Un algoritmo incremental para la obtención de particiones con datos mezclados

Aurora Pons-Porrata[1], José Ruiz-Shulcloper[2], Rafael Berlanga-Llavori[3], Yovanis Santiesteban Alganza[1]
[1] Universidad de Oriente, Santiago de Cuba (Cuba)
aurora@app.uo.edu.cu, yosva@csd.uo.edu.cu
[2] Instituto de Cibernética, Matemática y Física (Cuba)
recpat@cidet.icmf.inf.cu
[3]Universitat Jaume I, Castellón (España)
berlanga@lsi.uji.es

Resumen

En este trabajo se introduce un algoritmo incremental eficiente para estructurar un conjunto de datos en conjuntos compactos. Este algoritmo se apoya en algunas propiedades de los conjuntos compactos que son demostradas en el trabajo. El algoritmo propuesto crea una partición única del conjunto de datos, por lo que no depende del orden de presentación de los objetos. La complejidad computacional del algoritmo incremental presentado sigue siendo la misma que la de su variante no incremental y, por tanto, mucho más eficiente que la aplicación reiterada de la variante no incremental. El algoritmo propuesto puede ser utilizado en todas las tareas que requieran del agrupamiento de objetos en compactos y del procesamiento dinámico de la información, tales como, por ejemplo, la organización de la información, navegación, filtrado, ruteo y detección y seguimiento de sucesos en un flujo de documentos, entre otras aplicaciones. Es bueno resaltar que este algoritmo puede utilizarse, además, en problemas donde se deseen agrupar objetos de cualquier naturaleza, descritos por rasgos cuantitativos y cualitativos mezclados, incluso con ausencia de información.

Palabras clave: algoritmos incrementales, algoritmos de agrupamiento.

1. Introducción.

El agrupamiento de datos es uno de los problemas centrales en la Minería de Datos. Entre las técnicas de agrupamiento existentes, el criterio de agrupamiento en conjuntos compactos tiene la propiedad de que la semejanza entre un par de objetos de un mismo grupo es máxima. Los grupos obtenidos por este criterio son relativamente pequeños y densos.

Existen muchas aplicaciones donde este criterio resulta de utilidad: la determinación de zonas perspectivas de ciertos minerales, el análisis de datos en problemas de clasificación supervisada y otros. En estos problemas se procesan conjuntos de datos estáticos, es decir, que no varían.

Existen otros problemas en los que el conjunto de datos se incrementa en tiempo. Por ejemplo, el análisis del comportamiento de las facturas de una empresa, el estudio de las características sociales de los turistas que arriban a un hotel, identificación y seguimiento de sucesos en un flujo de noticias, etc. Para este tipo de aplicaciones se requiere de un algoritmo incremental que obtenga los conjuntos compactos.

En este trabajo se introduce un algoritmo incremental eficiente para estructurar un conjunto de datos en conjuntos compactos. Este algoritmo se apoya en algunas propiedades de los conjuntos compactos que son demostradas en el trabajo.

2. Conceptos y resultados necesarios.

Sea ζ una colección de objetos y S una función de semejanza entre objetos simétrica. Aquí sólo consideraremos este tipo de función de semejanza. Sea, además, β_0 un umbral de semejanza definido por el usuario.

Definición 1. *Se dice que dos objetos O_i y O_j son β_0-semejantes si $S(O_i, O_j) \geq \beta_0$ para todo O_j de la colección de objetos ζ se cumple que $S(O_i, O_j) < \beta_0$ entonces O_i se denomina β_0-aislado.*

Definición 2. *Diremos que $NU \subseteq \zeta$, $NU \neq \emptyset$ es una componente conexa si se cumple que* [1]:

1. $\forall O_i, O_j \in NU,\ O_i \neq O_j,\ \exists O_{i_1}, \ldots, O_{i_q} \in NU\ [O_i = O_{i_1} \wedge O_j = O_{i_q} \wedge \forall p = 1, \ldots, q\text{-}1\ S(O_p, O_{p+1}) \geq \beta_0]$.
2. $\forall O_i \in \zeta\ [O_j \in NU \wedge S(O_i, O_j) \geq \beta_0] \Rightarrow O_i \in NU$.
3. *Todo elemento β_0-aislado es una componente conexa (degenerada).*

Definición 3. *Diremos que $NU \subseteq \zeta$, $NU \neq \emptyset$, es un conjunto compacto si* [1]:

1. $\forall O_j \in \zeta\ [O_i \in NU \wedge \max_{\substack{O_t \in \zeta \\ O_t \neq O_i}} \{S(O_i, O_t)\} = S(O_i, O_j) \geq \beta_0] \Rightarrow O_j \in NU$.
2. $[\max_{\substack{O_i \in \zeta \\ O_i \neq O_p}} \{S(O_p, O_i)\} = S(O_p, O_t) \geq \beta_0 \wedge O_t \in NU] \Rightarrow O_p \in NU$.
3. $|NU|$ *es mínimo.*

Todo elemento β_0-aislado constituye un conjunto compacto (degenerado).

La condición 1 dice que todo objeto de *NU* tiene en *NU* a sus objetos más β_0-semejantes. La condición 2 dice que no existe fuera de *NU* un objeto cuyo objeto más β_0-semejante esté en *NU*. La condición 3 dice que *NU* es el conjunto más pequeño que cumple las condiciones 1 y 2.

El criterio de agrupamiento basado en los conjuntos compactos forma, al igual que el de las componentes conexas, una partición. Esta partición es, además, única para cada conjunto de datos dado.

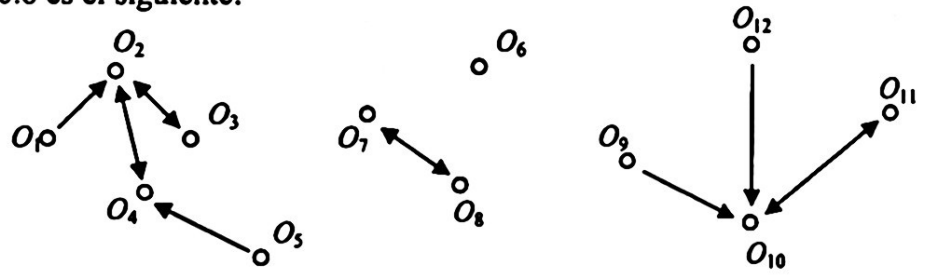
Ejemplo 1. Sean $\zeta=\{O_1,O_2,O_3,O_4,O_5,O_6,O_7,O_8,O_9,O_{10},O_{11},O_{12}\}$ y la matriz de semejanza siguiente, obtenida a partir de una función de semejanza S:

$$MS = \begin{array}{c} \\ O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \\ O_6 \\ O_7 \\ O_8 \\ O_9 \\ O_{10} \\ O_{11} \\ O_{12} \end{array} \begin{pmatrix} 1.0 & 0.85 & 0.75 & 0.78 & 0.63 & 0.70 & 0.51 & 0.38 & 0.43 & 0.27 & 0.22 & 0.13 \\ & 1.0 & 0.90 & 0.90 & 0.76 & 0.62 & 0.58 & 0.45 & 0.56 & 0.38 & 0.36 & 0.27 \\ & & 1.0 & 0.88 & 0.83 & 0.50 & 0.60 & 0.47 & 0.65 & 0.47 & 0.47 & 0.33 \\ & & & 1.0 & 0.85 & 0.58 & 0.68 & 0.55 & 0.63 & 0.45 & 0.42 & 0.30 \\ & & & & 1.0 & 0.48 & 0.73 & 0.65 & 0.80 & 0.65 & 0.58 & 0.50 \\ & & & & & 1.0 & 0.51 & 0.36 & 0.27 & 0.11 & 0.05 & 0.00 \\ & & & & & & 1.0 & .0.83 & 0.66 & 0.58 & 0.48 & 0.51 \\ & & & & & & & 1.0 & 0.66 & 0.63 & 0.53 & 0.61 \\ & & & & & & & & 1.0 & 0.83 & 0.80 & 0.70 \\ & & & & & & & & & 1.0 & 0.88 & 0.85 \\ & & & & & & & & & & 1.0 & 0.80 \\ & & & & & & & & & & & 1.0 \end{pmatrix}$$

Si consideramos $\beta_0 = 0.8$, entonces los conjuntos compactos son los siguientes: $NU_1=\{O_1,O_2,O_3,O_4,O_5\}$, $NU_2=\{O_9,O_{10},O_{11},O_{12}\}$, $NU_3=\{O_7,O_8\}$ y $NU_4=\{O_6\}$

Definición 4. *Sea S una función de semejanza entre objetos. Llamaremos grafo basado en la máxima β_0-semejanza según S, y lo denotaremos máx-S, al grafo orientado $G=(\zeta,E)$ cuyos vértices son los objetos de la colección ζ y existe un arco del vértice O_i al O_j si se cumple que O_j es el objeto más β_0-semejante a O_i.*

Ejemplo 2. El grafo *máx-S* correspondiente a la matriz de semejanza del ejemplo 1 con $\beta_0=0.8$ es el siguiente:



Proposición 1. *El conjunto de todos los conjuntos compactos de ζ coincide con el conjunto de todas las componentes conexas del grafo máx-S asociado a ζ sin tener en cuenta la orientación.*

Demostración. Es inmediato, a partir de la definición 4, que toda componente conexa del grafo *máx-S* asociado a ζ sin tener en cuenta la orientación, es un conjunto compacto de ζ.

Sea $C=\{O_{i_1},O_{i_2},...,O_{i_k}\}$, $k>1$, un conjunto compacto y $G=(\zeta, E)$ el grafo *máx-S* asociado a ζ sin tener en cuenta la orientación.

Supongamos que C no es una componente conexa de G. Por lo tanto, existen $O_{i_j},O_{i_r}\in C$, tales que entre ellos no existe un camino en G.

Entonces, construyamos el conjunto $NU=\{O_{i_j}\}$ y agreguemos a él todos los objetos O_{i_t} de C tales que:

$$\max_{\substack{O_{i_s}\in C \\ O_{i_s}\neq O_{i_j}}} \{S(O_{i_j},O_{i_s})\} = S(O_{i_j},O_{i_t}) \geq \beta_0$$

$$\max_{\substack{O_{i_s}\in C \\ O_{i_s}\neq O_{i_t}}} \{S(O_{i_s},O_{i_t})\} = S(O_{i_j},O_{i_t}) \geq \beta_0$$

Se repite el proceso para cada objeto de C agregado a NU hasta que no se pueda agregar a nadie más.

Por construcción de NU el grafo *máx-S* asociado a NU sin tener en cuenta la orientación es conexo, $O_{i_r}\notin NU$ por la suposición y, además, NU es el conjunto más pequeño que satisface las condiciones 1 y 2 de la definición de conjunto compacto. Como $NU\subset C$, por la condición 3 de la definición de conjunto compacto, C no puede ser un conjunto compacto.

Si C fuera un conjunto unitario, es decir, O_{i_j} es un objeto β_0-aislado, entonces una componente conexa de G sería $\{O_{i_j}\}$, pues no existe ningún arco de O_{i_j} cualquier otro vértice de dicho grafo. ∎

3. Nuevas propiedades de los compactos.

Definición 5. *Sean C un conjunto compacto de una colección de objetos ζ, S una función de semejanza y β_0 un parámetro definido por el usuario. Decimos que un objeto O está conectado con C si existe al menos un $O'\in C$ que cumpla una de las dos condiciones siguientes:*

1. *O' es el objeto más β_0-semejante a O según S.*
2. *O es el objeto más β_0-semejante a O' según S.*

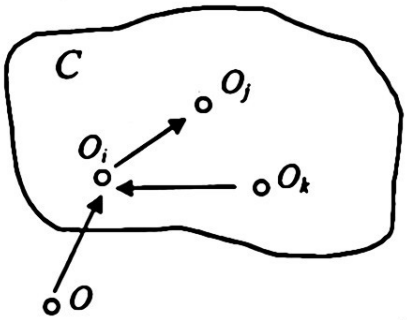
Un objeto O puede conectarse con un conjunto compacto C de varias formas. La figura 1 muestra todos los casos que pueden presentarse. En los casos I y II el objeto O no rompe ningún arco del grafo *máx-S* asociado a C. En los casos III y IV sí rompen arcos.

Sean ζ una colección de objetos, S una función de semejanza, $\wp$ el conjunto todos los conjuntos compactos de ζ, $C\in\wp$ un conjunto compacto, $G_c=(C,U)$ el grafo *máx-S* asociado a C y NG_c el grafo G_c sin tener en cuenta la orientación. Sea, además, O un objeto tal que $O\notin\zeta$.

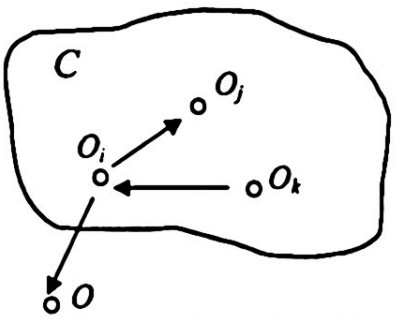
Lema 1. *Si O no está conectado con C, entonces C será un conjunto compacto de* $\zeta \cup \{O\}$.

Demostración. Sea G el grafo *máx-S* asociado a ζ sin tener en cuenta la orientación. Por la proposición 1, NG_c es una componente conexa de G. Si O no está conectado con C, por definición de componente conexa de un grafo, O no puede pertenecer a dicha componente conexa. Por tanto, NG_c será también una componente conexa del grafo *máx-S* asociado a $\zeta \cup \{O\}$ sin tener en cuenta la orientación y, en consecuencia, C será un conjunto compacto de $\zeta \cup \{O\}$. ∎

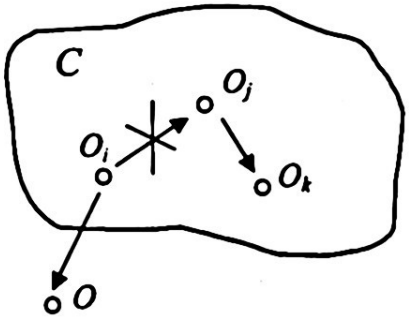
Corolario. *Si O es* β_0*-aislado, entonces el conjunto de todos los conjuntos compactos de* $\zeta \cup \{O\}$ *es* $\wp \cup \{\{O\}\}$.



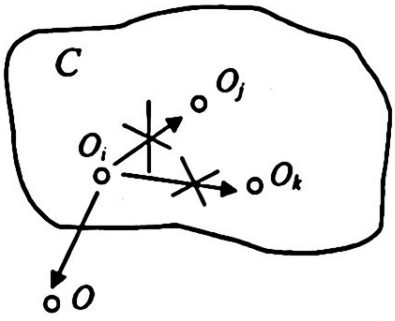
Caso I: O_i es el objeto más β_0-semejante a O.



Caso II: O y O_j son los objetos más β_0-semejantes a O_i.



Caso III: O es el objeto más β_0-semejante a O_i y O_j deja de serlo.



Caso IV: O es el objeto más β_0-semejante a O_i y O_j y O_k dejan de serlo.

Figura 1: Formas de conexión entre un objeto y un conjunto compacto.

Lema 2. *Si O está conectado con C y O no rompió ningún arco de* G_c*, entonces O y todos los objetos de C pertenecerán a un mismo conjunto compacto de* $\zeta \cup \{O\}$.

Demostración. Por la proposición 1, NG_c es una componente conexa del grafo *máx-S* asociado a ζ sin tener en cuenta la orientación. Si O no rompió ningún arco de G_c, entonces NG_c sigue siendo conexa en el grafo *máx-S* asociado a $\zeta \cup \{O\}$ sin tener en

cuenta la orientación. Si O está, además, conectado con C, entonces O pertenece a la misma componente conexa de los objetos de C y, por tanto, O y todos los objetos de C pertenecerán a un mismo conjunto compacto de $\zeta \cup \{O\}$. ∎

Definición 6. *Un punto de articulación de un grafo no orientado es un vértice v tal que cuando removemos v y todos los arcos incidentes en él, la componente conexa de v se divide en dos o más partes* [2].

Sean G' el grafo *máx-S* asociado a $\zeta \cup \{O\}$ y H, el subconjunto de objetos de C que pertenecen a la componente conexa de O en G' sin tener en cuenta la orientación.

Lema 3. *Si O está conectado con C y O rompe uno y sólo un arco de G_c (caso III), entonces si el conjunto $C \setminus H \neq \emptyset$, $C \setminus H$ es un conjunto compacto de $\zeta \cup \{O\}$.*

Demostración. Sea $O_i \in C$ el objeto al cual O es el más β_0-semejante y tal que se rompe su único arco a que incide al exterior. Si O_i no es un punto de articulación en NG_c, entonces al eliminar a el grafo sigue siendo conexo y, por tanto, $H=C$.

Si O_i es un punto de articulación en NG_c, entonces el grafo $G''=(C, U \setminus \{a\})$ sin tener en cuenta la orientación no es conexo. Si $C \setminus H \neq \emptyset$, entre cualesquiera dos objetos de $C \setminus H$ existe un camino en NG_c porque, de lo contrario, C no sería un conjunto compacto. Como O_i es un punto de articulación en NG_c, no existe en H ningún otro objeto conectado con $C \setminus H$.

Por tanto, $C \setminus H$ es una componente conexa en G' sin tener en cuenta la orientación, es decir, $C \setminus H$ es un conjunto compacto en $\zeta \cup \{O\}$. ∎

Teorema. *El conjunto de todos los conjuntos compactos de $\zeta \cup \{O\}$ es:*

$$\wp' = [\wp \setminus (N \cup R \cup M)] \cup \left[\bigcup_{i=1}^{|R|} \{R_i \setminus H_i\}\right] \cup \left[\{O\} \cup \bigcup_{i=1}^{|N|} N_i \cup \bigcup_{i=1}^{|R|} H_i \cup \bigcup_{i=1}^{|M|} J_i\right] \cup K$$

donde:

N es el conjunto de todos los conjuntos compactos de $\wp$ que están conectados con O y donde no se rompieron arcos y $N_i \in N$.

R es el conjunto de todos los conjuntos compactos de $\wp$ que están conectados con O y donde se rompió uno y solo un arco y $R_i \in R$.

M es el conjunto de todos los conjuntos compactos de $\wp$ que están conectados con O y donde se rompió más de un arco y $M_i \in M$.

H_i es el conjunto de todos los objetos de R_i que pertenecen a la misma componente conexa de O en G' sin tener en cuenta la orientación.

J_i es el conjunto de todos los objetos de M_i que pertenecen a la misma componente conexa de O en G' sin tener en cuenta la orientación.

K es el conjunto de todos los conjuntos compactos que pueden formarse en cada $M_i \setminus J_i$, $i=1,...,|M|$.

Demostración. Como dijimos anteriormente, un nuevo objeto $O \notin \zeta$ puede conectarse sólo de cuatro formas diferentes con los conjuntos $C \in \wp$. Es importante

observar que estos 4 casos no son excluyentes, incluso pudieran ocurrir todos a la vez varias veces cada uno. Por esta razón no es posible, ni necesario, analizar todas las posibles combinaciones. La ocurrencia de cada caso conllevará a una determinada modificación en el conjunto $\wp$.

La demostración de este teorema se apoya en los lemas 1, 2 y 3 y en la proposición 1.

A $\wp'$ pertenecerán por el lema 1, todos los conjuntos compactos de $[\wp \setminus (N \cup R \cup M)]$, pues con ellos el nuevo objeto O no está conectado.

Además, por el lema 3, dado un conjunto compacto R_i, el conjunto $R_i \setminus H_i$ mantiene su propiedad de ser compacto si $H_i \neq \varnothing$ y, por tanto, los conjuntos de $\left[\bigcup_{i=1}^{|R|}\{R_i \setminus H_i\}\right]$ también pertenecen a $\wp'$.

La componente conexa de G' a la que pertenece O $\left(\left[\{O\} \cup \bigcup_{i=1}^{|N|} N_i \cup \bigcup_{i=1}^{|R|} H_i \cup \bigcup_{i=1}^{|M|} J_i\right]\right)$ es por la proposición 1 un conjunto compacto. Note que a ella pertenecen todos los conjuntos compactos de N (según lema 2).

No es difícil ver que:

$$\zeta \setminus \left\{\left[\wp \setminus (N \cup R \cup M)\right] \cup \left[\bigcup_{i=1}^{|R|}\{R_i \setminus H_i\}\right] \cup \left[\{O\} \cup \bigcup_{i=1}^{|N|} N_i \cup \bigcup_{i=1}^{|R|} H_i \cup \bigcup_{i=1}^{|M|} J_i\right]\right\}$$

puede ser diferente del vacío. Esto es debido a la posibilidad de que O provoque la ruptura de más de un arco en al menos un conjunto compacto M_i. Es inmediato que no podemos afirmar que $M_i \setminus J_i$ es un conjunto compacto. Luego, a partir del conjunto $\left[\bigcup_{i=1}^{|M|}\{M_i \setminus J_i\}\right]$ se hace necesario construir el conjunto K de todos sus posibles conjuntos compactos. Observe que el cardinal de este conjunto es mucho más pequeño que el de $\zeta \cup \{O\}$. Con la inclusión de los conjuntos compactos existentes en K se construye completamente $\wp'$. ∎

4. Algoritmo compacto incremental.

Paso 1. Cada vez que se presenta un nuevo objeto O se calcula la similaridad con todos los objetos de los conjuntos compactos existentes.

Paso 2. Se seleccionan los objetos que son más β_0-semejantes a O y aquellos a los que O es el más β_0-semejante. Si no existen tales objetos, en virtud del corolario del lema 1, el conjunto $\{O\}$ es un nuevo conjunto compacto y terminar.

Paso 3. Los objetos seleccionados en el paso 2 pertenecen a conjuntos compactos de N, R ó M, luego:

a) Se crea un nuevo conjunto compacto formado por O y todos los objetos que pertenecen a la misma componente conexa de O en el grafo *máx-S* de $\zeta \cup \{O\}$ sin tener en cuenta la orientación. Note que esta componente conexa está formada por todos los objetos de los conjuntos de N y una parte de los objetos de cada uno de los conjuntos de R y M.
b) Estos objetos son eliminados de los conjuntos compactos a los que pertenecían.
c) Los conjuntos compactos que quedan vacíos se eliminan.
d) Las partes que quedaron de los conjuntos de R siguen siendo conjuntos compactos, en virtud del lema 3.

Paso 4. Puede suceder que al eliminar un objeto (que pertenece a la componente conexa de O) del conjunto compacto M_i de M al que pertenecía provoque que se quede inconexo y, por tanto, deje de ser un conjunto compacto. En cada parte de los conjuntos M_i que quedaron hacer:
a) Seleccionar un objeto.
b) Construir su componente conexa (por la proposición 1 sus objetos forman un conjunto compacto).
c) Eliminar los objetos de esa componente conexa de la parte de M_i.
d) Mientras queden objetos ir al paso 4a).

5. Análisis de la complejidad computacional.

En este algoritmo cada objeto O tiene asociado el grupo al que pertenece y tres informaciones: $A(O)$ que contiene el o los objetos más β_0-semejantes a O, es decir, $\{O' \in \zeta \,/\, S(O,O') = \max_{\substack{O'' \in \zeta \\ O'' \neq O}} \{S(O,O'')\} \wedge S(O,O') \geq \beta_0\}$, $SimilMax(O)$, el valor de esa máxima similitud y $De(O)$ que contiene los objetos a los que O es el más β_0-semejante, es decir, $\{O' \in \zeta \,/\, O \in A(O')\}$.

La complejidad computacional de este algoritmo es $O(n^2)$, pues en el paso 1 cada objeto se compara con todos los existentes. El paso 3 conforma la componente conexa a la que el nuevo objeto pertenece. Este paso tiene complejidad $O(e)$, pues se trabaja con listas de adyacencia [3]. Aquí e es la cantidad de aristas del grafo. Aunque desde el punto de vista teórico e es del orden n^2, en la práctica los grafos de máxima semejanza no son grafos completos y lo que ocurre con mucha frecuencia es que la cantidad de objetos más semejantes a uno dado es 1. Por eso, en nuestro caso, hemos estimado experimentalmente que $e=cn$, donde c es la cantidad máxima de objetos más semejantes a uno dado. En el paso 4 se reconstruyen las componentes conexas de los grupos que quedaron inconexos, lo cual, es $O(n+e)$.

Con relación a la complejidad espacial, nuevamente si tenemos en cuenta lo anterior, la complejidad sería $O(n)$, pues para cada objeto sólo tendríamos que almacenar el valor de su máxima semejanza y la lista de los objetos más β_0-semejantes, cuyo cardinal nuevamente podría aproximarse por una constante.

6. Experimentación.

La efectividad del algoritmo incremental propuesto ha sido evaluada usando dos conjuntos de datos.

Conjunto de datos 1.

El primer conjunto de datos que usamos es una porción de *flag* tomado de [4]. Este conjunto está formado por 50 banderas de países descritas por 23 rasgos: 8 numéricos, 12 booleanas y 3 cualitativas *k*-valentes. Ejemplos de estos rasgos son: número de barras horizontales y verticales, colores presentes, si tienen o no triángulos, número de diagonales y estrellas, etc. Se utilizaron los siguientes criterios de comparación para cada tipo de variable:

$$C(x_i(O), x_i(O')) = \begin{cases} 1 & si\ x_i(O) = x_i(O') \\ 0 & en\ otro\ caso \end{cases}, \text{ para las variables booleanas}$$

$$C(x_i(O), x_i(O')) = \begin{cases} 1 & si\ x_i(O), x_i(O') \in A_p \\ 0 & en\ otro\ caso \end{cases}, \text{ para las variables } k\text{-valentes.}$$

$$C(x_i(O), x_i(O')) = \begin{cases} 1 & si\ |x_i(O), x_i(O')| \le \varepsilon \\ 0 & en\ otro\ caso \end{cases}, \text{ para las variables numéricas}$$

A_p son subconjuntos de valores del rasgo x_i y ε un umbral. Ambos parámetros son dados por el usuario.

Como función de semejanza utilizamos la suma ponderada de los criterios de comparación por rasgos y como umbral de semejanza =0.9.

Para el subconjunto de datos formado por 48 banderas se obtuvieron los siguientes 26 agrupamientos:

G_1={Bulgaria,Venezuela}, G_2={Gambia,Costa Rica}, G_3={Uruguay}, G_4={Fiji}, G_5={Bhutan},G_6={Zimbabwe},G_7={Uganda},G_8={Guinea_Bissau},G_9={Panama}, G_{10}={Somalia,Mauritania}, G_{11}={NorthKorea},G_{12}={Greece}, G_{13}={Japan,Brazil}, G_{14}={Ethiopia,Afghanistan,Colombia,Ecuador}, G_{15}={Cameroon,Mexico,Belgium, Mali,Chad,Andorra,Senegal},G_{16}={Niger,India},G_{17}={Haiti,Canada},G_{18}={Belize}, G_{19}={Cuba,Mozambique,Puerto_Rico},G_{20}={Argentina,Egypt,Austria},G_{21}={USA} G_{22}={Australia, Tuvalu, New_Zealand}, G_{23}={South_Africa}, G_{24}={Malta}, G_{25}={Bahamas}, G_{26}={Sweden,Dominican_Republic}.

Al incorporarse la bandera de Francia se mantuvieron los mismos 26 grupos, incorporándose Francia al grupo 15:

G_{15}={France, Cameroon, Mexico, Belgium, Mali, Chad, Andorra, Senegal}

Todas las banderas de este grupo se caracterizan por tener 3 franjas verticales de 3 colores diferentes donde uno de ellos es el rojo.

Al incorporarse, finalmente, la bandera de España se forman 26 grupos pero, en este caso, se particiona el grupo G_{14} y se fusionan una parte de este grupo con el grupo G_{20}, quedando de esta forma:

G_{14}={Colombia, Ecuador},

G_{20}={Spain, Ethiopia, Afghanistan, Argentina, Egypt, Austria}

Atendiendo a los rasgos considerados para la descripción, la bandera de España es ahora la más parecida a la de Etiopía, por lo que las banderas de Colombia y Ecuador dejan de serlo. Por su parte la bandera más parecida a la de Afganistán es la de Etiopía, lo que justifica que éstas dos últimas se mantengan unidas en el nuevo agrupamiento G_{20}.

S(Etiopía,Colombia)=0.9828 S(Etiopía,Ecuador)=0.9828
S(Etiopía,Afghanistan)=0.9814 S(Etiopía, España)=0.9885

Esto provoca la partición del Grupo 14. Por otra parte, la bandera de Austria es la más parecida a la de España, S(Austria,España)=0.9985; lo que provoca la fusión de estos grupos al incorporarse en él la bandera de España.

Como ya hemos visto en el ejemplo, al incorporarse una nueva bandera al conjunto de datos, los grupos pueden cambiar en su composición y también pueden hacerlo en número.

Conjunto de datos 2.

El segundo conjunto de datos con el que experimentamos es una colección de 526 artículos de periódicos publicados en la sección de noticias internacionales del periódico español "El País" durante el mes de junio de 1999. En este tipo de problema, es claro el carácter dinámico de la colección de datos (en este caso, de documentos) y, por tanto, imprescindible la utilización de un algoritmo incremental que identifique los sucesos que acontecen. Los documentos, denotados aquí por el título del artículo, fueron representados en función de la frecuencia relativa de sus términos y de sus ocurrencias temporales [5]. A continuación mostramos algunos de los grupos formados utilizando como umbral de semejanza 0.25:

Grupo #1:
Kofi Annan cree que es prematuro ponerse a dar saltos de alegría.
La secuencia de la pacificación.
La ONU ve imposible el regreso de los refugiados antes del otoño.
Aznar dice que espera ver un Kosovo liberado de la represión.
Clinton deja abiertas todas las opciones.
La OTAN consigue imponer sus principales objetivos.
El plan para la paz en Yugoslavia y el regreso de los refugiados.
Siete discrepancias iniciales.
Los refugiados carecen de papeles para regresar y recuperar sus casas.
Aznar dice que el próximo paso debe ser colaborar con el Tribunal Internacional.
Proyecto de resolución de la ONU.
Solana se niega a recibir al líder de la guerrilla kosovar en su visita privada a la OTAN.
Aznar pide ante Rugova que no se creen falsas ilusiones sobre la independencia de Kosovo.

Grupo #2:
Amnistía Internacional destaca el 'caso Pinochet' como hito de los derechos humanos.
Baltasar Garzón amplía en 34 casos la acusación por torturas contra el ex dictador Pinochet.
Documentos oficiales secretos chilenos pueban la existencia de la Operación Cóndor.
La desaparición de personas era una 'industria', según un informe.
'Embustes de los señores de España'.
Repercusión política del caso.
El Supremo dice que la actitud de Fungairiño en el 'caso Pinochet' puede ser 'criticable'.

El primer grupo contiene las noticias sobre los acontecimientos de la guerra de Kosovo y el segundo, las que abordan el caso Pinochet. Observe que, incluso al analizar el contenido de los artículos por su título, y teniendo en cuenta que el umbral seleccionado es bajo, los grupos formados presentan una alta coherencia temática.

7. Conclusiones.

En este trabajo se presenta un nuevo algoritmo incremental de agrupamiento que crea una partición en conjuntos compactos de la colección de objetos.

La variante no incremental de este criterio de agrupamiento necesitaba almacenar la matriz de similaridad (de orden n^2) de los objetos de la colección lo que, sin dudas, constituye una gran limitación cuando se desea trabajar con colecciones dinámicas o grandes de objetos. La variante incremental desarrollada no necesita almacenar esta matriz.

Otra ventaja del algoritmo propuesto es que permite encontrar grupos con formas arbitrarias en oposición a los algoritmos incrementales como *K-Means* [6] y *Single-Pass* [7] que utilizan medidas centrales para generar los grupos, y como consecuencia, se restringe a los grupos a tener formas esféricas o elipsoidales.

Además el agrupamiento obtenido por el algoritmo propuesto es único, es decir, no depende del orden de presentación de los objetos, a diferencia del *K-Means* y el *Single-Pass*, que pueden producir diferentes agrupamientos al cambiar el orden de entrada de los objetos. Esta unicidad hay que entenderla de la siguiente manera: dados m objetos, la estructuración en conjuntos compactos es única, independientemente del orden en que los m objetos sean considerados. Es claro que en una población dinámica, la llegada de un nuevo objeto puede producir nuevas estructuraciones, pero de poblaciones diferentes. Sin embargo, en el caso de la estructuración en compactos, una vez que ha sido considerada una población determinada, los objetos que la conforman pudieron haber arribado en cualquier orden produciendo la misma estructuración.

El algoritmo propuesto no necesita fijar a priori el número de grupos a obtener, es decir, es aplicable a problemas de clasificación no supervisada libre.

El criterio de agrupamiento basado en los conjuntos compactos forma grupos disjuntos y más cohesionados y pequeños que los formados por las componentes conexas basadas solamente en la β_0-semejanza. Este último criterio es el utilizado en el algoritmo incremental *GLC* [8].

La complejidad computacional del algoritmo incremental presentado sigue siendo la misma que la de su variante no incremental, lo que es muy superior a la aplicación reiterada del algoritmo no incremental.

El algoritmo propuesto puede ser utilizado en todas las tareas que requieran del agrupamiento de objetos en compactos y del procesamiento dinámico de la información, tales como la organización de la información, navegación, filtrado, ruteo y detección y seguimiento de sucesos en un flujo de documentos, el análisis del comportamiento de las facturas de una empresa, el estudio de las características sociales de los turistas que arriban a un hotel, entre otras aplicaciones.

Es bueno resaltar que este algoritmo puede utilizarse, además, en problemas donde se deseen agrupar objetos de cualquier naturaleza, descritos por rasgos cuantitativos y cualitativos mezclados, incluso con ausencia de información.

Referencias

[1] Martínez-Trinidad, J. F., Ruiz-Shulcloper J., Lazo-Cortés, M. (2000). Structuralization of Universes, *Fuzzy Sets and Systems*, Vol. 112 (3), 485-500.

[2] Aho, A. V. , Hopcroft, J. E., Ullman, J. D. (1983). *Data Structures and Algorithms*, Addison-Wesley Publishing Company.

[3] Horowitz, E., Sahni, S. (1975). *Fundamentals of Data Structures*, Computer Science Press, Woodland Hills, California.

[4] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/

[5] Pons-Porrata, A., Berlanga-Llavori, R., Ruiz-Shulcloper, J. (2002). Temporal-Semantic Clustering of Newspaper Articles for Event Detection, *Pattern Recognition in Information Systems*, Eds. José M. Iñesta y Luisa Micó, ICEIS Press, 104-113.

[6] Larsen, B., Aone, C. (1999). Fast and Effective Text Mining Using Linear-time Document Clustering, *in KDD'99*, San Diego, California, 16-22.

[7] Hill, D. R. (1968). A vector clustering technique, *in Samuelson (ed.). Mechanized Information Storage, Retrieval and Dissemination*, North-Holland, Amsterdam.

[8] Sánchez-Díaz, G. (2001). *Desarrollo de algoritmos para el agrupamiento grandes volúmenes de datos mezclados*, Tesis doctoral, Centro Investigación en Computación, México.