XXVI Congreso Nacional de Estadística e Investigación Operativa
Úbeda, 6-9 de noviembre del 2001

CONTRIBUCIONES DESDE UNA PERSPECTIVA BASADA EN PROXIMIDADES AL *FUZZY K-MEANS CLUSTERING*

Francesc Oliva[1], Miquel de Cáceres[2], Xavier Font[2], Carles M. Cuadras[1]

[1]Departamento de Estadística
Facultad de Biología
Universidad de Barcelona

[2]Departamento de Biología Vegetal
Facultad de Biología
Universidad de Barcelona

RESUMEN

El *fuzzy K-means* es una generalización del algoritmo *K-means* en el ámbito de la lógica difusa. Si como información de origen se dispone únicamente de proximidades entre objetos, puede obtenerse una partición *fuzzy K-means* a partir de expresiones basadas en las mismas proximidades sin necesidad de disponer de coordenadas de representación de los objetos. Por otra parte, debido a que la participación de un individuo en un *cluster* influye en su posterior reasignación, se propone una modificación del algoritmo basada en una validación cruzada (*cross-validation*) que elimina dicho efecto. Todas las expresiones aportadas admiten una versión *crisp* del *K-means*.

Palabras y frases clave: *fuzzy K-means, partitioning, clustering, dissimilarities, principal coordinate analysis*.

1. INTRODUCCIÓN

K-means (MacQueen, 1967) es una técnica de análisis *cluster* que trata de establecer una partición en K grupos o *clusters* sobre un conjunto de N objetos $\{O_1,\ldots,O_N\}$ de los que disponemos de una información multivariante P-dimensional. Partiendo de la matriz de datos $\mathbf{X}_{N\times P}$, la función que se pretende minimizar en el proceso de clasificación es la suma total de cuadrados de los errores (*TESS*), cuya expresión viene dada por:

$$TESS_K = \sum_{k=1}^{K} E^2_{(k)} = \sum_{k=1}^{K}\sum_{i=1}^{N} I\left[O_i \subset C_k\right] e^2_{i(k)}$$

siendo $E^2_{(k)}$ la suma de cuadrados de los errores (*ESS*) para el *cluster* C_k, $I[O_i \subset C_k]=1$ si el objeto O_i ha sido asignado a C_k, $I[O_i \subset C_k]=0$ si O_i no ha sido asignado a C_k, y $e^2_{i(k)}$ la distancia euclídea al cuadrado de cada objeto al centroide de C_k:

$$e^2_{i(k)} = \sum_{j=1}^{P} (x_{ij} - \bar{x}_{(k)j})^2 = (\mathbf{x}_i - \bar{\mathbf{x}}_{(k)})'(\mathbf{x}_i - \bar{\mathbf{x}}_{(k)})$$

donde $\bar{x}_{(k)j} = \frac{1}{n_k} \sum_{i=1}^{N} I[O_i \subset C_k]\, x_{ij}$, $n_k = \sum_{i=1}^{N} I[O_i \subset C_k]$.

En la práctica, dada una partición inicial en *K clusters*, la técnica se basa en el siguiente algoritmo iterativo:

1. Cálculo de las posiciones de los centroides $\bar{\mathbf{x}}_{(k)}$ de los *K clusters*.
2. Para cada objeto, cálculo de su distancia a los *K* centroides, $e^2_{i(k)}$.
3. Reasignación de cada objeto al cluster cuyo centroide es el más próximo.

Es un hecho destacable que la solución (partición) final depende de la configuración inicial de los *clusters* elegida, siendo posible la convergencia a un mínimo local de *TESS*. Una opción recomendable y que suele ofrecer buenos resultados es la de realizar un análisis *cluster* jerárquico y elegir como partición inicial la obtenida con un nivel de disimilaridad que aplicado al árbol ultramétrico conduzca al número de grupos deseado.

Fuzzy K-means es una generalización en el ámbito de la lógica difusa del algoritmo *K-means*. La idea original de utilizar conjuntos difusos en técnicas de análisis *cluster* fue una propuesta de Ruspini (1969), desarrollada posteriormente por Bezdek (1974, 1981), Dunn (1974), Gustafson y Kessel (1979). Introduciendo la lógica difusa en el funcional de partición se obtiene la expresión (Bezdek, 1981, 1987):

$$TESS_{K,m} = \sum_{k=1}^{K} J^2_{(k),m} = \sum_{k=1}^{K} \sum_{i=1}^{N} u^m_{i(k)}\, e^2_{i(k)}$$

siendo $u_{i(k)}$ la pertenencia del elemento O_i al conjunto difuso C_k y $m \in (1,\infty)$ un exponente de *fuzziness* que determina la incidencia de las pertenencias difusas en la partición (cuanto más alto sea el exponente más difusa será la partición resultante). La pertenencia $u_{i(k)}$ de O_i a C_k se calcula mediante la expresión (Bezdek, 1981, 1987):

$$u_{i(k)} = \frac{1}{\sum_{l=1}^{K} \left[\frac{e_{i(k)}}{e_{i(l)}} \right]^{2/(m-1)}}$$

que cumple la restricción $\sum_{k=1}^{K} u_{i(k)} = 1$. Para el caso límite $m = 1$:

$$u_{i(k)} = \begin{cases} 1 & si\ e_{i(k)} = \min_{l}\{e_{i(l)}\} \\ 0 & si\ e_{i(k)} \neq \min_{l}\{e_{i(l)}\} \end{cases}$$

con lo que se obtiene la versión no difusa (*crisp*) de *K-means*.

Según lo expuesto, ambas modalidades (*fuzzy* y *crisp*) del algoritmo presuponen un espacio euclídeo *P*-dimensional y la distancia euclídea como medida de proximidad. Sin embargo, en muchas aplicaciones, ésta se ha revelado como una medida inadecuada. Un ejemplo de ello son los estudios en el ámbito de la ecología, donde se han propuesto otras funciones de similaridad o distancia que han resultado más útiles e interpretables para medir semejanzas o diferencias entre unidades ecológicas (ver, por ejemplo, Legendre y Legendre, 1998, para una descripción y discusión detallada del tema).

Debido a la restricción anterior, si la medida de proximidad que se desea utilizar es distinta de la distancia euclídea, se han propuesto diversas soluciones:

a) Para algunas proximidades que admiten una representación en un espacio euclídeo, es posible transformar las variables originales de modo que la distancia euclídea calculada sobre los datos transformados sea equivalente a la medida deseada respecto a los datos originales. Por ejemplo, es posible este enfoque con la distancia ji-cuadrado o la distancia de Hellinger (ver Legendre y Gallagher, 2001).

b) En otros casos, la proximidad admite una representación en un espacio euclídeo pero no existe una tranformación directa de los datos originales que permita reproducir la distancia deseada. La solución propuesta por algunos autores consiste en la realización de un análisis de coordenadas principales (*PCoA*), de modo que las coordenadas obtenidas serán las variables a utilizar en *K-means*.

c) Finalmente, la medida de proximidad puede no admitir una representación en un espacio euclídeo. En este caso, la solución utilizada por algunos autores consiste de nuevo en la realización de un análisis de coordenadas principales y la obtención de una representación en $\Re^M$, $M < P$ (ver p.e. Dufrêne y Legendre, 1997). Otra opción posible es la realización de un *non-metric scaling*.

Por otro lado, si $d_{ij} = d(O_i, O_j)$ es la distancia euclídea entre el elemento O_i y el elemento O_j, es bien conocida la relación:

$$E^2_{(k)} = \sum_{i=1}^{N} I[O_i \subset C_k]\ (\mathbf{x}_i - \bar{\mathbf{x}}_{(k)})'(\mathbf{x}_i - \bar{\mathbf{x}}_{(k)}) = \frac{1}{2n_k} \sum_{i,j=1}^{N} I[O_i \subset C_k]\, I[O_j \subset C_k]\, d_{ij}^2$$

En base a esta equivalencia, se pueden encontrar diversas referencias sobre la posibilidad de utilizar *K-means* a partir de matrices de distancias (Dufrêne y Legendre, 1997; Legendre y Legendre, 1998; Kaufman y Rousseeuw, 1990; *S-Plus 6.0*, 2000). Sin embargo no hemos encontrado *software* disponible que implemente una versión de *K-means* que evite el cálculo de las coordenadas de los centroides, ni referencias bibliográficas que describan las expresiones explícitas necesarias para dicha implementación. Por otra parte, ¿se puede generalizar el resultado a cualquier otra medida de proximidad? Según diversas referencias, ello no es posible (Kaufman y Rousseeuw, 1990; Legendre y Gallagher, 2001) o no se manifiesta explícitamente (p.e. Legendre y Legendre, 1998; Marsili-Libelli, 1991).

En el presente trabajo se presenta una aproximación basada en distancias que generaliza la utilización de *K-means* y *fuzzy K-means* a prácticamente cualquier medida de proximidad. A partir de una matriz de disimilaridades, las expresiones aportadas permiten calcular la distancia a los centroides sin necesidad de obtener explícitamente las coordenadas de los mismos.

Por otra parte, el algoritmo de reasignación de los objetos en *K-means* (y *fuzzy K-means*) presenta un inconveniente: la distancia de un objeto al cluster al que ha sido previamente asignado tiene 'sesgo', tanto mayor cuanto menos objetos contenga dicho cluster. Efectivamente, la distancia se reduce debido al 'efecto atractor' que realiza el objeto sobre el centroide por su pertenencia *a priori* al *cluster*. En el caso límite, un *cluster* que en una determinada iteración contenga un único objeto, va a quedar definitivamente inmovilizado. Presentamos también en este trabajo una modificación de las distancias que elimina el 'efecto atractor' mencionado, equivalente a la extracción momentánea del objeto del *cluster*.

2. *K-MEANS* Y *FUZZY K-MEANS* A PARTIR DE MATRICES DE DISIMILARIDADES: *DB K-MEANS*

Sea **X** un vector aleatori *P*-dimensional definido sobre un espacio de probabilidad (Π, A, $\boldsymbol{P}$) que toma valores $S \subset \Re^P$ con función de densidad de probabilidad f respecto a una medida adecuada λ. Consideremos $d(\cdot,\cdot)$ una función de disimilaridad definida sobre las parejas de elementos de Π, tal que su cuadrado sea integrable en *S*. La *variabilidad geométrica* de X respecto a $d(\cdot,\cdot)$ (Cuadras y Fortiana, 1995) viene definida por:

$$V_d(\mathbf{X}) = \frac{1}{2} E\left[d^2(\mathbf{X}_1, \mathbf{X}_2) \,\middle|\, \mathbf{X}_1, \mathbf{X}_2 \in S\right] = \frac{1}{2} \int_{S \times S} d^2(\mathbf{x}_1, \mathbf{x}_2)\, f(\mathbf{x}_1)\, f(\mathbf{x}_2)\, \lambda(d\,\mathbf{x}_1)\, \lambda(d\,\mathbf{x}_2)$$

Dado $\mathbf{x}_o \in \Re^P$, definimos la proximidad de $\mathbf{x}_o$ a la población Π con respecto a $d(\cdot,\cdot)$ como (Cuadras *et al*., 1997):

$$\phi_d^2(\mathbf{x}_o, \Pi) = \int_S d^2(\mathbf{x}_o, \mathbf{x})\, f(\mathbf{x})\, \lambda(d\,\mathbf{x}) \; - V_d(\mathbf{X})$$

Si existe una representación de $d(\cdot,\cdot)$, es decir, existe una función $\psi : R^P \rightarrow L$ ($L, \langle\cdot,\cdot\rangle$ simboliza un espacio euclídeo o Hilbert con producto escalar $\langle\cdot,\cdot\rangle$, de manera que $\|\mathbf{u}\|^2 = \langle\mathbf{u},\mathbf{u}\rangle$ es la norma natural para todo $\mathbf{u} \in L$), entonces:

a) $V_d(\mathbf{X}) = E(\|\psi(\mathbf{X}) - E(\psi(\mathbf{X}))\|^2)$

b) $\phi_d^2(\mathbf{x}_o) = \|\psi((\mathbf{x}_o) - E(\psi(\mathbf{X}))\|^2$

suponiendo que $E(\psi(\mathbf{X}))$ y $E(\|\psi(\mathbf{X})\|^2)$ sean finitas.

En el caso muestral, dadas muestras $\mathbf{x}(k)_1, \ldots, \mathbf{x}(k)_{n_k}$ de Π_k, $k = 1, \ldots, K$, la distancia de un elemento $\mathbf{x}_o$ al grupo Π_k será (Cuadras *et al.*, 1997):

$$\hat{\phi}_d^2(\mathbf{x}_o, \Pi_k) = \frac{1}{n_k}\sum_{h=1}^{n_k} d^2(\mathbf{x}_o, \mathbf{x}(k)_h) - \hat{V}_d(k)$$

$$\hat{V}_d(\Pi_k) = \frac{1}{2n_k^2}\sum_{h,l} d^2(\mathbf{x}(k)_h, \mathbf{x}(k)_l)$$

La generalización que proponemos para *K-means* consiste en considerar las relaciones $e_{i(k)}^2 = \hat{\phi}_d^2(\mathbf{x}_i, \Pi_k)$, $E_{(k)}^2 = n_k \hat{V}_d(\Pi_k)$, de modo que en el caso *crisp* tenemos:

$$E_{(k)}^2 = \frac{1}{2n_k}\sum_{i,j=1}^{N} I[O_i \subset C_k]\, I[O_j \subset C_k]\, d_{ij}^2$$

$$e_{i(k)}^2 = \frac{1}{n_k}\sum_{h=1}^{N} I[O_h \subset C_k]\, d_{ih}^2 - \frac{1}{2n_k^2}\sum_{h,l=1}^{N} I[O_h \subset C_k]\, I[O_l \subset C_k]\, d_{hl}^2$$

En el caso *fuzzy*, pueden determinarse expresiones equivalentes a las anteriores. Se obtiene que el *ESS* del *cluster* C_k es*:*

$$J_{(k),m}^2 = \frac{\sum_{i,j=1}^{N} u_{ik}^m u_{jk}^m d_{ij}^2}{2\sum_{i=1}^{N} u_{ik}^m}$$

Finalmente, la distancia $e_{i(k)}^2$ del objeto O_i al centroide del cluster C_k es:

$$e_{i(k)}^2 = \frac{1}{\sum_{h=1}^{N} u_{hk}^m}\sum_{h=1}^{N} u_{hk}^m d_{ih}^2 - \frac{1}{2\left(\sum_{h=1}^{N} u_{hk}^m\right)^2}\sum_{h,l=1}^{N} u_{hk}^m u_{lk}^m d_{hl}^2$$

3. *LOO DB K-MEANS*: CORRECCIÓN DEL 'EFECTO ATRACTOR'

Tal como hemos explicado, es evidente que los objetos ejercen un 'efecto atractor' sobre los centroides (en el caso *crisp*, únicamente sobre el centroide del *cluster* al cual ha sido asignado). Con el objetivo de eliminar este efecto, proponemos calcular la distancia de un elemento a cada centroide eliminando dicha influencia. La propuesta tiene una evidente analogía con una validación cruzada o *leave-one-out* (*LOO*). La expresión que se obtiene es la siguiente:

$$e_{i(k)}^{2\,(LOO)} = \left(\frac{\sum_{h=1}^{N} u_{hk}^{m}}{\left(\sum_{h=1}^{N} u_{hk}^{m} \right) - u_{ik}^{m}} \right)^2 e_{i(k)}^{2}$$

En el caso de la versión *crisp*, la expresión es entonces:

$$e_{i(k)}^{2\,(LOO)} = \left(\frac{n_k}{n_k - I[O_i \subset C_k]} \right)^2 e_{i(k)}^{2}$$

Es evidente que con la corrección propuesta el *TESS* será mayor, debido a la eliminación de la influencia del elemento sobre el centroide. En el caso *crisp*, la relación entre ambos valores de *TESS* es bien sencilla:

$$TESS_{K}^{(LOO)} = \sum_{k=1}^{K} E_{(k)}^{2\,(LOO)} = \sum_{k=1}^{K} \left(\frac{n_k}{n_k - 1} \right)^2 E_{(k)}^{2}$$

En definitiva, ambos valores no deben compararse directamente, puesto que obedecen a criterios distintos de evaluación de las distancias de los elementos a los *clusters*. Por otra parte, cabe resaltar que la corrección conlleva la desaparición de un *cluster* cuando contenga un solo objeto. En este caso, sería recomendable un estudio detallado del individuo para detectar si se trata de un *outlier* (por ejemplo, comprobar el efecto que tendría su eliminación sobre el *ESS* del *cluster* al que finalmente haya sido asignado).

4. EJEMPLO DE APLICACIÓN

Con el objeto de mostrar un ejemplo de la utilización de *db K-means* y *db fuzzy K-means*, se han seleccionado 118 inventarios de comunidades vegetales pertenecientes a 15 asociaciones de las alianzas *Xerobromion erecti* y *Mesobromion erecti*, correspondientes a pastos xerófilos y mesófilos de tendencia medioeuropea (ver Font, 1993). Para cada inventario, se dispone de la medida de abundancia (escala de Braun-Blanquet) de 499 especies. A partir de los datos se han calculado distintas medidas de proximidad de interés

fitosociológico. Considerando únicamente la presencia o ausencia de las especies, se han utilizado dos medidas de similaridad para datos binarios:

- *Simple Matching Coefficient* (*SMC*)

$$s_1(O_1, O_2) = \frac{a+d}{a+b+c+d}$$

- Índice de Jaccard

$$s_2(O_1, O_2) = \frac{a}{a+b+c}$$

donde a es el número de dobles presencias, d las dobles ausencias y $b+c$ es el número de especies para el que ambas comunidades no coinciden. Se ha utilizado la transformación de Gower (1966)

$$d_{ij} = (s_{ii} + s_{jj} - 2\, s_{ij})^{1/2}, \quad s_{ij} = s(O_i, O_j)$$

para transformar las similaridades en distancias.

A partir de los datos cuantitativos de las abundancias, se ha aplicado la transformación combinada propuesta por van der Maarel (1979), recomendada en diversos estudios posteriores (p.e. Hajdu, 1981). Una vez transformados los datos, se han utilizado las siguientes medidas de distancia entre los inventarios ($\mathbf{X}(O_i) = \mathbf{x}_i$ simboliza el vector de abundancias transformadas de las 499 especies para el inventario O_i):

- Distancia Euclídea

$$d_1(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{P} (x_{1j} - x_{2j})^2}$$

- Distancia de Bray-Curtis

$$d_2(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{P} \left| x_{1j} - x_{2j} \right|}{\sum_{j=1}^{P} (x_{1j} + x_{2j})^2}$$

- Distancia de Hellinger

$$d_3(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{P} \left[\sqrt{\frac{x_{1j}}{x_{1+}}} - \sqrt{\frac{x_{2j}}{x_{2+}}} \right]^2}, \quad x_{i+} = \sum_{j=1}^{P} x_{ij}$$

A partir de las expresiones presentadas en este trabajo, se ha ejecutado el *db K-means* y el *db fuzzy K-means* ($m = 1{,}1$) para generar particiones en 15 grupos. Como configuración inicial se ha elegido la partición generada a partir de un análisis *cluster* jerárquico de Ward (1963), debido a que la función que optimiza en cada paso es

equivalente al *TESS*. Todos los cálculos se han repetido utilizando la corrección propuesta del 'efecto atractor' (*loo db K-means* y *loo db fuzzy K-means*).

Para medir la recuperación de la estructura de grupos original se han comparado las particiones obtenidas con la clasificación fitosociológica en asociaciones. Dicha comparación se ha realizado mediante el índice de Rand corregido (Hubert and Arabie, 1985). Los resultados se presentan en la tabla 1 para el caso *crisp* y en la tabla 2 para el caso *fuzzy*. En cada tabla se muestra el índice de Rand y el mínimo de la función optimizada *TESS* o $TESS^{(LOO)}$.

	db K-means		***loo db K-means***	
Medida de proximidad	**Rand**	**TESS**	**Rand**	**TESS(LOO)**
Distancia euclídea	0,84	20408	0,88	26183
SMC	0,74	2023	0,69	2642
Índice de Jaccard	0,83	25,10	0,81*	33,33*
Distancia Bray-Curtis	0,87	18,32	0,87	23,93
Distancia Hellinger	0,90	51,99	0,83*	67,94*

Tabla 1: *K-means* basado en distintas funciones de proximidad $(K = 15)$. Los valores marcados con un asterisco indican que se ha producido la pérdida de un *cluster* en el proceso de clasificación.

	db Fuzzy K-means		***loo db Fuzzy K-means***	
Medida de proximidad	**Rand**	**TESS**	**Rand**	**TESS(LOO)**
Distancia euclídea	0,89	20080	0,91	25120
SMC	0,76	2011	0,72	2474
Índice de Jaccard	0,84	25,26	0,80	31,74
Distancia Bray-Curtis	0,94	18,10	0,91	23,09
Distancia Hellinger	0,94	51,44	0,89	64,66

Tabla 2: *fuzzy K-means* basado en distintas funciones de proximidad ($K = 15$, $m = 1{,}1$).

Observamos que el mejor ajuste (índice de Rand) a la clasificación fitosociológica es para la distancia de Hellinger y la de Bray-Curtis, mientras que el peor lo ha obtenido el *SMC* (en realidad, el *SMC*, una vez transformado en distancia, equivale a la distancia euclídea con los datos binarios). Por otra parte, la versión *fuzzy* para $m = 1{,}1$ ha obtenido en todos los casos un ajuste mejor a la estructura de grupos original.

En dos casos (Jaccard y Hellinger), un *cluster* se ha quedado con un único objeto y ha sido eliminado en *loo db K-means*, con lo que la partición se ha reducido a 14 *clusters*.

Aún así, el índice de Rand puede calcularse, puesto que no requiere que las particiones tengan un número de grupos idéntico.

En la tabla 2 sorprende el peor comportamiento de *loo fuzzy K-means* frente al *fuzzy K-means* sin corrección. Esto es debido a que la introducción de la corrección *loo* genera particiones más difusas como consecuencia del aumento de las distancias de los objetos a los centroides, de modo que los resultados para un mismo exponente de *fuzziness* (*m*) no son comparables.

Para mostrar este aumento de la *fuzziness*, se ha ejecutado *db fuzzy K-means* y su correspondiente versión *loo* con exponentes *m* comprendidos entre 1,025 y 1,3, utilizando la distancia de Hellinger. En la figura 1 se puede observar cómo los valores máximos del índice de Rand en las dos versiones se alcanzan para valores de *m* distintos.

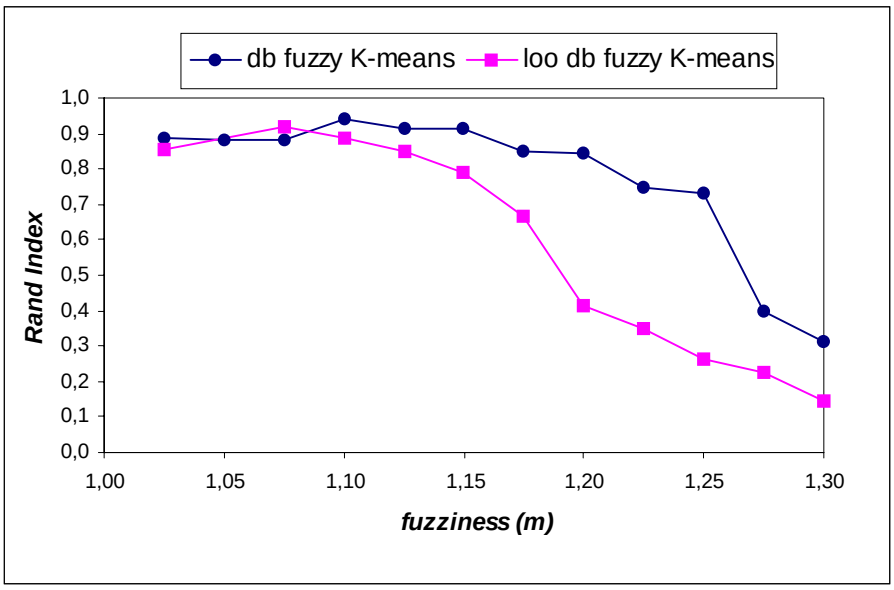

Figura 1 Relación del exponente de *fuzziness* (*m*) con el ajuste de la partición resultante medido con el índice de Rand corregido. Resultados obtenidos utilizando la distancia de Hellinger.

Para un mismo valor de *m*, *loo db fuzzy K-means* siempre generará una partición final más difusa que *db fuzzy K-means*, con lo que sus resultados no son comparables. En la figura 2, se muestra la relación del ajuste (índice de Rand) respecto a la *fuzziness* de la partición medida con el índice de Dunn normalizado (Dunn, 1976). Puede observarse que para una misma *fuzziness* de la partición el ajuste de ambas estrategias de optimización es parecido.

El ejemplo que se ha mostrado es solo una ilustración del comportamiento, a grandes rasgos, de las propuestas de este estudio. Se hace necesario realizar otros estudios

de eficiencia en la recuperación de estructuras de grupos en otros casos aplicados y en estudios de simulación mediante generación de *clusters* artificiales.

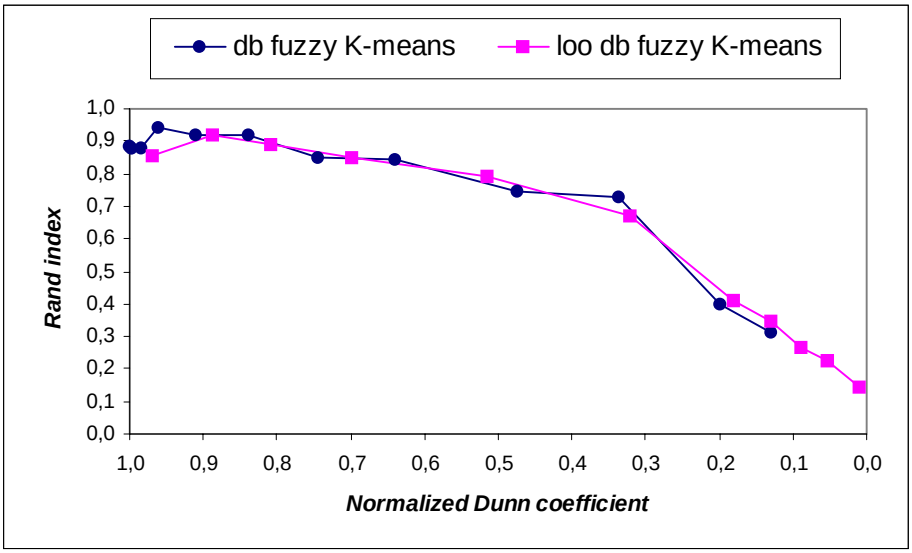

Figura 2 Relación del coeficiente de Dunn normalizado con el ajuste de la partición resultante medido con el índice de Rand corregido. Resultados obtenidos utilizando la distancia de Hellinger.

5. CONCLUSIONES

Se han propuesto sendas extensiones de *K-means* y *fuzzy K-means* para su uso a partir de cualquier matriz de proximidades entre objetos sin necesidad de un cálculo explícito de las coordenadas de los centroides. Así mismo, se ha propuesto una corrección para evitar la influencia del individuo sobre la posición de los centroides en su reasignación.

6. AGRADECIMIENTOS

El presente trabajo se ha realizado con el soporte del "Comissionat per a Universitats i Recerca" (1999SGR00059), del "Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya" (2001 FI 00269) y del "Ministerio de Ciencia y Tecnología" (BFM 2000-0801).

7. REFERENCIAS

Bezdek, J.C. (1974): "Numerical taxonomy with Fuzzy sets". *J. Math. Biol.* 1 (1), 57-71.

Bezdek, J.C. (1981). *Pattern Recognition with Fuzzy Objective Function*. Plenum Press.

Bezdek, J.C. (1987): "Some non-standard clustering algorithms" en: Legendre, P. & Legendre, L. *Developments in Numerical Ecology. NATO ASI Series, Vol. G14*. Springer-Verlag.

Cuadras, C.M, Fortiana, J. (1995): "A continuous metric scaling solution for a random variable". *Journal of Multivariate Analysis* 52, 1-14.

Cuadras, C.M, Fortiana, J, Oliva, F. (1997): "The proximity of an individual to a population with applications in discriminant analysis". *Journal of Classification* 14, 117-136.

Data Analysis Division (2000), *S-Plus 6.0: Guide to Statistics*, MathSoft, Inc.

Dufrêne, M., Legendre P. (1997): "Species assemblages and indicator species: The need for a flexible asymetrical approach". *Ecological Monographs* 67(3), 345-366.

Dunn, J.C. (1974): "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact, Well Separated Clusters". *J. Cybern.* 3, 32-57.

Dunn, J.C. (1976): "Indices of partition fuzziness and the detection of clusters in large data sets" en: Gupta, M. (ed) *Fuzzy automata and decision processes* . Elsevier.

Font, X. (1993): *Estudis geobotànics sobre els prats xeròfils de l'estatge montà dels Pirineus*. Arxius de la secció de ciències CV. Secció de ciències biològiques. Institut d'Estudis Catalans.

Gower, J.C. (1966): "Some distance properties of latent roots and vector methods used in multivariate analysis". *Biometrika* 53, 325-338.

Hajdu L.J. (1981): "Graphical comparison of resemblance measures in phytosociology". *Vegetatio* 48, 47-59.

Hubert L., Arabie P. (1985): "Comparing partitions". *Journal of Classification* 2, 193-218.

Kaufman L., Rousseuw P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: Willey & Sons Inc.

Legendre, P., Gallagher, E.D. (2001): "Ecologically meaningful transformations for ordination of species data". *Oecologia.* In press.

Legendre P, Legendre L. (1998): *Numerical Ecology. Second english edition.* Elsevier.

MacQueen J. (1967): "Some methods for classification and analysis of multivariate observation". pp. 281-297 en: L.M. Le Cam & J. Neyman (eds.) *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1*. University of California Press, Berkeley.

Marsili-Libelli, S. (1991): "Fuzzy clustering of ecological data". pp. 173-184 en: E. Feoli & L. Orlóci (eds.) *Computer Assisted Vegetation Analysi*. Kluver Academic Publishers.

Ruspini, E. (1969): "A New Approach to Clustering". *Inf. Control* 15, 22-32.

van der Maarel E. (1979): "Transformation of cover-abundance values in phytosociology and its efects on community similarity". *Vegetatio* 39(2), 97-114.

Ward J.H. (1963): "Hierarchical grouping to optimize an objective function". *J. Amer. Stat. Ass.* 58, 236-244.