simple. Se necesitan como parámetros iniciales: el número de grupos K y K centroides iniciales que se pueden elegir por parte del usuario u obtenerse por medio de algún procesamiento previo. Se asigna entonces cada punto a su centroide más cercano y así se obtienen los grupos iniciales. A partir de estos grupos se recalculan los centroides y se hace una nueva reasignación. El proceso se vuelve a repetir hasta que los centroides no cambian. La descripción formal del algoritmo es la siguiente:

1. Sean $\{x_1...x_n\}$ n patrones de un espacio d-dimensional E, representados por matriz de datos $[x_{il}, i = \{1,..n\}$ y $l = \{1,..d\}]$. En el espacio E hay definida una función de proximidad $p(.,.)$ que, para fijar ideas, supondremos establecida en términos de disimilaridad (que puede ser o no una distancia).

   Elegir K y seleccionar un conjunto $c_1...c_K$ de centroides iniciales en el mismo espacio. Si $\{G_1,...G_k\}$ nota el conjunto de grupos que vamos a obtener, inicialmente $G_j = \emptyset\ \forall j \in \{1,...K\}$

2. $\forall i \in \{1,..n\}$:

   - calcular $j_i \in \{1,..K\}$ tal que:

   $$p(x_i, c_{j_i}) = min_{j\in\{1,..k\}}(p(x_i, c_j)) \tag{1.3}$$

   - Hacer $G_{j_i} = G_{j_i} \cup \{x_i\}$

3. Obtener los nuevos centroides haciendo:

   $$\forall j \in \{1,..K\}\ \forall l \in \{1,..,d\}\ cn_{jl} = \sum_{x_i \in G_j} x_{il}/|G_j|$$

   donde $|G_j|$ representa el cardinal del grupo G_j, es decir, el número de elementos que contiene.

4. Si $cn_j = c_j \forall j \in \{1,..K\}$, parar. En caso contrario:

   - Hacer $cn_j = c_j \forall j \in \{1,..K\}$
   - Hacer $G_j = \emptyset\ \forall j \in \{1,...K\}$
   - Ir a 2.

La versión original de este método consideraba la distancia euclídea, dado lo cual, y teniendo en cuenta que los centriodes están definidos como la media de los elementos de cada grupo, en cuyo caso lo que se minimiza en la expresión 1.3 es la desviación con respecto a la media de los elementos del grupo, o lo que también de denomina la suma de los errores cuadráticos. No obstante se pueden considerar cualquier función de distancia o índice de semejanza, cambiando en este último caso, el criterio de minimización por el de maximización. De hecho se han utilizado con bastante frecuencia la distancia de Manhatann (ver tabla 1.3.2), considerando como centroide la mediana de los grupos, y la medida de similaridad del coseno, (ver expresion 1.1).

Ya hemos comentado con anteriroridad que el método de las K-medias es probablemente el más popular y aplicado de todos los particionales. No obstante hay que tener en cuenta que es fuertmente dependiente los parámetros de entrada, es decir, del número K de grupos considerados y de los centroides que se eligen de partida.

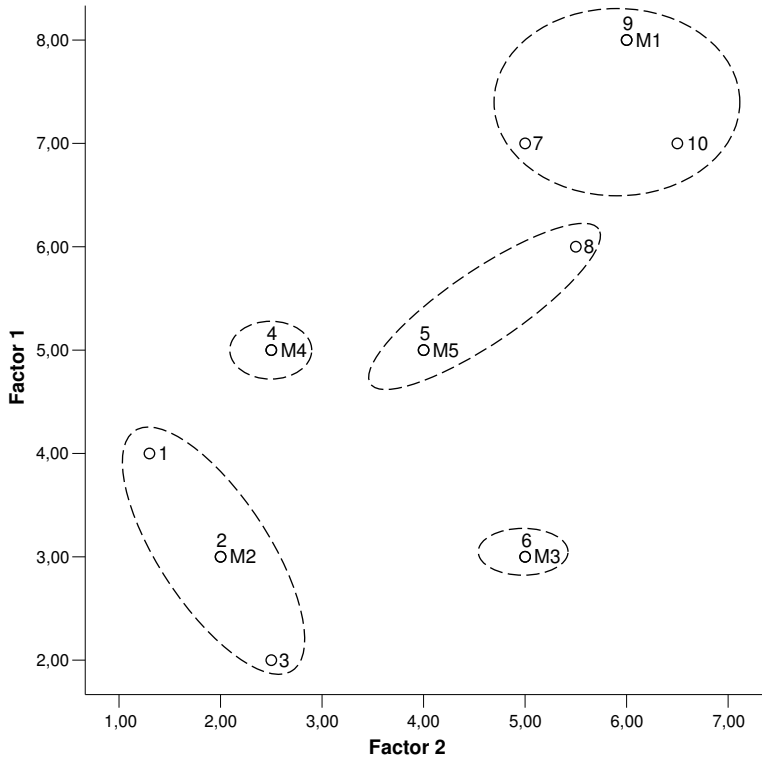


Figura 1.5: Ejemplo del método de las K-medias con 5 grupos

En las figuras 1.4 y 1.5 se presentan dos ejemplos de agrupamiento de 10 items cuya matriz de distancia aparece en la tabla 1.3. En el primero se consideran 3 grupos y en el segundo 5, y como puede comprobarse hay una gran diferencia en los resultados.

Dada la gran dependencia de este método de sus parámetros iniciales una buena medida de la bondad del agrupamiento es justamente la suma total de la proximidad que se minimiza:

$$SSE = \sum_{j=1}^{j=K} \sum_{x_i \in G_j} p(x_i, c_j)/n$$

En el caso de que trabajemos con distancia euclídea, esta expresión es la del error cuadrático global. Obviamente este valor será tanto menor cuanto más grupos consideremos; pero esta opción puede no ser la más adecuada, por ello existen algunas técnicas de postprocesamiento [20], que permiten mejorar el agrupamiento obtenido y técnicas de detección de "items extraños" (outliers). Tal y como se comenta en [13] un procedimiento para fijar los parámetros de partida consiste en realizar un agrupamiento jerárquico previo y partir con alguno de los agrupamientos que proporcione.

A título de ejemplo hemos seleccionado el agrupamiento en tres grupos

$$P_1 = \{\{1,2,3\},\{4,5,6\},\{7,8,9,10\}$$

proporcionado por el enfoque jerárquico usando el método de enlace completo (ver figura 1.2) y el de cinco grupos

$$P_2 = \{\{1,2,3\},\{4\},\{5\},\{6\},\{7,8,9,10\}\}$$

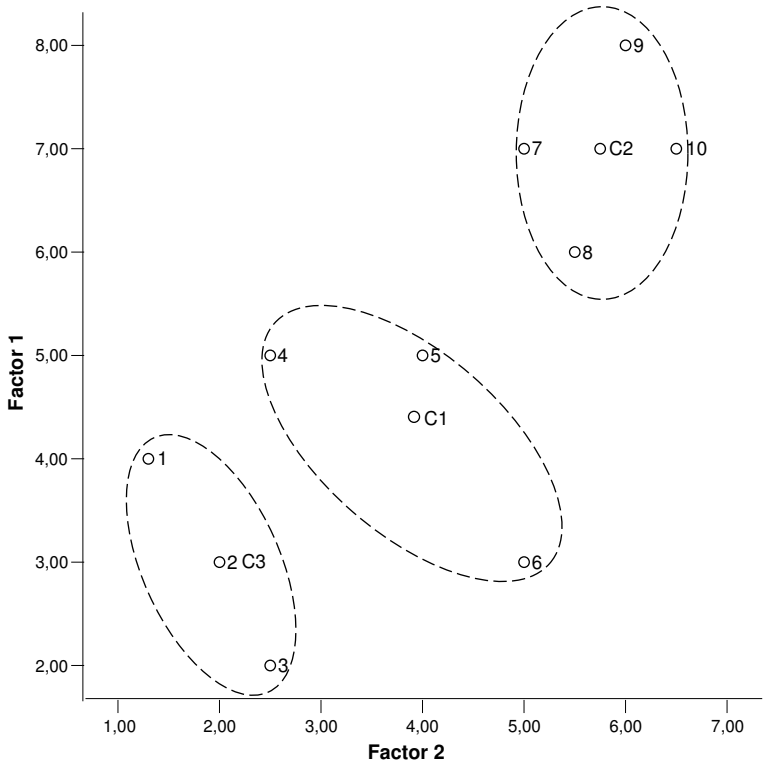


Figura 1.6: Ejemplo del método de las K-medias con 3 grupos, datos de partida elegidos

N. de grupos	Sin selección previa	Con selección previa
3	1.083	0.980
5	0.650	0.590

Tabla 1.5: Ejemplos de medida SSE en distintas aplicaciones del método de las k-medias

que es uno de los agrupamientos resultado de aplicar el método de enlace simple (ver figura 1.3), las figuras 1.6 y 1.7 , muestran los resultados obtenidos después de aplicar el método de las k-medias, tomando como centroides iniciales los de estos agrupamientos. Como puede verse, el algoritmo no ha modificado los grupos de partida, y los resultados son diferentes de los que se muestran en las figuras 1.4 y 1.5, en los que la selección de los centroides se hizo de forma aleatoria. La tabla 1.5.2 muestra los valores de SSE para cada caso, como puede verse el valor disminuye con el número de grupos y la elección inicial usando un agrupamiento obtenido por un método jerárquico mejora esta medida de bondad.

1.5.3 DBSCAN: un método basado en el análisis de densidad

Tal como se ha comentado, los métodos de agrupamiento basados en la densidad analizan regiones del espacio de alta densidad que están separadas por otras de baja densidad. Obviamente todos los métodos de esta categoria se basan en el concepto de "densidad" de una región, habiendose desarrollado diversas formas de medir dicha densidad, algunas de las cuales tienen fundamento estadístico. Parten del conocimiento de la distribución de probabilidad conjunta entre items y grupos

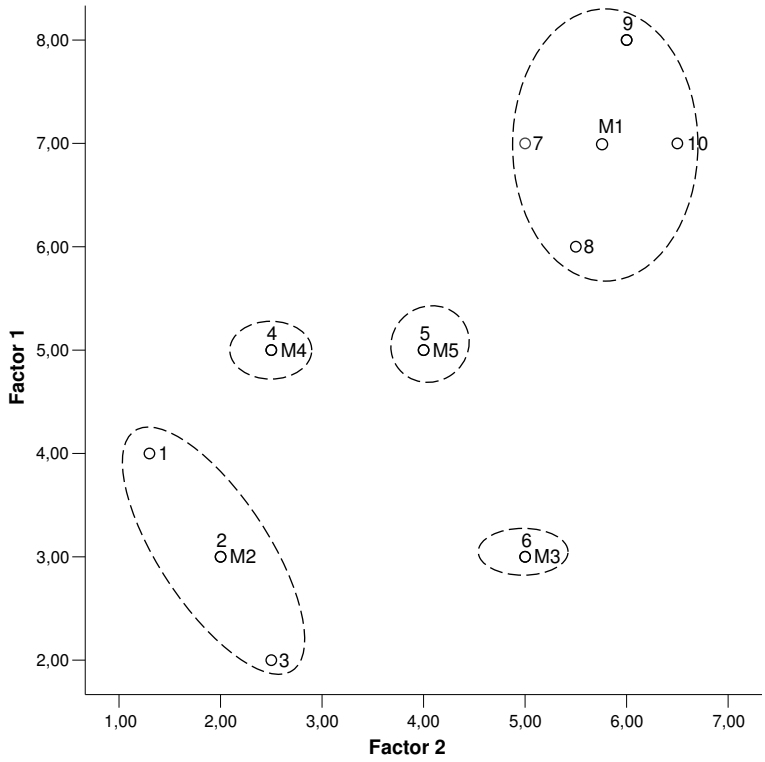


Figura 1.7: Ejemplo del método de las K-medias con 5 grupos, datos de partida elegidos

para obtener la distribución de probabilidad asociada a cada grupo y a partir de ella desarrollan un procedimiento de asignación de items. Los métodos basados en Análisis Discriminante y los de Mezcla de densidades (ver [11],[12]) pertenecen a esta categoria

De forma diferente trabaja DBSCAN, uno de los más interesantes y utilizados, que si bien es bastante simple nos va a servir para ilustrar muchas de las ideas subyacentes en este enfoque.

DBSCAN utiliza la denominada *densidad basada en centros*, la cual se estima para un punto (item o patrón) concreto contando el número de puntos que caen dentro de un entorno centrado en él de radio *eps* fijado . En la figura 1.8, se presenta una distribución de puntos en un espacio bidimensional y varios entornos con radio eps centrados en algunos puntos. Podemos ver que la densidad en el punto A es de 7 mientras que la que hay en D es de 2. Está claro que la densidad depende del radio que se tome, si este es suficientemente grande, la densidad de cada punto será n, (el número total de puntos); por el contrario si es suficiente pequeño la densidad tendrá valor 1 en cualquier punto.

Fijado un valor de radio *eps* y un número de puntos mínimo: $MinPt$, suficiente para considerar que un entorno tiene la densidad adecuada para formar un grupo, todo punto del espacio de patrones se puede clasificar en como:

Punto Núcleo. Es el centro de un entorno de radio *eps* que tiene más de $MinPt$ puntos. En la figura 1.8 son punto núcleo el A y el B. Se considera que pertenece al interior de un grupo.

Punto Frontera. Es aquel que se encuentra en un entorno de radio *eps* que tiene como centro un punto núcleo. En la figura 1.8 se pueden observar varios puntos frontera. Puede ocurrir que un punto frontera pertenzca al entorno de varios puntos núcleo, este es el caso del punto C

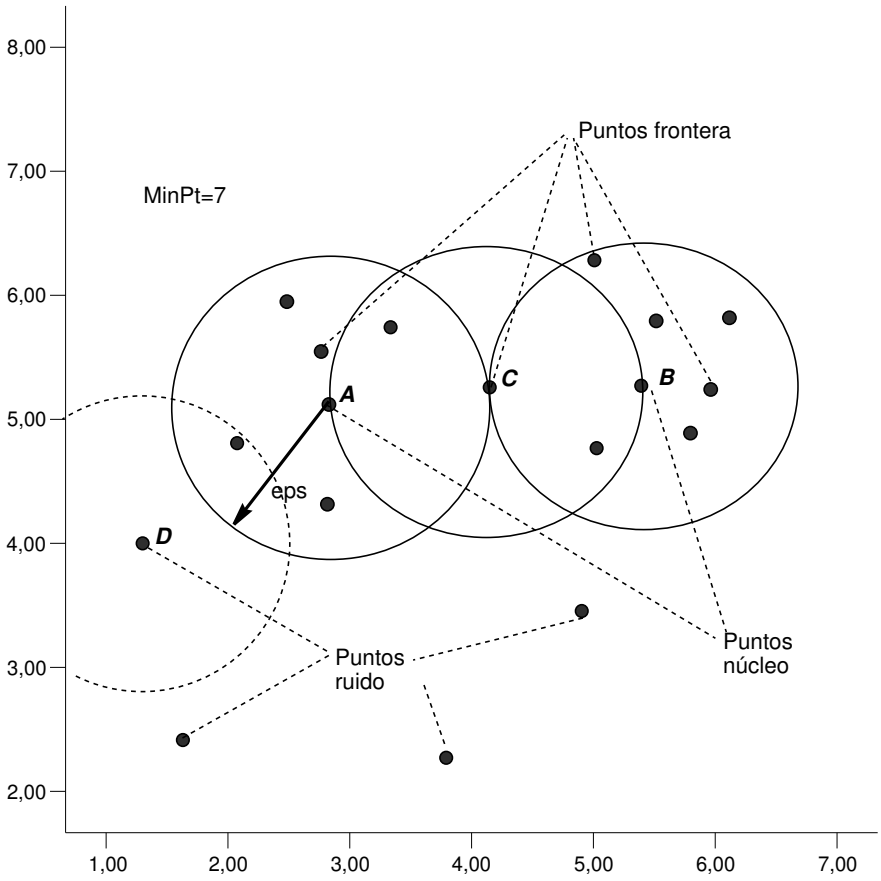


Figura 1.8: Conceptos básicos referentes a DBSCAN

en la figura 1.8.

Punto ruido. Es aquel que no es núcleo, ni frontera. Se supone que va a estar en regiones muy poco densas y que no va a formar parte de ningún grupo. En la figura 1.8 aparecen varios puntos frontera, entre ellos el D.

Con base en estos tres tipos de punto, el ciclo básico de trabajo de DBSCAN es muy sencilla:

- Se colocan en un mismo grupo todos los puntos núcleo que distan entre sí menos de eps, aceptando un criterio de transitividad, es decir si la distancia entre dos puntos núcleo $f(n_1, n_2) \leq eps$ y para otro punto núcleo $f(n_2, n_3) \leq eps$ entonces n_1, n_2 y n_3 pertencen al mismo grupo. Se dice entonces que n_1 y n_2 (o n_2 y n_3) son "directamente densidad alcanzables" mientras que n_1 y n_3 se dice que son "densidad alcanzables"

- También se asignan al mismo grupo todos los puntos frontera asociados a cada punto núcleo. Habrá que dar algún criterio para el caso en que un cierto punto pertenzca a los entornos de dos núcleos que no están en el mismo grupo.

- Se eliminan todos los puntos ruido.

DBSCAN funciona realmente de forma iterativa, ya que, fijados sus parámetros eps y $MinPt$, va considerando punto a punto, estableciendo su entorno de radio eps y viendo si es o no núcleo,

en el caso de que lo sea se construye un grupo con dicho entorno y se buscan otros núcleo sean densidad alcanzables a partir de él, si existe alguno, el grupo generado inicialmente se une a aquel al que pertenzca este núcleo. El proceso termina cuando ningún punto puede ser añadido a ningún grupo.

DBSCAN es un algoritmo potente y sencillo que puede optimizarse para espacios de baja dimensionalidad y que produce grupos complejos para los cuales no hay ninguna hipótesis de "centralidad" o "globularidad" como en el caso del método de las k-medias. No obstante depende fuertemente de los valores fijados para *eps* y $MinPt$. Este es un problema general de todos los métodos presentados; pero en el caso de DBSCAN el problema se agudiza ya que cuando se tiene un gran volumen de datos y una alta dimensionalidad puede ocurrir que haya zonas del espacio de patrones muy densas, con grupos que incluyan muchos puntos y otras con una densidad más baja y con grupos menos densos. Si se toman los mismos parámetros para ambas zonas los puntos de la segunda se considerarán como ruido, con lo que se pierde mucha información.

Existen generalizaciones de DBSCAN que permiten evitar en lo posible estos problemas. OPTICS [1] trabaja con regiones de densidad variable, considerando valores de *eps* menores. Un estudio detallado del OPTICS se puede encontrar en [13]. Algunas consideraciones adicionales acerca de la complejidad de DBSCAN y sus problemas de uso se pueden encontrar en [20]. Un estudio bastante completo de las extensiones de DBSCAN puede encontrarse en [2] y en [13].

1.6 Nuevos resultado y extensiones

1.6.1 Resultados recientes sobre agrupamiento jerárquico

Las técnicas de agrupamiento jerárquico siempre han tenido el inconveniente de ser muy costosas desde el punto de vista computacional como consecuencia de operar con la matriz de proximidad completa. Este inconveniente se agrava cuando se trabaja con un gran cantidad de patrones, ya que, en su versión clásica, , estas técnicas obtienen el conjunto de todos los posibles agrupamientos, desde el inicial con n grupos hasta el último con un sólo grupo. Las primeras versiones que tratan de evitar estos inconvenientes son los algoritmos llamados DIANA y AGNES [17]. El primero es de tipo divisivo y el segundo aglomerativo pero en ambos se especifica el número de grupos que se desea como una condición de terminación de modo que los algoritmos finalizan cuando se alcanza este número.

También se considera un gran inconveniente de los métodos jerárquico clásicos el hecho de que, cuando se toma la decisión de unir dos grupos (en los aglomerativos) o dividir uno (en los divisivos), no se puede volver atrás, ya que cada nuevo paso de las iteraciones parte del agrupamiento generado en el paso anterior. Este procedimiento tan rígido puede llevar a resultado erróneos, además de que la forma en que se toman las decisiones de unir o separar implica evaluar un buen número de items o grupos y por tanto gran coste computacional. Un enfoque muy interesante que se ha seguido para resolver estos problemas es mejorar la calidad de los grupos obtenidos utilizando otras técnicas de agrupamiento, realizando de este modo un proceso de múltiple fases. A continuación comentamos resumidamente los métodos más conocidos que siguen este enfoque.

BIRCH Balanced Iterative Reducing and Clustering using Hierarchies [23]. Inicialmente particiona los patrones de forma jerárquica utilizando estructuras de árboles y posteriormente aplica otros algoritmos de agrupamiento para refinar el resultado. Este algoritmo se basa en el uso de resúmenes de los datos (estadísticos suficientes) para sustituir a los datos originales y es facilmente escalable y muy usado en problemas de Minería de Datos. Se introducen dos conceptos:

- El descriptor del grupo, **CF,**(clustering feature). Para cada grupo con M elementos CF se define como un triple:

$$CF = (M, LS, SS)$$

donde Ls es la suma de las componentes del vector y SS la suma de los cuadrados de sus componentes. Un CF almacena un resumen estadístico de la información asociada a un grupo.

- El árbol de descriptores , **CF-tree**, que no es más que un árbol equilibrado cuyos nodos son CF y que se va construyendo conforme los grupos se van generando. El criterio de construcción de este árbol contempla dos parámetros: el factor de ramficación B, que especifica el número máximo de hijos que puede tener un nodo y el umbral T que fija el diámetro máximo de los grupos almacenados en las hojas del árbol.

El algoritmo trabaja en dos fases, en la primera se explora la base de datos para construir un CF-tree inicial en memoria principal, que puede verse como una compresión multinivel de los datos y que trata de preservar la estructura de agrupamiento inherente a los datos. En la segunda se aplica un algoritmo de agrupamiento a las hojas del CF-tree.

BIRCH trata de producir los mejores grupos con los recursos computacionales disponibles y el análisis experimental ha probado la escabilidad lineal del algoritmo con respecto al número de items y la buena calidad de los resultados obtenidos. No obstante dado que cada nodo de un CF-tree sólo puede almacenar un úmero limitado de items, es posible que no corresponda a un grupo real. Además en el caso de que los grupos no sean esféricos puede trabajar mal ya que está basado en la noción de diámetro para controlar la frontera de los grupos.

CURE Clustering Using REpresentatives [11]. La mayoría de los algoritmos de agrupamiento favorecen la creación de grupos de forma esférica y se ven muy afectados por la existencia de items extremos (outliers). CURE, que integra algoritmos jerárquicos y particionales, intenta resolver estos inconvenientes.

Este método emplea un enfoque muy novedoso de agrupamiento jerárquico que se encuentra a mitad de camino entre los métodos de enlace ponderado y de enlace completo, ya que en lugar de utilizar un único punto, o todos ellos para representar un grupo, se elige un número fijo de puntos representativos. Estos puntos se generan elegiendo primeramente un conjunto de puntos bien distribuidos sobre el grupo y posteriormente reduciendo la distancia de estos al centro del grupo un determinado factor (factor de reducción). Los grupos con puntos representativos más cercanos se unen en cada paso del algoritmo.

Dado que usa más de un punto representativo por grupo CURE se ajusta bien a la geometría de formas no esféricas y elimina el efecto de enlace simple. El proceso de reducción permite eliminar puntos extremos. El algoritmo trabaja con grandes bases de datos utilizando un proceso adicional de muestreo. La muestra obtenida se divide en un número de grupos mayor que el que se desea obtener, con estos grupos se realiza la determinación de los puntos representativos, una reducción para eliminar extremos y un proceso aglomerativo hasta obtener el número de grupos deseados.

CURE produce grupos de alta calidad en presencia de extremos y formas complejas de distintos tamaños y es escalable a grandes bases de datos sin sacrificar la calidad de los grupos. Además necesita relativamente pocos parámetros: tamaño de la muestra, número de grupos deseados, y factor de reducción y no es muy sensible a la variación de los mismos.

Otro algoritmo aglomerativo desarrollado por los mismos autores [12] es **ROCK**, orientado al manejo de atributos categóricos

CHAMELEON [16] . Explora un modelo dinámico de agrupamiento jerárquico. En su proceso de agrupamiento se unen dos grupos si su interconectividad y su cercaní se corresponde con la conectividad interna y la cercanía de los items que están en ellos. El proceso de unión basado en un modelo dinámico facilita el descubrimiento de grupos naturales y homogéneos, y se aplica a todo tipo de datos una vez que se ha establecido una adecuada función de similaridad.

Este algoritmo intenta evitar las debilidades de CURE y ROCK. CURE (al igual que otros métodos relacionados) ignora la información acerca de cómo están interconectados los elementos de dos grupos que se van a unir. ROCK, por su parte, se fija en la interconectividad de los elementos de los dos grupos pero no hace uso de la distancia entre grupos como consecuencia del uso del método de enlace ponderado y otros esquemas relacionados.

CHAMELEON utiliza primero un algoritmo divisivo basado en grafos para agrupar los datos en un número relativamente grande de pequeños grupos. Después utiliza un algoritmo jerárquico aglomerativo para obtener los grupos finales. Para determinar la pareja de grupos más cercanos se usa tanto la interconectividad como la cercanía de los grupos, así como las características de cohesión internas de los grupos.

Para más información acerca de estos algoritmos, se puede consultar también [2] y [13]

1.6.2 Resultados recientes sobre métodos particionales prototípicos: los métodos de k-medoides

Como ya hemos comentado, el método de las k-medias es muy sensible a la presencia de items extraños, dado que un item con atributos de valor muy grande puede distorsionar la distribución de los datos y el valor del centroide correspondiente. Por estas razones se ha propuesto una alternativa al uso del centroide en la que este se sustituye como prototipo de cada grupo por un punto del mismo considerado representativo. Este punto que se denomina *medoide* es el que está colocado de forma más central en el grupo y todo el proceso iterativo del método de las k-medias se desarrolla ahora minimizando las sumas de las distancias de cada punto a los medoides considerados. Existen varios algoritmos basados en esta idea, algunos de los cuales pasamos a comentar.

PAM [17] Es un de los algoritmos más antiguos y resulta bastante sencillo.

- Se parte de una selección inicial de k medoides. Como en el caso de las k-medias, estos generan una partición en k grupos del conjunto total de items.
- Para cada grupo obtenido se intenta reemplazar el medoide asociado a él por algún punto del mismo grupo que sea más idóneo. Esto se hace considerando cada punto del grupo como candidato y calculando la distancia total del resto los elementos del grupo a dicho punto. Si hay mejora con respecto a la del mediode actual, este es sustituido por el punto en cuestión, en caso contrario se mantiene.
- Se recalculan los grupos en base a los nuevos medoides seleccionado y se vuelve con ellos al paso anterior .
- El proceso termina cuando no se cambian los medoides en una iteración.

Este método es más robusto que el de las k-medias en presencia de puntos extraños, pero el proceso es computacionalmente mucho más costoso que el de las k-medias. De hecho, si

consideramos un conjunto de n items el costo computacional de cada iteración es de $O(k(n-k))^2$. En cualquier caso, sigue siendo muy dependiente del número de grupos y de la selecćon inicial de los prototipos.

CLARA [17] PAM trabaja bien para conjuntos de datos pequeños; pero no se adpata bien a grandes bases de datos. Para esta situación, la alternativa propuesta por sus diseñadores es CLARA, basado en un proceso de muestreo. La idea básica es que, si se toma una muestra suficientemente aleatoria y representativa, los medoides obtenidos con ella serán los mismos que los conjunto total. Con esta idea, CLARA realiza sucesivos muestreos y aplica PAM a cada uno de ellos. Los conjuntos de medoides obtenidos se analizan sobre el conjunto total, seleccionando que el nos dé menor distancia global. Con estos de partida se genera un nuevo proceso de muestreo y una nueva iteración. La complejidad de cada iteración es ahora de $O(kS^2 + k(n-k)$, donde S nota el tamaño de la muestra.

La efectividad de CLARA depende del tamaño del la muestra, si bien es fácilmente escalable y puede trabajar con grandes bases de datos.No obstante, hay que destacar que CLARA no obtiene un conjunto óptimo de medoides ya que estos se extraen de datos muestreados y que por tanto puede ocurrir que un punto óptimo no se considere nunca. En este sentido es muy importante la aleatoridad del proceso de muestreo.

CLARANS [19] Para mejorar la calidad y escalabilidad de CLARA se ha propuesto un nuevo algoritmo desarrollado en el contexto de las bases de datos espaciales. CLARANS es también de tipo k-medoide y combina técnicas de muestreo con PAM. La diferencia con CLARA radica en que el proceso de muestreo no se hace para cada conjunto de medoides sino que se realiza cada vez que se calcula uno de ellos. La idea se considerar la búsqueda de los medoides óptimos como un proceso de búsqueda en un árbol donde cada nodo es un conjunto de k medoides. Para cada nodo concreto con su conjunto de prototipos y su agrupamiento asociado, el agrupamiento obtenido reemplazando un medoide del mismo por algún otro punto se denomina entorno. El proceso prueba una serie de entornos generando puntos aleatoriamente, si encuentra un entorno mejor que el agrupamiento considerado, el algoritmo se mueve a este nodo y comienza de nuevo a probar, en caso contrario se considera que se ha llegado a un óptimo local. Cuando se llega a un óptimo local, el algoritmo comienza con un nuevo conjunto de nodos obtenidos por medio de un muestreo aleatorio y una aplicación de PAM. El algoritmo termina cuando se han alcanzado un número suficiente de mínimos locales (datos experimentales recomiendan tomar dicho número igual a 2). La complejidad de CLARANS es de $O(n^2)$. En [10]se presenta una mejora de este algoritmo obtenida mediante el uso de R^*-árboles.

1.7 Técnicas de agrupamiento difuso

En todos los métodos de agrupamiento presentados hasta ahora se ha supuesto, al menos implicitamente, la hipótesis de que el agrupamiento es exclusivo, en el sentido de que un patrón se asigna a un grupo y sólo a un grupo. En otras palabras, los patrones se particionan en conjuntos disjuntos. Naturalmente, si los grupos son compactos y están bien separados esta es la mejor opción, pero el problema aparece cuando los grupos tienen puntos comunes e incluso se solapan, en cuyo caso las fronteras de cada grupo no están definidas y existen puntos que pueden pertenecer a un grupo o a otro y tenemos entonces conjuntos cuyas fronteras están mal definidas o "borrosas". La teoría de subconjuntos difusos (fuzzy sets) desarrollada en la década de los 60 permite que un patrón pertenezca a un grupo con un cierto "grado de pertenencia"con valores entre 0 y 1 (es decir en el

intervalo [0,1]). Para grupos ordinarios (grupos "crisp") el grado de pertenencia para un punto solo puede ser 1 o 0 según el punto esté o no en el grupo y es obvio que se trata de un caso particular de grupo difuso

En estas condiciones cada grupo difuso $C_j;\ j \in \{1,..,K\}$ tiene asociada una "función de pertenencia (que extiende la idea de función indicador de un conjunto ordinario):

$$C_j : X \longrightarrow [0,1],$$

siendo $X = \{x_1, ...x_N\}$ el espacio de patrones. El valor$u_{ij} = C_j(x_i)$ representa el grado de pertenencia del punto x_i al grupo C_j y será tanto mayor cuanto mayor confianza exista en que x_i pertenece a C_j, con$u_{ij} = 1$ si estamos totalmente seguros de dicha pertenencia. Los valores u_{ij} constituyen la "matriz de pertenencia" que notaremos U.

Aunque es bastante habitual imponer la condición (de partición difusa):

$$\sum_{j=1}^{K} u_{ij} = 1, \forall i \in \{1,..N\},$$

hay que hacer notar que el grado de pertenencia no tiene el mismo sentido que una probabilidad. Bajo una hiótesis probabilistica un punto pertenece solamente a un grupo que se determina por medio de un experimento aleatorio, mientras que con una hipótesis de lógica difusa un punto puede pertenecer a dos grupos a la vez, el grado de pertenencia se puede interpretar como el grado de compatibilidad del punto x_i con el grupo C_j, entendido este como el resultado de una propiedad (o un conjunto de propiedades) expresadas de forma imprecisa. Este enfoque es muy útil cuando se intenta una interpretación de los grupos, ya que, en muchos casos, las descripciones de los grupos obtenidos en un problema concreto serán de tipo impreciso por serlo las etiquetas que los caracterizan. Por ejemplo si se intenta agrupar un conjunto de coches itentado obtener los de gama "alta", "media" o "utilitario" o si se intenta agrupar un conjunto de parcelas atendiendo a las prácticas de cultivo que realizan sobre ellas.

Los problemas de agrupamiento han sido ampliamente tratados por medio de conjuntos difusos desde 1966. Existen buenos libros sobre el tema, de entre los que destacaremos [3] por ser el más recientedonde y porque además contiene una amplia bibliografía.

1.7.1 Agrupamientos difusos particionales

Casi todos los algoritmos de agrupamiento basados en conjuntos difusos hacen uso del concepto de partición difusa, lo que permite considerarlos como particionales con fronteras mal definidas entre grupos. No obstante y como veremos posteriormente esta teoría está estrechamente relacionada con el agrupamiento jerárquico. Existen distintos enfoques para diseñar algoritmos particionales de tipo difuso, la mayor parted de los cuales son generalizaciones más o menos directas directas del método de las k-medias. Una forma genérica de describir estos algoritmos se describe a continuación.

Generalización difusa de un algoritmo particional

1. Seleccionar una partición difusa inicial de N objetos en K grupos seleccionando una matriz de pertenencia U.

2. Calcular los "centros" de los grupos difusos asociados a U mediante la expresión:

$$c_j = \sum_{i=1}^{N} u_{ij} x_i$$

3. Utilizando U calcular el valor óptimo de una función objetivo que habitualmente será una ponderación de alguna forma de error cuadrático y que tiene la forma:

$$E^2(U) = \sum_{i=1}^{N}\sum_{j=1}^{K} u_{ij}||x_i - c_k||^2 \tag{1.4}$$

En este punto debemos resolver un problema de optimización sobre los valores de pertenencia u_{ij}, las cuales estarán sujetas a restricciones departición difusa,

$$\sum_{j=1}^{K} u_{ij} = 1;\ u_{ij} \geq 0 \tag{1.5}$$

o bien a condiciones algo más suaves como:

$$max_{j\in\{1..K\}} u_{ij} = 1;\ u_{ij} \geq 0 \tag{1.6}$$

4. Repetir desde el paso 2 hasta que los valores de U no cambien significativamente.

Las diferencias entre unas y otras variantes de este proceso general se encuentran, de una parte, en distintas variantes de la función objetivo 1.4 y de otra en distintas formas de calcular el óptimo de la misma.

El algoritmo particional difuso más popular es el el método conocido como "fuzzy c-means", que es una adaptación directa del de k-medias siguiendo el esquema anterior(ver[3]). Algunas variantes del mismo se pueden encontrar en [7]. A título de ejemplo citaremos las versiones que se obtienen considerando :

1. Que el centro de un grupo difuso no es su media sino su valor más representativo es decir:

$$\forall j \in \{1..K\}\ ;\ c_j = x_{lj} \mid u_{lj} = max_{i\in\{1,..n\}} u_{ij}$$

2. Otras funciones de distancia diferentes la de la expresión 1.4, concretamente, dada una partición difusa definida por la matriz U, se define la distancia entre dos puntos asociada al grupo C_j como:

$$\forall x, y\ d_j(x,y) = min(u_j(x), u_j(y))||y - x||^2$$

En estas condiciones la distancia expresada en 1.4 e transforma en:

$$S^2(U) = \sum_{i=1}^{N}\sum_{j=1}^{K} d_j(x_i, c_j) \tag{1.7}$$

y la forma de dicha expresión depende de cómo se tome el valor de $u_j(c_j)$, en el caso de que consideremos $u_j(c_j) = 1$ obtenenmos la expresión 1.4

Otras variantes del algoritmo utilizan una función de distancia adaptativa que depende de cada iteración ya que utiliza "diámetro"de cada clase en la expresión de la distancia , en [7] se pueden encontrar más detalles. Otra referencia importante dentro de este punto son [3], ambos libros contienen a su vez conjuntos de referencias muy amplios acerca de este tema.

Existen otros métodos de agrupamiento difuso que pueden aplicarse cuando lo que se conoce no es la relación entre items sino que se tiene una cierta medida de la conexión entre items y clases.

Esta situación es similar la que se tiene cuando, en el caso de agrupamiento particional no difuso se posee información acerca de la distribución de probabilidad de cada grupo. Es decir, de la probabilidad de que un determinado patrón pertenezca a una clase. Como comentamos en el apartado que ha dado origen a los métodos de agrupamiento "crisp" de tipo estadístico (Análisis Discriminante, Máxima Verosimilitud etc..)

En el caso de los método de agrupamiento basados en conjuntos difusos la generalización se hace en dos sentidos:

A Considerando que los grupos son difusos,

B Considerando que la asociación entre items y clases no está dada por una distribución de probabilidad conjunta sino por medio de un concepto más general que es el de una .ªsignación Básica de Probabilidad", es decir una "Evidencia" en sentido de la Teoría de Dempster- Shaffer.

El modelo general contiene los siguientes elementos:

1. $X = \{x_1..x_n\}$ conjunto finito de items a agrupar

2. $C = \{C_1..C_K\}$ conjunto de grupos o clases de of X, supondremos que $\{C_i\}$ son subconjuntos difusos de X

3. Una "asignaciónbásica de probabilidad" (BAP) m sobre $\mathcal{P}(\mathcal{X} \times \mathcal{C})$, representando la información disponible acerca de la relación item/grupo. m nos mide la probabilidad de que ciertos conjuntos items y de clases aparezcan conjuntamente.

4. Una función de costos de "mal agrupamiento" que nos mide el costo de considerar que un item pertenece a una clase cuando pertenece a otra.

$$r : C \times C \to R^+ \cup \{0\}$$

r puede representarse como una matriz $K \times K$ $\left[r_{i_j}\right]$

En estas condiciones lo que se busca es una familia de funciones de pertenencia $u = \{u_1, ..., u_n\}$ para las clases de manera que se haga lo menor posible alguna función de pérdida, por ejemplo:

$$R_i(u) = \sum_{A \subset X} m(A/i) \sum_l r_{li} \sup_{x \in A} u_l(x)$$

$$T_i(u) = \sum_{A \subset X} m(A/i) \sum_k l r_{li} \inf_{x \in A} u_l(x)$$

donde $m(A/i)$ son medidas condicionadas. También se pueden limitar los posibles valores de las funciones de pertenencia introduciendo restricciones del tipo 1.5 end 1.6.

Los resultados correspondientes a estos modelos se pueden encontrar en [6], algunos casos particulares se proponen en [21] y [5]

1.7.2 Agrupamientos jerarquicos y conjuntos difusos

La Teoría de Conjuntos Difusos ha resultado ser una herramienta muy adecuada para ser usada en problemas de clasificación y particularmente para generar potentes métodos de agrupamiento. Ahora bien, en el caso jerárquico la relación va más alla de la mera consideración de herramienta relación como consecuencia de la identidad entre un agrupamiento jerárquico con el concepto de relación de similitud difusa.

De la misma forma que el concepto de conjunto difuso extiende el de conjunto clásico, el concepto de relación difusa extiende el de relación clásica, concretamente el de relación de similitud extiende el de relación de equivalencia. Las definiciónes formales son:

Dados dos referenciales X, e Y se define *Relación Difusa* sobre $X \times Y$, R como un subconjunto difuso tal que:

$$\mu_R : X \times Y \longrightarrow [0,1]$$

Semáticamente una relación difusa refleja conexiones imprecisas (graduales) entre elementos de dos conjuntos.

Si $X \equiv Y$ una relación difusa R se dice de *similitud* si verifica las propiedades:

1. *Reflexiva:* $\forall x \in X, \mu_R(x,x) = 1$
2. *Simétrica:* $\forall x, y \in X, \mu_R(x,y) = \mu_r(y,x)$
3. *Max-min transitiva* $\forall x, y \in X, \mu_R(x,y) \geq max_{z \in X}(\mu_R(x,z) \bigwedge \mu_R(z,y))$

La importancia de las relaciones de similitud difusa radica en la siguientes propiedades:

1. Sea $\alpha \in [0,1]$ y sea la relación no difusa $R_\alpha = (x,y) | \mu_R(x,y) \geq alpha$, si $\mu_R(.,.)$ es de similitud R_α es de equivalencia, es decir induce una partición en el conjunto x
2. Sean α y β valores entre 0 y 1 tales que $alpha \leq beta$, sean $\mathcal{P}_\alpha$ y $\mathcal{P}_\beta$ las particiones inducidas en X por R_α y R_β respectivamente, se verifica que:

   $$\forall B \in \mathcal{P}_\beta \; \exists A \in \mathcal{P}_\alpha \text{ tal que} B \subseteq A$$

   es decir si $alpha \leq beta$, entonces $\mathcal{P}_\alpha$ y $\mathcal{P}_\beta$ son "particiones encajadas".

De estas dos propiedades se deduce fácilmente los conceptos de agrupamiento jerárquico y relación de similitud difusa son equivalentes de forma que de todo agrupamiento jerárquico se puede obtener una relación de similitud difusa y de toda relación de similitud difusa se puede obtener un agrupamiento jerárquico. Esta identificación de conceptos permite utilizar las medidas y propiedades asociadas a conjuntos difusos para dar criterios de búsqueda de particiones en agrupamientos jerárquicos. En [4], [7] o [3] puede encontrarse discusiones detalladas sobre el tema, así como acerca de la relación de los agrupamientos jerárquicos difusos, con jerarquías de sistemas de control.

Bibliografía

[1] M. Akerst, M. Breuning, H.P. Kriegel, and J. Sanders. Optics: Ordering points to identify clustering structures. In *Proc. of ACM-SIGMOD Int. Conf.*, pages 49–60, Philadelphia USA 1999, 1999.

[2] P. Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[3] J. Bezdeck, J. Keller, R. Krisnapuram, and N. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing.* Springer Verlag, 2005.

[4] M. Delgado, A.F. Gomez-Skarmeta, and M.A. Vila. On the use of hieralchical clustering in fuzzy modellling. *International Journal on Approximate Reasoning*, 14:237–257, 1996.

[5] M. Delgado, A.F. Gomez-Skarmeta, and M.A. Vila. On the use of probability and possibility measures in fuzzy clustering. In *Proc. of Fuzz-IEEE 97 July , 1997, Barcelona, Spain*, pages 143–148. IEEE Pub., 1997.

[6] M. Delgado, A.F. Gomez-Skarmeta, and M.A. Vila. Pattern recognition with evidential knowledge. *International Journal of Intelligent Systems*, 14(2):145–164, 1999.

[7] D. Dumitriescu, B. Lazzerini, and L.C. Jain. *Fuzzy sets and their applications to clustering and training.* CRC Press, Boca Raton, London, New York, Washington D.C., 2000.

[8] B.S. Duran and P.L. Odell. *Cluster Analysis: a survey.* Springer Verlag, Berlin, Heidelberg, New York, 1974.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X.Xu. A density.based algorithm for discovering clusters in large spatial databases. In *Proc. 2nd. Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, pages 67–82, Oregon Usa August 1966, 1996.

[10] M. Ester, H.-P. Kriegel, and X.Xu. A database interface for clustering in large spatial databases. In *Proc. 1st. ACM SIGKDD*, pages 94–99, Montreal Canada, 1994.

[11] S. Guha, R. Rastogi, and K. Shim. CURE: an efficient clustering algorithm for large databases. In *1998 ACM-SIGMOD Int. Conf. Management of Data proceedings*, pages 73–84, Sydney, Autralia March 1996, 1998.

[12] S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345–366, 2000.

[13] J. Hang and M. Kamber. *Data Mining: Concepts and Techniques.* Morgan Kauffman, 2000.

[14] A. K. Jain, M.N. Murty, and P.J. Flynn. Data clustering : A review. *ACM Computing Survey*, 31:264–323, 1999.

[15] K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, New Jersey, 1988.

[16] G. Karypis, E.-H.Han, and V. Kumar. CHAMELEON: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.

[17] L. Kauffman and P.J. Rousseeuw. *Finding Group in Data: an Introduction to Cluster Analysis*. John Wiley and Sons, New York, 1990.

[18] G. N. Lance and W.T. Williams. A general theory of classificatory sorting strategies: Ii. clustering systems. *Computer Journal*, 10:271–277, 1967.

[19] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In Jorgeesh Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *20th International Conference on Very Large Data Bases, September 12–15, 1994, Santiago, Chile proceedings*, pages 144–155, Los Altos, CA 94022, USA, 1994. Morgan Kaufmann Publishers.

[20] N. T. Pang, Steinbach M., and V. Kumar. *Introduction to Data Mining*. Addison Wesley, New Jersey, 2006.

[21] M.A. Vila and M. Delgado. Problems of classification in a fuzzy environment. *Fuzzy Sets and Systems*, 9(83):229–239, 1983.

[22] D. R. Wilson and T. R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.

[23] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *1996 ACM-SIGMOD Int. Conf. Management of Data proceedings*, pages 103–114, Montreal Canada June 1996, 1996.