

Implementación del método de grupos estables en el paquete FactoClass de R

CARLOS ANDRÉS ARIAS^a, DIANA CAROLINA ZÁRATE-DÍAZ^b, CAMPO ELÍAS PARDO^c

DEPARTAMENTO DE ESTADÍSTICA, FACULTAD DE CIENCIAS, UNIVERSIDAD NACIONAL DE COLOMBIA, BOGOTÁ, COLOMBIA

Resumen

Se presenta la función *stableclus*, programada en R e incluida en el paquete *FactoClass*, para ejecutar el método de formas fuertes o grupos estables a partir de las coordenadas factoriales de un análisis previo. Se resume el método y se hacen dos ejercicios utilizando la función construida, que permiten ver sus bondades y deficiencias. Con los ejemplos se corrobora que el método es inestable para decidir el número de clases. Una vez establecido el número de clases, en los dos ejemplos, se obtienen resultados más cercanos, a los de la clasificación mixta, cuando se hace un filtrado previo utilizando menos ejes factoriales para la clasificación con *K*-medias.

Palabras clave: software estadístico, métodos de clasificación, *K*-medias, análisis multivariado.

1. Introducción

El método *K*-medias (Hartigan & Wong 1979) es de gran utilidad para los métodos de clasificación, ya que, permite hacer particiones de grandes conjuntos de datos. Sin embargo, este algoritmo tiene dos inconvenientes: se debe dar el número de clases con sus puntos iniciales y converge a óptimos locales que dependen de los puntos iniciales. En la figura 1 se presenta un resumen del algoritmo de agregación alrededor de centros móviles (sinónimo de *K*-medias en este documento), como se encuentra en (Lebart et al. 2006, p.254).

Para solucionar, al menos parcialmente, los inconvenientes de los métodos de agregación alrededor de centros móviles, se puede utilizar el método de formas fuertes o grupos estables (Lebart et al. 2006, p.254). Los grupos estables se identifican realizando varias particiones, cada una con diferentes puntos iniciales y seleccionando aquellos grupos en los cuales los individuos se mantienen asignados a la misma clase. En la sección 2 se presenta una descripción del método.

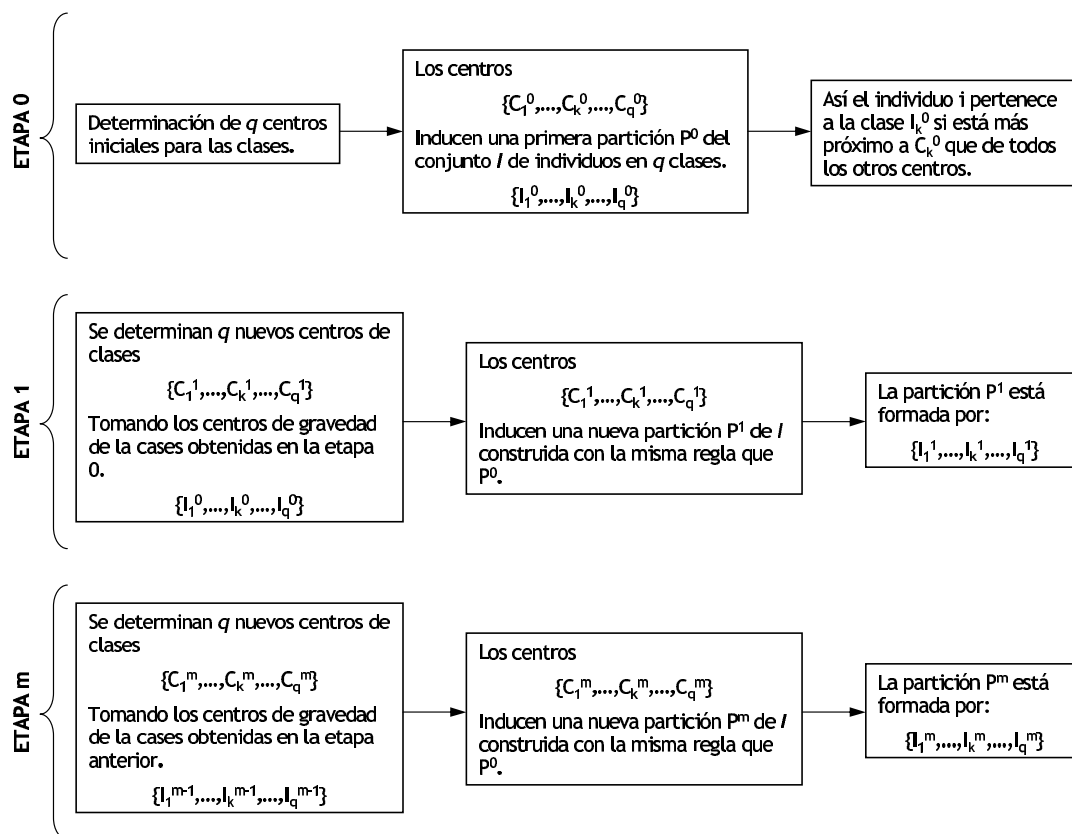
El método de grupos estables forma parte del procedimiento de clasificación mixta, descrito en Lebart et al. (2006, pp.287-294) y disponible en los programas SPAD (Lebart et al. 1999) y DtmVic (Lebart 2008). En el desarrollo de este trabajo se programa en R (R Development Core Team 2009) la función *stableclus* para ejecutar el procedimiento de grupos estables y se incluye dentro del paquete FactoClass (Pardo & DelCampo 2007).

En la sección 4 se compara el método con la clasificación mixta en dos ejemplos académicos, incluidos en el paquete *FactoClass* y finalmente se dan algunas conclusiones en la sección 5. En el apéndice se muestran los comandos de R, utilizados para la obtención de resultados gráficos y numéricos de los ejemplos.

^aEstudiante de Estadística. E-mail: caariasr@unal.edu.co

^bEstudiante de Estadística. E-mail: dczarated@unal.edu.co

^cProfesor Asociado. E-mail: cepardot@unal.edu.co

FIGURA 1: Esquema del algoritmo *K-medias*.

2. Formas fuertes

El método de formas fuertes, sugerido inicialmente por Diday (1972) (citado por Lebart et al. (2006)), surge como una solución para la debilidad que hay en el algoritmo de las *K-medias*. Sin embargo, este método únicamente da una solución parcial al problema.

La técnica de formas fuertes consiste, básicamente, en llevar a cabo varias particiones del conjunto inicial, utilizando cada vez diferentes centros y manteniendo como grupos estables los conjuntos de elementos (o individuos) que en cada una de las particiones realizadas fueron siempre asignados a la misma clase.

Siguiendo la idea que se expone en Lebart et al. (2006), los pasos para encontrar los grupos estables son:

1. Determinación de s particiones cada una a partir de un conjunto $\{C_{i1}, \dots, C_{iq}\}$ con $i = 1, \dots, s$.
2. Para el conjunto de las s particiones $\{P_1, P_2, \dots, P_s\}$ cada una en q clases, en la partición producto, la clase indexada por $\{k_1, k_2, \dots, k_s\}$ contiene los individuos pertenecientes a la clase k_1 de la partición P_1 , luego a la clase k_2 de P_2 hasta llegar a la clase k_s de P_s .
3. Los grupos estables se forman con las clases que contengan más de un individuo de la partición producto.

Para ilustrar el procedimiento, se usa el ejemplo presentado en Lebart et al. (2006). La idea es obtener 6 grupos homogéneos de un conjunto de 1000 individuos. Inicialmente se hace una partición del conjunto en 6 clases alrededor de centros móviles, posteriormente se repite el procedimiento dos veces, obteniendo para cada una de las particiones los resultados presentados en la tabla 1.

TABLA 1: Partición de 1000 individuos en 6 grupos homogéneos

Partición	Clase						
	1	2	3	4	5	6	
1	127	188	229	245	151	60	1000
2	232	182	213	149	114	110	1000
3	44	198	325	99	130	204	1000

De acuerdo con los pasos mencionados anteriormente, ahora se debe encontrar la partición producto. Como se tienen 3 particiones cada una con 6 clases, la partición producto tiene en total $3^6 = 216$ clases. Los individuos en cada una de las 216 clases son los que en cada una de las particiones quedaron siempre agrupados juntos. En la tabla 2 se presentan los 50 grupos no vacíos del total de 216 grupos.

TABLA 2: Grupos estables

Grupos 1 - 10	168	118	114	107	88	83	78	26	22	16
Grupos 11 - 20	15	14	12	12	12	11	10	7	7	7
Grupos 21 - 30	6	6	4	4	4	4	3	3	3	3
Grupos 31 - 40	3	3	3	2	2	2	2	2	2	2
Grupos 41 - 50	1	1	1	1	1	1	1	1	1	1

En general el número de clases obtenidas en la partición producto es muy grande y de hecho, muchas de ellas presentan frecuencias bajas. Por lo tanto, una sugerencia es tomar las clases que tengan frecuencias significativas. El número de estas agrupaciones se pueden identificar con un salto en un gráfico de barras de las frecuencias de los grupos como está en la figura 2. En este caso los primeros 7 agrupamientos tienen frecuencias significativas Lebart et al. (2006).

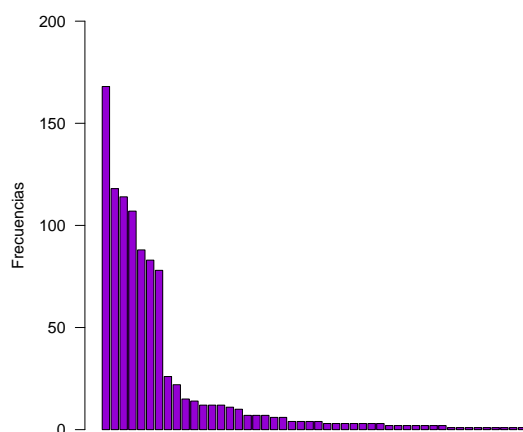


FIGURA 2: Histograma de las frecuencias de los grupos estables del ejemplo

3. La función *stableclus*

La función se programa para ejecutar el procedimiento descrito en la sección anterior, utilizando la función `kmeansW` de *FactoClass*, que provee los puntos iniciales, cuando solo se dá el número de clases. La

función *stableclus* recibe un objeto *dudi*, del paquete *ade4* (Dray & Dufour 2007), de donde se toman las coordenadas factoriales y los pesos de las filas (`dudi$li` y `dudi$lw`). Los otros parámetros básicos son el número de particiones (`part`) y el número de clases en cada partición (`k.clust`). Se incluye el parámetro `clus` para indicar el número de clases de la partición a construir, el valor por defecto es `NULL` en cuyo caso se le pide al usuario que lo suministre durante la ejecución. Finalmente se incluyen dos parámetros lógicos: `bplot` para producir o no el histograma de las formas fuertes y `kmns` para realizar, si se desea, una consolidación final con *K*-medias. El comando para llamar la función es:

```
stableclus(dudi,part=2,k.clust=2,ff.clus=NULL,bplot=TRUE,kmns=FALSE)
```

4. Ejemplos

Se utiliza la función *stableclus* para analizar un poco el procedimiento de formas fuertes a través de dos ejemplos incluidos en *FactoClass* con los nombres *BreedsDogs* y *ColorAdjective*.

4.1. Clasificación de razas de perros

Se tienen 27 razas de perros, que se caracterizan por 6 variables ordinales: tamaño, peso, velocidad, inteligencia, afectividad y agresividad, ejemplo utilizado por Fine (1996), para ilustrar el uso combinado del análisis de correspondencias múltiples y los métodos de clasificación. El objetivo del ejercicio es clasificar a las razas de perros según las 6 características físicas y psíquicas.

Se realiza previamente un análisis de correspondencias múltiples (ACM) y se utilizan los 10 ejes factoriales para llevar a cabo una clasificación jerárquica con el método de Ward. Se corta el árbol para obtener una partición en 4 clases, que se muestra en la figura 3, sobre el primer plano factorial del ACM. Esta partición se usa como referencia para comparar con la obtenidas con el método de formas fuertes utilizando la función *stableclus*.

Luego de muchos ensayos se decide presentar en este documento el resultado de realizar 9 particiones, usando sendas repeticiones del comando `stableclus(acm,3,3,4,TRUE,TRUE)` (3 particiones, en 3 clases y partición final en 4 clases, mostrando el histograma de las formas fuertes y realizando un *k*-medias final de consolidación), utilizando 3 ejes factoriales del ACM previo (ver el código en el apéndice). La decisión de tomar 3 ejes se sustenta en la forma del diagrama de valores propios (`barplot(acm$eig)`) y del criterio de Benzécri (Lebart et al. 2006, p. 218) que argumenta que solo los ejes con inercia superior al inverso del número de variables son “factores directos”. Se observa efectivamente, que se pueden identificar grupos estables en cada uno de los 9 histogramas (figura 4), pero no se logra siempre la identificación de las 4 clases que se obtienen con la clasificación jerárquica. La comparación con la partición de referencia se hace mediante tablas de contingencia cruzando cada partición obtenida por formas fuertes con la lograda utilizando clasificación jerárquica. Cada cruce corresponde a una línea de la tabla 3, en ésta se observa que de las nueve particiones 4 coinciden con la obtenida con la clasificación jerárquica y que la clase 1 es la misma en las 9 particiones. Este es apenas un resultado, pero el lector puede comprobar, por ejemplo, el número de particiones coincidentes por cada 100 ejecutando el código que se da en el apéndice. En tres ejecuciones del código obtuvimos 21, 31 y 24 coincidencias de 100, respectivamente.

4.2. Ejemplo adjetivos y colores

Se realiza el mismo ejercicio de comparar 9 particiones obtenidas por formas fuertes con la partición, denominada de referencia, lograda mediante la combinación de análisis de correspondencias simples (ACS) y clasificación mixta. El objetivo en el ejemplo colores es resolver la pregunta ¿Qué adjetivos se asocian con cada color?, a partir de una tabla de contingencia que cruza 89 adjetivos por 11 colores. La pregunta se responde muy bien con la partición en 7 clases obtenida con la función *FactoClass*, usando los 10 ejes del ACS previo. En la figura 5 se muestra la proyección de las 7 clases sobre el primer plano factorial del ACS.

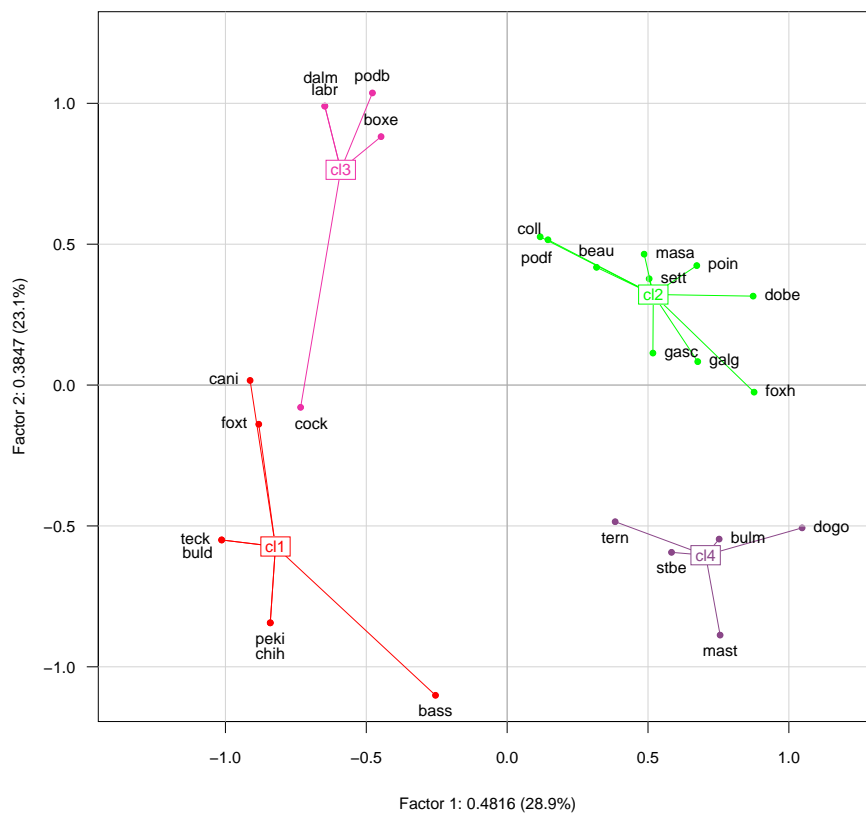


FIGURA 3: Clases de referencia sobre el primer plano factorial del ACM, del ejemplo razas de perros

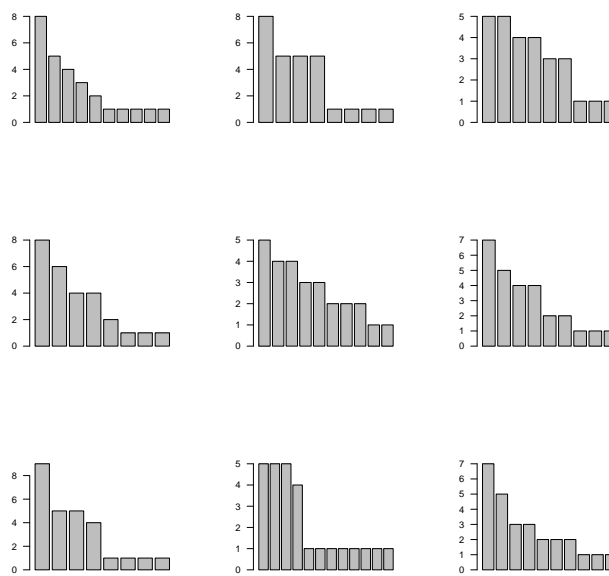


FIGURA 4: Diagramas de barras mostrando las formas fuertes de 9 ejecuciones de la función *stableclus* en el ejemplo razas de perros

TABLA 3: Cruce de nueve particiones con la partición de referencia del ejemplo *razas de perros*

Partición	Clase 1 7				Clase 2 10				Clase 3 5				Clase 4 5			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	P1	0	0	7	0	0	0	0	10	0	5	0	0	5	0	0
P2	0	0	0	7	10	0	0	0	0	0	5	0	0	5	0	0
P3	0	0	0	7	0	7	3	0	0	0	4	1	5	0	0	0
P4	0	0	0	7	9	0	1	0	0	0	5	0	0	5	0	0
P5	0	0	0	7	0	8	2	0	0	0	4	1	5	0	0	0
P6	0	0	0	7	0	5	5	0	4	0	0	1	0	2	3	0
P7	0	7	0	0	0	0	0	10	0	0	5	0	5	0	0	0
P8	0	0	7	0	0	10	0	0	5	0	0	0	0	0	0	5
P9	0	0	0	7	1	1	8	0	5	0	0	0	0	5	0	0

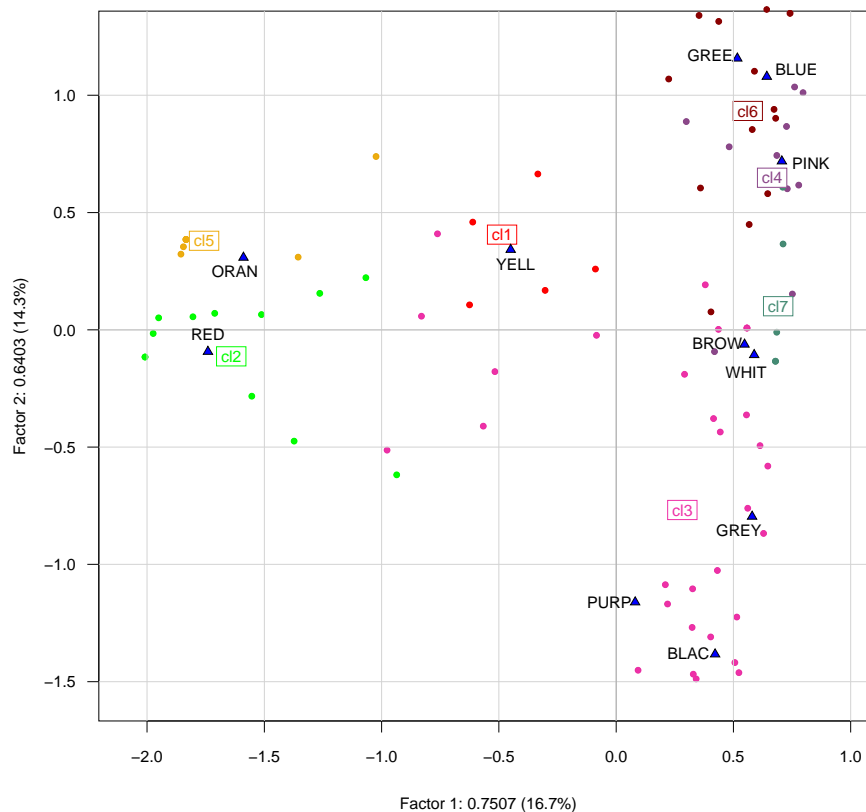


FIGURA 5: Partición de referencia para ejemplo adjetivos y colores sobre el primer plano factorial del ACS

Para la obtención de las formas fuertes se decide tomar 6 ejes del ACS previo, en este ejercicio 5 ejes tienen valores propios superiores al promedio, pero el eje 6 separa los adjetivos que se asocian al color naranja de los que se asocian al rojo. Se presenta el resultado de repetir 9 veces la función `stableclus(acs, 3, 4, 7, TRUE, TRUE)` (acs previo reteniendo 6 ejes, 3 particiones de 4 clases y partición final en 7 clases, se muestra el histograma de formas fuertes y se hace consolidación final con K -medias). La figura 6 muestra los diagramas de barras para las formas fuertes de las nueve repeticiones, mostrando de nuevo la inestabilidad para decidir el número de clases. La tabla 4 muestra las 9 tablas de contingencia, cada una en una columna, al cruzar las particiones obtenidas con formas fuertes con la partición de referencia. En este ejemplo no se encuentra con formas fuertes, en las 9 repeticiones, una partición igual a la obtenida con clasificación mixta, pero se encuentran siempre las clases 1, 4 y 5, aunque es un

resultado que puede ser distinto cada vez que se ejecute el procedimiento. Se realizaron 300 repeticiones del procedimiento de formas fuertes y en ninguna se encontró coincidencia con la partición de referencia, sin embargo son particiones parecidas.

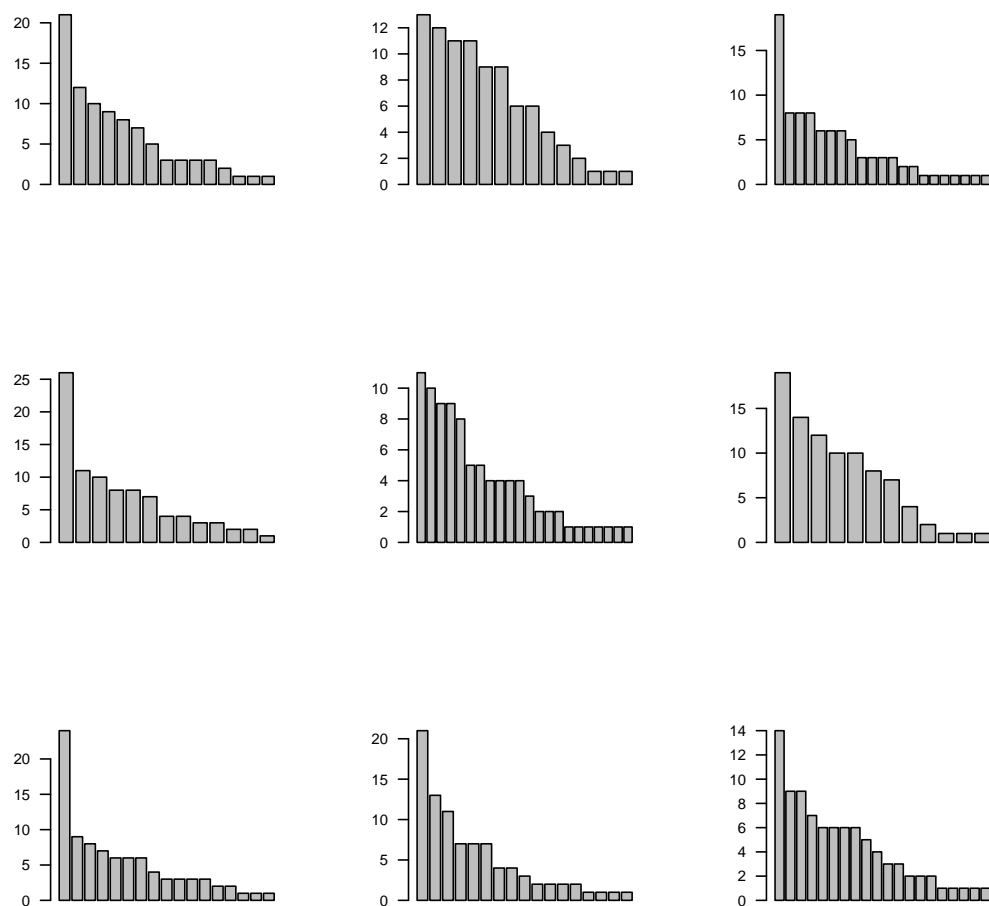


FIGURA 6: Diagramas de barras mostrando las formas fuertes de 9 ejecuciones de la función *stableclus* en el ejemplo adjetivos y colores

TABLA 4: Cruce de 9 particiones de formas fuertes lo la de referencia para el ejemplo adjetivos y colores

Cl H	Frec	Cl FF	P1	P2	P3	P4	P5	P6	P7	P8	P9
1	7	1	0	0	0	0	0	0	0	0	0
		2	0	0	0	0	0	0	0	7	0
		3	0	0	7	0	0	7	0	0	7
		4	0	0	0	0	0	0	0	0	0
		5	0	0	0	7	0	0	0	0	0
		6	0	0	0	0	0	0	0	0	0
		7	7	7	0	0	7	0	7	0	7
2	12	1	11	0	0	0	0	0	0	12	0
		2	0	0	11	0	0	0	12	0	0
		3	0	0	1	12	0	0	0	0	0
		4	0	12	0	0	0	0	0	0	0
		5	0	0	0	0	12	0	0	0	12
		6	0	0	0	0	0	0	0	0	0
		7	1	0	0	0	0	12	0	0	0
3	32	1	0	8	25	0	10	0	0	3	0
		2	15	0	0	16	0	1	0	3	0
		3	11	0	0	0	4	1	6	0	12
		4	2	0	0	6	0	8	14	17	7
		5	0	4	0	10	0	22	0	9	0
		6	4	13	0	0	15	0	0	0	13
		7	0	7	7	0	3	0	12	0	0
4	10	1	0	0	0	0	0	0	10	0	0
		2	0	10	0	0	10	0	0	0	0
		3	0	0	0	0	0	0	0	10	0
		4	0	0	0	0	0	0	0	0	0
		5	10	0	0	0	0	0	0	0	0
		6	0	0	10	10	0	10	0	0	0
		7	0	0	0	0	0	0	0	0	10
5	8	1	0	0	0	8	0	8	0	0	8
		2	0	0	0	0	0	0	0	0	0
		3	0	0	0	0	8	0	0	0	0
		4	0	0	8	0	0	0	0	0	0
		5	0	8	0	0	0	0	0	0	0
		6	8	0	0	0	0	0	8	0	0
		7	0	0	0	0	0	0	0	8	0
6	15	1	11	0	0	0	2	0	0	0	0
		2	0	0	0	0	0	1	0	0	14
		3	3	3	0	0	0	0	0	0	1
		4	1	0	0	8	13	14	0	0	0
		5	0	0	1	1	0	0	6	3	0
		6	0	0	0	0	0	0	0	12	0
		7	0	12	14	6	0	0	9	0	0
7	5	1	0	0	0	0	0	0	0	0	0
		2	0	0	0	0	0	5	0	0	1
		3	0	5	0	0	0	0	0	0	0
		4	5	0	0	0	1	0	0	3	4
		5	0	0	5	0	0	0	5	0	0
		6	0	0	0	0	0	0	0	2	0
		7	0	0	0	5	4	0	0	0	0

5. Conclusiones

Se corrobora la utilidad relativa de la búsqueda de formas fuertes o grupos estables para decidir el número de clases en una partición que se desea obtener, en el sentido que mejora un poco los resultados cuando se utiliza solamente una vez el algoritmo de K -medias.

En los dos ejemplos el procedimiento de formas fuertes, una vez decidido el número de clases, da mejores resultados cuando se hace un filtrado previo tomando menos ejes (3 para el ejemplo de razas de perros y 6 para el de colores y adjetivos).

Este ejercicio permite corroborar la fortaleza de los métodos de clasificación mixta y su uso combinado con los métodos factoriales, lo mismo que la utilidad de realizar un filtrado utilizando los primeros ejes que se puedan considerar informativos (Lebart et al. 2006, cap. 6).

El acercamiento que se logra con las formas fuertes a las particiones de la clasificación mixta justifica su uso para mejorar las clasificaciones previas a los métodos jerárquicos, cuando el volumen de datos no permite utilizar a estos últimos directamente.

La función *stableclus* es útil como herramienta académica para visualizar los problemas de los métodos de agregación alrededor de centros móviles y queda disponible dentro del paquete *FactoClass*.

Referencias

- Dray, S. & Dufour, A. (2007), 'The ade4 package: implementing the duality diagram for ecologists', *Journal of Statistical Software* **22**(4), 1–20.
- Fine, J. (1996), *Iniciación a los análisis de datos multidimensionales a partir de ejemplos*, Notas de clase, Montevideo.
- Hartigan, J. A. & Wong, M. A. (1979), 'A K-means Clustering Algorithm', *Applied Statistics* **28**(100–108).
- Lebart, L. (2008), 'DtmVic: Data and Text Mining - Visualization, Inference, Classification. Exploratory statistical processing of complex data sets comprising both numerical and textual data.', Web.
*<http://egsh.enst.fr/lebart/>
- Lebart, L., Morineau, A., Lambert, T. & Pleuvret, P. (1999), *SPAD. Système Pour L'Analyse des Données*, Paris.
*<http://www.spad.eu>
- Lebart, L., Piron, M. & Morineau, A. (2006), *Statistique exploratoire multidimensionnelle. Visualisation et inférence en fouilles de données*, 4 edn, Dunod, Paris.
- Pardo, C. & DelCampo, P. (2007), 'Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass', *Revista Colombiana de Estadística* **30**(2), 231–245.
*www.estadistica.unal.edu.co/revista
- R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*<http://www.R-project.org>

Apéndice A. Comandos de R para los ejemplos

```

=====
# Simposio de Estadística (2009). Ejemplos de Arias,Zarate & Pardo (2009)
# Comandos de R para el ejemplo de formas fuertes
#-----
# Ejemplo razas de perros
library(FactoClass)      # carga el paquete
data(BreedsDogs)        # carga los datos
BDact <- BreedsDogs[-7]  # elimina la variable 7
# obtención de la partición de referencia y acm para formas fuertes con 3 ejes (nf=3)
FCbd <- FactoClass( BDact, dudi.acm, k.clust = 4,scanFC = FALSE,nf=3, nfcl = 10)
cl <- FCbd$cluster      # vector con la partición de referencia
xfig(encoding="latin1", width = 8, height = 8) # primer plano con clases y exportacion a xfig
plotFactoClass(FCbd,cframe=1,col.row=c("red","green","maroon2","orchid4"),Tcol=FALSE)
dev.off()
# obtención de 9 particiones no se obtienen las mismas porque los puntos iniciales son aleatorios
acm <- FCbd$dudi
par(mfrow=c(3,3)) # las 9 gráficas en una grilla de 3x3.
tabla <- rep(c(table(cl)),each=4)
tabla <- rbind(tabla,rep(1:4,4))
for (i in 1:9){
  ff <- stableclus(acm,3,3,4,TRUE,TRUE)
  fcl <- ff$class
  tabla <- rbind(tabla,t(c(table(fcl,cl))))
}
rownames(tabla) <- paste("P",-1:9,sep="")
colnames(tabla) <- rep(paste("H",1:4,sep=""),each=4)
dev.copy2eps() # bajo Linux grabar actual en archivo Rplot.eps
xtable(tabla,digits=rep(0,17))
# conteo de particiones coincidentes con la de referencia en 100 ejecuciones de stableclus
pv <- NULL
for (i in 1:100){
  sort(table(stableclus(acm,3,3,4,FALSE,TRUE)$class))-> p
  pv <- rbind(pv,p)
}
sum(rowSums(pv==matrix(c(5,5,7,10),100,4,TRUE))==4)
=====
# ejemplo adjetivos y colores: particion de referencia y acs para formas fuertes reteniendo 6 ejes
data(ColorAdjective)
FCcol <- FactoClass(ColorAdjective, dudi.coa,nf=6,nfcl=10,k.clust=7,scanFC = FALSE)
acs <- FCcol$dudi
cl <- FCcol$cluster
xfig(encoding="latin1", width = 8, height = 8)# primer plano con clases y exportacion a xfig
plotFactoClass(FCcol,ucal=100,cframe=0.9,cstar=0,
  col.row=c("red","green","maroon2","orchid4","darkgoldenrod2","dark red","aquamarine4"))
text(FCcol$dudi$co,colnames(ColorAdjective),cex=0.8,pos=3)
dev.off() # cierre device Xfig
# particiones con formas fuertes
par(mfrow=c(3,3)) # las 9 gráficas en una grilla de 3x3.
tabla <- rep(c(table(cl)),each=7)
tabla <- rbind(tabla,rep(1:7,7))
for (i in 1:9){
  ff <- stableclus(acs,3,4,7,TRUE,TRUE)
  fcl <- ff$class
  tabla <- rbind(tabla,t(c(table(fcl,cl))))
}
dev.copy2eps() # bajo Linux grabar actual en archivo
rownames(tabla) <- paste("P",-1:9,sep="")
colnames(tabla) <- paste(rep(paste("H",1:7,sep=""),each=7),1:49,sep="")
xtable(t(tabla),digits=rep(0,12)) # tabla para LaTeX
pv <- NULL # cien repeticiones de formas fuertes
for (i in 1:100){
  sort(table(stableclus(acs,3,4,7,FALSE,TRUE)$class))-> p
  pv <- rbind(pv,p)
}
sum(rowSums(pv==matrix(sort(table(cl)),100,7,TRUE))==7)
=====

```