ESTADÍSTICA ESPAÑOLA
Vol. 43, Núm. 148, 2001, págs. 209 a 224

Nuevos métodos para el diseño de clusters no jerárquico. Una aplicación a los municipios de Castilla y León

por
MAURICIO BELTRAN PASCUAL
Institut Balear d'Estadística
Comunidad Autónoma de les Illes Balears

JOAQUÍN A. PACHECO BONROSTRO
Departamento.de Economía Aplicada. Universidad de Burgos
Facultad de Ciencias Económicas y Empresariales

RESUMEN

En este trabajo se proponen algunas modificaciones en algoritmos de Búsqueda Local existentes en la literatura para el problema de diseño de clusters, como el conocidísimo K-medias o el recientísimo J-medias propuesto por Hansen y Mladenovic, (1999). Estas modificaciones consisten en el diseño de un método ávido-aleatorio en dos fases para la construcción de una solución inicial de partida, que siguiendo la filosofía de los metaheurísticos GRASP, aseguren la diversidad y calidad en las soluciones obtenidas. Las soluciones finales resultantes son mejores que las que se llegan usando otras soluciones iniciales. Posteriormente se aplican estas modificaciones a un estudio sobre la actividad económica de los municipios de Castilla y León, comparándose los resultados con los obtenidos por otros métodos incluyendo los usados por conocidos paquetes estadísticos como el SPSS.

Palabras clave: clusterización, soluciones ávido-aleatorias, clasificación de municipios.

Clasificación AMS: 68T20, 62H30

1. INTRODUCCIÓN

Sea $X = \{x_1. x_2..... x_N\}$ un conjunto de N puntos en R^q y m un número entero positivo predeterminado. El problema de diseño de clusters con el criterio de mínima suma de cuadrados (MSSC, *Minimun Sum-of-Squares Clustering*) consiste en encontrar una partición de X en m subconjuntos disjuntos (clusters) de forma que la suma de las distancias euclídeas al cuadrado de cada punto al centroide de su cluster sea mínima.

Más concretamente, sea P_m el conjunto de todas las particiones de X en m conjuntos; cada partición $P \in P_m$ viene definida de la forma $P = (C_1. C_2.... C_m)$, donde C_i denota cada uno de los clusters que forman P; entonces el problema puede formularse como sigue:

$$\min_{P \in P_m} \sum_{i=1}^{m} \sum_{x_l \in C_i} \left\| x_l - \bar{x}_i \right\|^2$$

donde el centroide $\bar{x}_i$ se define como:

$$\bar{x}_i = \frac{1}{n_i} \sum_{x_l \in C_i} x_l \text{ ; para } i = 1 .....m.$$

siendo $n_i = |C_i|$; o equivalentemente

$$\min \sum_{l=1}^{N} \left\| x_l - \bar{x}_{c(l)} \right\|^2$$

donde c(l) es el cluster al que pertenece el elemento l.

El diseño de clusters es una de las herramienta más conocidas en el Análisis de Datos, (más concretamente dentro del Reconocimiento de Patrones). Es usada como técnica de análisis exploratorio previo que intenta determinar si en un conjunto de casos (que se identificarían con los puntos de X) existe una posible estructura que permita clasificar dichos casos en subconjuntos o clusters. El problema que aquí se trata se puede enmarcar dentro del diseño de clusters no jerárquico y

tiene muchas aplicaciones en Ciencias Sociales, Económicas y Naturales. Se sabe que es NP-Hard, (Brucker, (1978)).

Existen en la literatura varios métodos exactos para el MSSC, como el descrito en Koontz y otros, (1975), Diehr, (1.985); algunos de los cuales, como el propuesto en du Merle y otros, (1997), consiguen resolver problemas de hasta 150 puntos.

Para problemas de tamaño superior sigue siendo necesario la aplicación de algoritmos heurísticos. Quizás los más populares sean los basados en procesos de Búsqueda Local, como los conocidísimos K-medias, (Jancey, (1966)) y H-medias, (Howard, (1966)). En un recientísimo trabajo Hansen y Mladenovic, (1999), proponen un nuevo proceso de Búsqueda Local, denominado J-medias así como variantes de los ya existentes: H-medias+, HK-medias.... Como ocurre en buena parte de los problemas combinatorios desde hace algunos años está de moda el diseño de Metaheurísticos que usan estos movimientos vecinales o locales, como Temple Simulado (Klein y Dubes, (1989)), Búsqueda Tabú, (Al-Sultan, (1995)), Genéticos, (Babu y Murty, (1993)) o la más reciente Búsqueda en Vecindarios Variables, (du Merle y otros, (1997), y Hansen y Mladenovic, (1999)).

En este trabajo se muestran algunas experiencias con heurísticos, más concretamente el diseño de un método ávido-aleatorio en 2 fases para la obtención de diversas soluciones iniciales de relativa calidad, y que combinado con algún algoritmo de Búsqueda Local pueda dar lugar a conjuntos de soluciones diversas y de calidad, -como en los Metaheurísticos tipo GRASP-, (Feo y Resende (1995)).

El trabajo se estructura de la siguiente manera: en la sección siguiente se describen algunos de los algoritmos de Búsqueda Local ya existentes que se emplean en este trabajo; en la tercera sección se describe el método para la obtención de soluciones iniciales en dos fases; en la cuarta se muestran los resultados de diferentes experimentos computacionales para contrastar la eficacia y eficiencia de este método; en la quinta se aplican los procedimientos y algoritmos analizados a un estudio de la actividad económica de los municipios de Castilla y León; finalmente en la sexta se establecen las conclusiones.

2. PRINCIPALES ALGORITMOS DE BÚSQUEDA LOCAL

Como se acaba de mencionar en esta sección se van a describir los algoritmos con los que se va a trabajar en este trabajo: los populares *H-medias* y *K-medias*, así como los más recientes *H-medias+*, *HK-medias*, *J-medias* y *J-medias+*.

En todos los casos se inician obteniendo una solución, (o partición), inicial aleatoria (C_1. C_2..... C_M)

Algoritmo H-Medias

Repetir

Calcular centroides $\bar{x}_i$. para i=1….m (Paso 1)

Reasignación de puntos a centroides más cercanos (→ nuevos clusters) (Paso 2) hasta que no haya modificaciones

Variante H-Medias+

Repetir

Ejecutar (Paso 1) *y* (Paso 2) *de* Algoritmo H-Medias

Si no hay modificaciones:

– Comprobar si existen clusters vacíos (degeneración)

– Si hay k clusters vacíos sustituirlos por k nuevos clusters unipuntuales con los k puntos más alejados de los centroides de sus clusters hasta que no haya modificaciones

Algoritmo K-Medias

El popular algoritmo de las *K-medias* consiste, básicamente, en mover en cada paso un punto de su cluster a otro diferente. En cada paso se ejecuta el mejor de estos movimientos. El algoritmo finaliza cuando no hay mejora.

Repetir

$\forall$ i =1..m. $\forall$ $x_j \notin C_i$ calcular v_{ij} aumento(*1*) en la función objetivo de reasignar x_j a C_i

Si $v_{i^*j^*} = \min v_{ij} < 0$ entonces reasignar x_{j^*} a C_{i^*}

hasta $v_{i^*j^*} \geq 0$

De Späth, (1980) se obtienen las siguientes fórmulas para simplificar los cálculos: Sea C_l el cluster al que pertenece x_j el valor de v_{ij} se calcula de la forma siguiente

$$v_{ij} = \frac{n_i}{n_i + 1} \cdot \left\| \bar{x}_i - x_j \right\|^2 - \frac{n_l}{n_l - 1} \cdot \left\| \bar{x}_l - x_j \right\|^2 ;$$

(1) Es decir, v_{ij} indica aumento en la función objetivo si es positivo y disminución si es negativo. Por tanto, el proceso para cuando no existe v_{ij} negativos

y cuando el cambio de x_j desde C_l hasta C_i se lleva a cabo los centroides se actualizan fácilmente

$$\bar{x}_l = \frac{n_l \bar{x}_l - x_j}{n_l - 1}\ ;\ y\ \bar{x}_i = \frac{n_i \bar{x}_i + x_j}{n_i + 1}\ .$$

Variante HK-Medias

Ejecutar H-Medias+

Ejecutar K-Medias

Un óptimo local para *K-medias* lo es también para *H-medias* (y *H-medias+*), pero no necesariamente al revés, por tanto se ejecuta *K-medias* posteriormente a *H-medias+*.

J-Medias

Repetir

$\forall$ j =1....N/ x_j no coincide con ningún centroide actual; añadir un cluster ficticio C_{M+1} de centroide x_j; y $\forall$ i=1....m calcular J_{ij} el aumento en la función objetivo de 'disolver' el cluster C_i y reasignar cada uno de sus elementos al cluster de centroide más cercano entre los restantes (incluido C_{M+1}) (Jump-moviento)

Si $J_{i^*j^*}$ = min $J_{ij} < 0$ entonces 'deshacer' C_{i^*} y reasignar los elementos de C_i al cluster de centroide más cercano entre los restantes (incluyendo C_{M+1} de centroide x_{j^*}), redefinir $C_i = C_{M+1}$ y eliminar C_{M+1}.

hasta $J_{i^*j^*} \geq 0$.

Finalmente, *J-medias+* consiste en aplicar *HK-medias* una vez que se ejecuta cada Jump-movimiento en *J-medias*.

3. CONSTRUCCIÓN DE SOLUCIONES ÁVIDO-ALEATORIAS EN DOS FASES

Las dos fases de este método consisten básicamente en: 1) seleccionar m puntos (puntos-semilla) de X suficientemente alejados entre sí, que serán considerados los centroides iniciales y 2) ir asignando los puntos de X a cada uno de los clusters. Cada uno de estas dos fases se realiza iterativamente, paso a paso, según algún criterio o función que ayude a medir la 'bondad' de cada elección (punto-semilla en la primera fase, punto y cluster al que es asignado en la segunda).

Estas funciones de 'bondad' no necesariamente llevan a la mejor solución final, ya que solo miden la bondad en cada paso concreto, es decir, a corto plazo. Por tanto, siguiendo la idea fundamental de los Metaheurísticos GRASP, en cada paso, no se va a elegir necesariamente el mejor elemento según está función, sino que se selecciona aleatoriamente un elemento entre los de mayor 'bondad'.

A continuación se describe cada una de las dos fases con detalle:

3.1. Elección de m puntos semilla

Como se acaba de comentar se van a elegir m puntos semilla en X lo suficientemente alejados con la esperanza de que pertenezcan a cluster diferentes en la solución óptima. A continuación se muestra el seudo código de este proceso iterativo

Procedimiento Determinación_Ptos_Semilla

Calcular $\bar{x}$ (centroide de todos los puntos de X)

Calcular j* = arg max $\left\{ \left\| x_j - \bar{x} \right\| / j = 1,\ldots,n \right\}$

Hacer $S = \{j^*\}$ (S conjunto de los índices de los puntos-semilla escogidos)

Calcular $H(j) = \min\left\{ \left\| x_j - x_l \right\|^2 / l \in S \right\}$ para j=1....N; $j \notin S$

Mientras |S| < m hacer:

Calcular Hmax = max {H(j) / $j \notin S$}

Formar el conjunto L = {j / $j \notin S$, H(j) > alfa·Hmax}

Elegir aleatoriamente j* entre los elementos de L

Hacer $S = S \cup \{j^*\}$

Actualizar H(j) : Si H(j) < $\left\| x_j - x_{j^*} \right\|^2 \Rightarrow H(j) = \left\| x_j - x_{j^*} \right\|^2$. para j=1....N; $j \notin S$

En cada paso se va a elegir aleatoriamente, como nuevo punto semilla, un elemento entre los puntos más alejados a los puntos semilla ya seleccionados.

Obsérvese en la figura siguiente como eligiendo en cada paso el punto con mayor valor de H(j) no se llega necesariamente a la mejor distribución de los puntos-semilla.

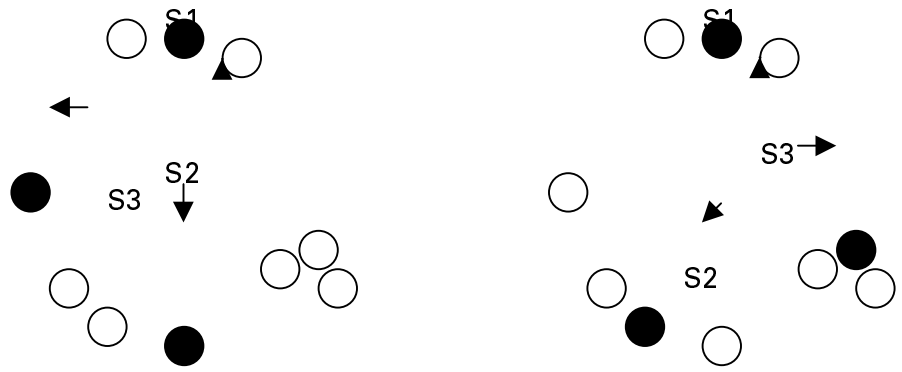


Figura1. La elección de cómo semillas de los puntos con mayor H(j) lleva en el dibujo de la izquierda a la elección de S1, S2 (punto más alejado de S2) y S3 (punto con mayor mínima distancia a S2 y S1). Sin embargo si se elige como S2 el segundo punto más alejado de S1 pueden llegarse a una elección de puntos-semilla más equilibrada de cara a la formación de clusters

Por tanto, una forma de elección aleatoria pero que tenga en cuenta la función de 'bondad', (en este caso H(j)), asegura variedad en las soluciones obtenidas, mantiene cierto grado de 'calidad' en las mismas y en muchas ocasiones superan a la solución obtenida por la elección del mejor movimiento (mayor H(j)). Para llevar a cabo esto se forma una lista de elementos élite y se elige aleatoriamente uno de esa lista con la misma probabilidad. Concretamente se forma la lista con los elementos j cuya bondad (H(j)) supera a la máxima (Hmax) por un coeficiente *alfa*, que en este caso vale 0'8.

3.2. Asignación de elementos a clusters

Una vez seleccionados los m puntos semilla éstos hacen de centroides iniciales. La asignación directa de cada punto a su centroide más cercano da lugar a relativas buenas soluciones iniciales. Sin embargo, se ha creído conveniente mejorar dicha asignación mediante un procedimiento paso a paso que tenga en cuenta la modificación de los centroides y el aumento real de la función objetivo según se van asignando los elementos.

Para cada elemento no asignado x_j. el incremento de la función objetivo de asignarle en el cluster i es:

$$e_{ij} = \frac{n_i}{n_i + 1} \cdot \left\| \bar{x}_i - x_j \right\|^2;$$

entonces se define P(j) = min $\{e_{ij} / i=1...m\}$ y r(j) = arg P(j).

La función P(j) va a ser usada como función de 'no-bondad' y hace de guía del siguiente.

Procedimiento asignación de puntos

Hacer A = {∅}

Iniciar P(j) y r(j) j=1.....N

Mientras |A| < N hacer:

Calcular Pmin= min {P(j) / j =1.....N. $j \notin A$ }

Formar L = {j/ $j \notin A$, P(j)*alfa < Pmin}

Elegir j* aleatoriamente entre los elementos de L

Hacer $C_{r(j^*)} = C_{r(j^*)} \cup \{x_{j^*}\}$. $A = A \cup \{j^*\}$ actualizar $n_{r(j^*)}$ y $\bar{x}_{r(j^*)}$

Actualizar P(j). j =1....N, $j \notin A$

La inserción de un punto x_{j^*} en un clúster modifica el centroide correspondiente y puede hacer cambiar el clúster al que inicialmente se asignan los puntos que quedan por insertar, (es decir, el valor de r(j)).

4. RESULTADOS COMPUTACIONALES

En esta sección se muestran los resultados de una serie de pruebas realizadas para contrastar la eficacia y eficiencia de cada uno de los algoritmos y modificaciones propuestos en las 3 secciones anteriores. Como instancia de prueba se ha usado la denominada de la librería TSPLIB, con N = 1.060 puntos y con diferentes números de clusters, m = 10, 20, 30, ...,150, como se hace en Hansen y Mladenovic, (1999).

A continuación se muestran los resultados. Cada uno de las 3 subsecciones siguientes corresponden respectivamente a las pruebas realizadas con los algoritmos propuestos en las secciones 3, 4 y 5.

Los algoritmos utilizados en esta sección son los siguientes:

− SA (Solución Aleatoria): más concretamente, se generan aleatoriamente m centroides (con distribución uniforme en el intervalo de valores de X) y se asignan directamente cada punto al centroide más cercano.

− Variante *HK-medias*

– Algoritmo *J-medias*

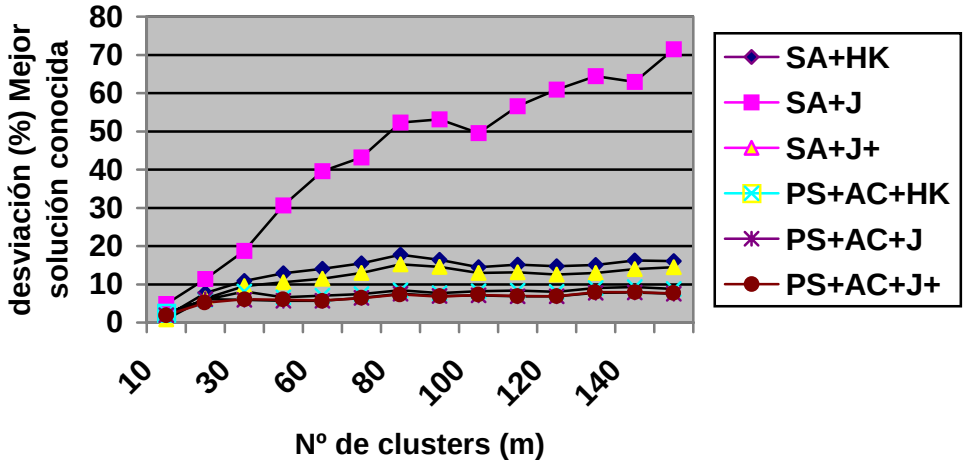
– Variante *J-medias+*.

– PS : Procedimiento *Determinación_Ptos_Semilla* descrito en la sección 3.1. Se considera que la solución obtenida es la resultante de asignar cada punto al punto-semilla más cercano.

– AC: Procedimiento *Asignación de puntos* descrito en la sección 3.2.

Se han realizado una serie de pruebas. Cada una de estas pruebas consiste en la generación de una solución aleatoria (ejecución de SA) y de otra solución ávido-aleatoria y, a partir de dichas soluciones, ejecutar los algoritmos HK-medias, J-medias y J-medias+. Para cada nivel o número de clusters m se han repetido 20 veces dichas pruebas. Para facilitar el análisis de los resultados se muestra una gráfica con la evolución de las desviaciones medias respecto a la mejor solución conocida (los valores de estas mejores soluciones vienen facilitados en Hansen y Mladenovic, (1999), excepto para m = 40):

A partir de los resultados de esta gráfica y los cuadros anteriores se pueden extraer las siguientes conclusiones:



– La calidad de las soluciones obtenidas por SA se deterioran a medida que aumenta el número clusters. Por el contrario PS y AC dan soluciones de una calidad que se mantiene igual para todos los valores de m (alejamiento de la mejor solución conocida en torno al 30% y 15% respectivamente); esto hace que al aplicar los mismos procedimientos de Búsqueda Local, se lleguen en general a

mejores soluciones que cuando se parte de soluciones con centroides iniciales aleatorios, (SA).

5. APLICACIÓN A UN ESTUDIO SOBRE ACTIVIDAD ECONÓMICA EN LOS MUNICIPIOS DE CASTILLA Y LEÓN

Una vez contrastada la eficacia de la estrategia propuesta en problemas teóricos de librerías conocidas, nos preguntamos si dicha mejora también se refleja en estudios más prácticos y reales.

Para ello vamos a utilizar los procedimientos descritos en este trabajo en un análisis de los municipios de Castilla y León según su desarrollo económico. Se seleccionan una serie de variables y se toman todos los casos (2.245 municipios). A partir de estos datos se hace un análisis factorial. Las puntuaciones factoriales se usan para el diseño de los clusters.

5.1. Las variables del análisis y sus fuentes

La información referente al ámbito municipal es relativamente escasa en comparación con otros ámbitos geográficos: provincial, regional o nacional. Aún así, recabando en todas las fuentes oficiales y privadas se puede recopilar un numeroso conjunto de variables.

En el punto de partida del análisis efectuado se disponía, sin ser exhaustivos, de variables relativas a la renta municipal, estimaciones y datos del Impuesto de Renta de las Personas Físicas, número de licencias comerciales por sectores procedentes del Impuesto de Actividades Económicas, indicadores y datos de la Encuesta de Infraestructura y Equipamientos Locales, número de líneas telefónicas, número de oficinas de entidades bancarias, superficies agrarias cultivadas por tipos de cultivos, número de vehículos matriculados, variables de educación, variables procedentes del Nomenclator del Instituto Nacional de Estadística tales como: superficie municipal, altitud y distancia a la capital y, por supuesto, todos los datos de los Censos de Población y Vivienda.

Con la amplia información disponible se efectuaron diferentes ensayos factoriales hasta que se llegó a un conjunto de variables que puede entenderse adecuado respecto al análisis que se pretende realizar.

Las diez variables que al final se utilizan en el análisis factorial y el análisis cluster son las siguientes:

– Edad media de la población (emedia)

– Índice de dependencia global (idepenglo)

– Índice de dependencia senil (iepense)

– Rendimientos medios declarados (rendi)

– Estimación de la renta per cápita (renta)

– Número total de entidades financieras por cada mil habitantes (tfinanp)

– Número de licencias comerciales divididas por la población (licipp)

– Número de vehículos por cada mil habitantes (vehiculp)

– Número de líneas telefónicas por mil habitantes (telefonp)

– Número de parados divididos por la población (parop)

5.2. Resultados del análisis

El número de factores elegidos fueron de cuatro, que explican el 73,41% de la variabilidad total, lo que puede interpretarse como un porcentaje bastante aceptable.

KMO y prueba de Bartlett

Medida de adecuación muestral Kaiser-Meyer-		,699
Prueba de esferici-dad de Bartlett	Chi-cuadrado aproximado	11060,93
	gl	45
	Sig.	,000

El índice KMO que nos compara los coeficientes de correlación de Pearson con los coeficientes de correlación parcial entre variables es del 0,699 lo que nos garantiza que el análisis factorial es un procedimiento adecuado.

El test de Bartlett con un valor de 11.060,93 y un grado de significación p = 0,000 rechaza que la matriz de correlaciones sea una matriz identidad, lo que conlleva a seguir con el análisis factorial.

Además de los resultados comentados se observan valores muy bajos en las matrices antiimagen, y también los MSA (Measures of Sampling Adequacy) son

bastante altos, lo que nos lleva a concluir a priori que resulta pertinente el análisis factorial y que esto puede proporcionarnos conclusiones satisfactorias.

A continuación, se presenta la matriz de componentes rotados que nos ayuda a interpretar los resultados obtenidos.

Matriz de componentes rotados [a]

	Componente			
	1	2	3	4
emedia	,779	-,206	-,110	,189
idepenglo	,910	-,231	-4,8E-02	7,45E-02
idepense	,912	-,260	-,102	,139
RENDI	-,204	,729	,133	9,34E-02
RENTA97P	7,06E-03	,292	,237	,790
TFINANP	-7,6E-02	-6,5E-02	,849	,209
LICIP	-3,8E-02	,257	,767	-,183
VEHICULP	-,237	,737	6,12E-03	7,96E-02
TELEFONP	,620	,506	,144	-6,0E-02
PAROP	-,270	6,92E-02	,162	-,789

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 7 iteraciones.

5.3. Diseño de clusters: Análisis Comparativo

Como se ha mencionado anteriormente la matriz de puntuaciones factoriales obtenidas va a ser usada como 'banco de pruebas' para contrastar la eficacia de la estrategia propuesta en este trabajo para obtener soluciones iniciales, con datos reales (2245 casos y 4 variables-factores). Se elige un número de cluster igual a 3.

Para ello, se considera el procedimiento de Búsqueda Local *HK-medias*. Se realizan una serie de 10 pruebas. Cada una de estas pruebas consiste en: generar 1 solución aleatoria (SA) y ejecutar *HK-medias*; análogamente se generan 10 soluciones ávido-aleatorias (PS+AC) y se ejecuta *HK-medias* tras cada una de ellas A continuación se muestran los estadísticos de las soluciones obtenidas (mínimo, media y máximo) y los tiempo medios de computación (en segundos)

	SA+HK	*PS+AC+HK*
Mínimo	6190,997300	6190,997300
Medio	6219,040531	6190,997300
Máximo	6470,600847	6190,997300
T.Computación	1,2735	8,5656 +1,3656 = 9,9312

Como se ve en esta tabla, para un número de clusters más bajo (3 como en este caso) también se aprecian las mejoras con el empleo del método ávido-aleatorio propuesto para obtener soluciones iniciales.

Además de obtener resultados medios mejores también se mejora en la consistencia. Si se observa con soluciones iniciales aleatorias HK-medias consigue en alguna ocasión llegar al mejor resultado obtenido en estas pruebas (6190,9973); sin embargo, cuando se parte de soluciones ávido-aleatorias HK-medias consigue el mismo resultado en todas las ocasiones.

Finalmente, comparamos los estadísticos de esta mejor solución (Función objetivo = 6190,9973) con los de la obtenida por el algoritmo de las K-medias del paquete SPSS ver. 9.0 (Función objetivo = 6222,7076)

Mejor Solución Trabajo:

Variable	*Análisis de la Varianza*		
	Entre Grupos	*Dentro de los Grupos*	*F*
Factor 1	17,741071	0,985073	18,009912
Factor 2	244,160628	0,783183	311,754378
Factor 3	698,260107	0,378279	1845,88654
Factor 4	434,339543	0,613607	707,846248

Solución K-medias de SPSS

	Análisis de la Varianza		
	Entre Grupos	*Dentro de los Grupos*	*F*
Factor 1	44,621861	0,961104	46,427713
Factor 2	124,756838	0,889651	140,231273
Factor 3	685,097279	0,390016	1756,58852
Factor 4	524,170223	0,533508	982,496489

Aunque el número de clusters sea pequeño se aprecian muy ligeras diferencias. La solución obtenida en nuestro análisis discrimina mejor en el conjunto de los factores dado que la suma de la varianza entre grupos es mayor, y la suma de la varianza intragrupos menor.

6. CONCLUSIONES

La aportación realizada en este trabajo ha consistido, básicamente, en el diseño de un método rápido (ávido-aleatorio) para obtener soluciones de relativa calidad, que puedan servir de punto de partida en algoritmos de Búsqueda Local.

El método ávido-aleatorio para obtener soluciones ha resultado interesante, desde un punto de vista algorítmico, por los siguientes aspectos:

– Aportan soluciones más eficientes respecto a la mejor solución conocida en torno a un 15% (incluso para valores de m grandes).

– Como consecuencia, cuando se aplican a estas soluciones un procedimiento de Búsqueda Local, se llega a resultados mucho mejores que si se partiera de una solución con elección de centroides totalmente aleatoria (SA). Aunque ocurre en todos los casos, esto es especialmente significativo cuando el número de clusters es mayor.

– Este método, además, da lugar a soluciones diversas. Este aspecto resulta interesante cuando se dispone de tiempo suficiente para repetir el procedimiento varias veces (como en las estrategias GRASP).

La interpretación acerca del desarrollo socioeconómico de los municipios de la región es similar con los nuevos algoritmos de clasificación empleados. Cabe mencionar, no obstante, algunas diferencias en favor de las nuevas técnicas:

– Las nuevas técnicas consiguen una clasificación más equitativa en cuanto a la distribución de los municipios por clusters, por lo que la comparación del desarrollo socioeconómico entre los clusters tiene más sentido.

– El cluster de los municipios más desarrollados económicamente en la nueva clasificación incluye, además de los municipios obtenidos por el algoritmo K-Medias, una serie de municipios con una cierta relevancia en el contexto de la región, cuya actividad económica permite una agrupación con los anteriores.

– La nueva clasificación separa de los antiguos clusters correspondientes a los municipios con menor desarrollo socioeconómico, un grupo de localidades caracterizadas por un nivel extremadamente bajo en la segunda componente (variables

como número de teléfonos y vehículos). Ello justifica su separación del resto en atención a este rasgo específico de su desarrollo socioeconómico.

En definitiva, podemos señalar que la clasificación final de los municipios de la región es más adecuada desde el punto de vista económico, puesto que crea agrupaciones más interpretables en función de rasgos determinados, -como número de licencias comerciales o el nivel y la calidad de sus comunicaciones-, que influyen en el concepto de desarrollo socioeconómico.

Desde un punto de vista práctico, el uso del método que se ha desarrollado y analizado ofrece mejores soluciones y, por tanto, se obtienen clusters o grupos más homogéneos, como en el ejemplo que se ha expuesto en la sección anterior. Esto, obviamente, favorece enormemente los análisis que se realicen, permitiendo sacar conclusiones de los estudios mucho más claras y rigurosas desde el punto de vista estadístico.

Por último, también es evidente que, al obtenerse soluciones más precisas, se optimizan los recursos económicos. Un ejemplo de esto sería que, una vez clasificados los municipios, se distribuyeran fondos para el desarrollo económico según las agrupaciones obtenidas. Si se efectúa un óptimo reparto de municipios, los costes económicos asociados a la toma de decisiones es menor, por lo que un buen método de clasificación redunda en una mayor eficiencia económica. Esto es válido para todos los sectores en los cuales exista un coste económico asociado a la toma de decisiones sobre los elementos que constituyen los clusters.

REFERENCIAS

AL-SULTAN,K.H. (1995). «A Tabu Search Approach to the Clustering Problem». *Pattern. Recogn*. 28, 1.443-1.451.

BABU,G.P. AND MURTY,M.N. (1993). «A Near-Optimal Initial Seed Value Selection in K-means Algorithm using Genetic Algorithms». *Pattern. Recogn. Lett*., 14, 763-769.

BRUCKER, P. (1978). «On the Complexity of Clustering Problems». *Lecture Notes in Economics and Mathematical Systems* 157, 45-54.

CANO,F.J. (1999) «Análisis de clusters o de conglomerados". *I Jornadas de Matemáticas*. Burgos, Octubre 1.999.

DIEHR,G. (1985). «Evaluation of a Branch and Bound Algorithm for Clustering». *SIAM J.Sci.Statist.Comput*., 6, 268-284.

FEO, T. A. AND RESENDE, M. G. C. (1995): «Greedy Randomized Adaptive Search Procedures». *Journal of Global Optimization*. Vol. 2, pp 1-27.

GRÖTSCHEL, M. AND WAKABAYASHI,Y. (1989). «A cutting plane algorithm for a clustering problem». *Mathematical Programming* 45, 59-96.

HANSEN,P. AND MLADENOVIC,N., (1999). «J-Means: A new Local Search Heuristic for Minimun Sum-of-Squares Clustering». *Les Cahiers de GERAD*, G-99-14, Montreal, Canada.

HERRERO, L.C. (1992): «Criterios multivariantes para el estudio del desarrollo económico y la organización del espacio en Castilla y León». Junta de Castilla y León.

HOWARD,R. (1966). «Classifying a Population into Homogeneous Groups». In Lawrence, J.R. (eds.), *Operational Research in the Social Sciences*. Tavistock Publ., London.

JANCEY,R.C. (1966). «Multidimensional Group Analysis». *Australian J. Botany* 14, 127-130.

KLEIN,R.W. AND DUBES,R.C. (1989). «Experiments in Projection and Clustering by Simulated Annealing». *Pattern. Recogn.* 22, 213-220.

KOONTZ,W.L.G., NARENDRA,P.M. AND FUKUNAGA,K., (1975). «A Branch and Bound Clustering Algorithm». *IEEE Trans.Computers* C-24, 908-915

MERLE,O DU., HANSEN,P., JAUMARD,B. AND MLADENOVIC,N. (1997). «An Interior Point Algorithm for Minimun Sum of Squares Clustering». *Les Cahiers de GERAD*, G-97-53, Montreal, Canada.