Murillo F., Alex; Arroyo P., Guillermo. *Clasificación por particiones utilizando medidas difusas.* **Tecnología en marcha.** Vol. 12, no. especial.

CLASIFICACION POR PARTICIONES UTILIZANDO MEDIDAS DIFUSAS

Alex Murillo F.*
Guillermo Arroyo P.**

Este artículo presenta una propuesta de clasificación de objetos, utilizando el algoritmo de clasificación jerárquica aglomerativa como base. Se define el concepto de conjunto difuso extendido, para establecer un control difuso en cada iteración del algoritmo antes mencionado. Además, para distribuir una asignación de atributos a los objetos de cada clase de la partición establecida, se utiliza un proceso de transformación que convierte la clase en una relación disfuncional.

INTRODUCCION

En este trabajo se presenta un modelo de solución para el problema de confeccionar horarios de profesores y secciones. Este problema[3] se ubica dentro de la categoría de los problemas NP, pues una solución se verifica en tiempo (n^2+m^2), donde n es el número de profesores y m el número de secciones.

Los enfoques para solucionar este problema, en su gran mayoría, son heurísticos, la simulación en estos casos podría ser de gran ayuda; sin embargo, en este artículo se pretende enmarcar el problema desde otro ámbito, tratando de particionar el problema en subproblemas, que puedan ser resueltos en una forma más eficiente. El problema de unir las soluciones particulares para obtener una solución del problema inicial no se toca en este artículo, pero sí se realiza un estudio muy minucioso de cómo particionar el problema, y para ello se utiliza una variante del método de clasificación jerárquica aglomerativa[2].

* Escuela de Matemática, Universidad de Costa Rica
** Departamento de Física, Universidad Nacional.

Descripción del Problema

Sea un grafo bipartito G = <V,E>, donde $V = N \cup M$, $E = N \times M$, m es el número de elementos de M y n es el número de elementos de N, ver Figura 1.

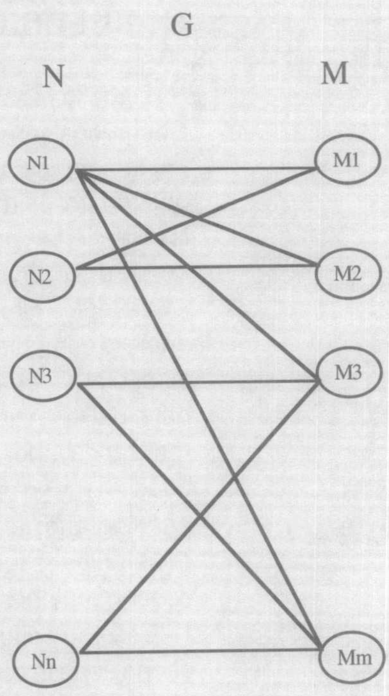

FIGURA 1. Grafo Bipartito G.

A cada elemento de los conjuntos M y N se le asocia una matriz de s x t, un arco del grafo G relaciona un elemento de N con un elemento de M de la siguiente forma:

$$Ni \overset{e}{-\,-\,-\,-\,-\,-\,-\,-} Mj$$

donde el peso del arco e es el número de elementos por asignar tanto en la matriz de Ni como en la matriz de Mj, que son del mismo tamaño, por ejemplo para e = 4, una posible distribución de las matrices Ni y Mj se detalla en el Cuadro 1, donde las coordenadas ocupadas por Ni y Mj son las mismas, esto quiere decir que en la matriz Mj se establecen los elementos con los cuales están relacionados en el conjunto N y en la matriz Ni se establecen los elementos con que se están relacionando en el conjunto M, el tamaño sxt de la matriz asignada a Ni y a Mj es fijo.

CUADRO 1. Asignación del Arco Ni - - - - - Mj

Matriz de Ni

Mj				
Mj			Mj	
			Mj	

Matriz de Mj

Ni				
Ni			Ni	
			Ni	

En el segundo aparte, se realiza una extensión de la teoría de conjuntos difusos, que será utilizada en la variante del método de clasificación jerárquica aglomerativa. Esta extensión consiste en cambiar el universo X de un conjunto difuso, a un subconjunto Y de partes de X (P (X)) y describir la transformación de un conjunto difuso en X a un conjunto difuso en el nuevo universo Y.

Más adelante, se describe el método de clasificación jerárquica aglomerativa, y cómo se utiliza la extensión descrita en el segundo aparte, para modificar el método de clasificación.

Finalmente, se realiza un análisis de la necesidad de crear una relación disfuncional, tratando a cada clase como un rectángulo de la relación, y se sugiere un algoritmo para la asignación, dada una relación disfuncional.

DEFINICION DE CONJUNTO DIFUSO EXTENDIDO

En la siguiente definición se describe la extensión de un conjunto difuso.

Definición 1. Considere un universo finito X, y sea

$$\tilde{A} = \{ (x, \mu_{\tilde{A}}(x)) \,/\, x \in X\}$$

un conjunto difuso en X. Un conjunto difuso extendido en X es un conjunto difuso en Y $\subseteq$ **P**(X), y se denota por $\tilde{A}^e$. Esto es:

$$\tilde{A}^e = \{ (y, \mu_{\tilde{A}}e(y)) \,/\, y \in Y\}$$

donde:

$$\mu_{\tilde{A}}e(y) = 1, \text{ si } \Sigma_{(x \in y)}\, \mu_{\tilde{A}}(x) \quad 1$$
$$\mu_{\tilde{A}}e(y) = \Sigma_{(x \in y)}\, \mu_{\tilde{A}}(x), \text{ en otro caso.}$$

Ejemplo 1. Se denota la clausura de 100 por [100] ("los salarios cercanos a los ¢ 100 000"), además, el álgebra respectiva es [100] + [200] = [300].

Se define un universo X de la siguiente forma:

X={ [100], [200], [300], [400], [500]}.

Sea Ã = "Ingreso alto", un conjunto difuso, esto es:

Ã = 0,2/[100] + 0,4/[200] + 0,6/[300] + 0,8/[400] + 1/[500]

entonces, si

$$Y = \{ \{[100]\}, \{[200]\}, \{[300]\}, \{[400]\}, \{[500]\}, \{[100], [200]\}, \{[100], [300]\}, \{[200], [300], [400]\} \}$$

se tiene

$$\tilde{A}^e = 0{,}2/[100] + 0{,}4/[200] + 0{,}6/[300] + 0{,}8/[400] + 1/[500] + 0{,}6/\{[100],[200]\} + 0{,}8/\{[100],[300]\} + 1/\{[200],[300],[400]\}$$

$\tilde{A}^e$ = "Ingreso del núcleo familiar alto",

donde {[100],[200]} significa que en un núcleo familiar hay una persona que tiene dos trabajos, en uno de los cuales gana cerca de ¢100 000 y en el otro recibe un salario cercano a los ¢200 000; por lo que el núcleo familiar tendría un ingreso total cercano a los ¢300 000. Otra posible interpretación de {[100],[200]} es que en el núcleo familiar, un integrante tiene un ingreso cercano a los ¢200 000 y otro miembro tiene un ingreso cercano a los ¢100 000, por lo que el núcleo familiar tiene un ingreso total cercano a los ¢300 000.

Observación. Si $\tilde{A}^e$ es un conjunto difuso extendido en X, entonces, es un conjunto difuso en el universo $Y \subseteq P(X)$.

Definición 2. Sea $\tilde{A}^e$ un conjunto difuso extendido en X. El conjunto de nivel α es:

$$A^e_\alpha = \{ y \in Y \subseteq \mathbf{P}(X) / \mu_{\tilde{A}}{}^e(y) \quad \alpha \}.$$

Ejemplo 2. Sea el universo X= {A, B, C, D, ..., K} y considere el conjunto difuso

$$\tilde{N} = 0{,}17/A + 0{,}17/B + 0{,}17/C + 0{,}17/D + 0{,}17/E + 0{,}17/G.$$

Sea el nuevo universo Y una partición de X, por ejemplo:

$$Y = \{\{A,B\}, \{C,D\}, \{E\}, \{F,H\}, \{G\}, \{I,J\}, \{K\}\},$$

entonces,

$$\tilde{N}^e = 0{,}34/\{A,B\} + 0{,}34/\{C,D\} + 0{,}17/\{E\} + 0{,}17/\{G\}.$$

Si $\alpha = 0{,}3$, entonces

$$N^e_{0,3} = \{ \{A,B\}, \{C,D\} \}.$$

CLASIFICACION JERARQUICA UTILIZANDO EXTENSION DIFUSA

Descripción del método

El algoritmo de clasificación jerárquica aglomerativa[2], construye en cada iteración, a partir de grupos unitarios, una jerarquía indexada, de acuerdo con un criterio de agregación, el cual usa un concepto de disimilitud adecuado a los objetos por clasificar. Esto quiere decir que si se tiene un conjunto M de objetos, el número de grupos unitarios es $|M|$ como estado inicial del algoritmo; luego de la primera iteración, los objetos se distribuyen en uno o más grupos, donde el número de grupos es menor que $|M|$.

En el problema inicial descrito en la introducción, los Ni están relacionados con los Mj, se supone que los Mj son los objetos por clasificar, esto implica, que un Ni está relacionado con todos los Mj en alguna medida de [0,1], por lo que se puede decir que para una clasificación dada, los $\tilde{N}_i^e$ son conjuntos difusos extendidos sobre el universo M, si un Ni no está relacionado con un Mj, entonces $\mu_{\tilde{N}i}e(Mj) = 0$.

Sea w el porcentaje de elementos de N que están relacionados en un α por ciento en una sola clase, que se le puede llamar ´calidad de la clasificación aproximada a un nivel α´, de acuerdo con esto se define al valor w como:

$$w = |\{Ni^e_\alpha / Ni^e_\alpha \quad \phi, i:=1,\dots,n\}| / |N| \qquad (1)$$

donde Ni^e_α es el conjunto de nivel de $\tilde{N}i$.

Ejemplo 3. En el Cuadro 2, se describe una relación de variables con objetos, las filas representan al conjunto N de variables, las columnas representan al conjunto M de objetos del ejemplo 2. Para realizar la clasificación se utilizó el índice de disimilitud de Chebychev[2], este es:

$$d(x_i,x_j) = \max_{k=1..19} |x_{ki} - x_{kj}|$$

donde $x_i, y_j \in \{A,B,C, ... K\}$.

Por ejemplo en el Cuadro 2, se tiene que:

$d(A,B) = 0{,}33$, $d(A,C) = 0{,}5$.

La agregación es vecino más cercano de Jardine y Sibson[2], y se define como:

$J_1(h1,h2) = \min\{d(x_i,x_j)\}$,
$x_i \in h1$, $x_j \in h2$
para todo $h1, h2 \in P(M)$.

Los datos se tomaron de un ejemplo real.

Condición Inicial

De acuerdo con el Cuadro 2, los Ni varían de 1 a 19 y los Mi varían de A a K; en este caso los Ñi son los conjuntos difusos sobre el universo M, además, los grupos son unitarios, de tal forma que los conjuntos difusos Ñi son los siguientes:

$\tilde{N}_1 =$ 0,17/A + 0,17/B + 0,17/C + 0,17/D+ 0,17/E + 0,17/G
$\tilde{N}_2 =$ 0,33/I + 0,33/J + 0,33/H
$\tilde{N}_3 =$ 0,5/F + 0,5/H
$\tilde{N}_4 =$ 0,17/F + 0,17/G + 0,17/H + 0,17/I + 0,17/J + 0,17/K
$\tilde{N}_{18} =$ 0,09/A + 0,09/B + 0,09/C +0,09/D + 0,09/E + 0,09/F + 0,09/G + 0,09/H + 0,09/I + 0,09/J + 0,09/K
$\tilde{N}_{19} =$ 0,33/B + 0,33/C + 0,33/D.

Iteración 15

En la iteración número 15 del algoritmo, la clasificación establece la siguiente partición del universo:

M^{15}={{A,B},{C,D},{E},{F,H},{G},{I,J},{K}},

CUADRO 2. Datos del Problema

	A	B	C	D	E	F	G	H	I	J	K
1	0,17	0,17	0,17	0,17	0,17		0,17				
2									0,33	0,33	0,33
3						0,5		0,5			
4						0,17	0,17	0,17	0,17	0,17	0,17
5									0,5	0,5	
6						0,33	0,33	0,33			
7	0,17	0,17	0,17	0,17	0,17						0,17
8	0,5	0,5									
9							0,33	0,33	0,33		
10			0,2	0,2					0,2	0,2	0,2
11							0,25	0,25		0,25	0,25
12	0,2	0,2	0,2		0,2	0,2					
13					0,1	0,15	0,15	0,15	0,15	0,15	0,15
14	0,25	0,25	0,25	0,25							
15			0,1			0,15	0,15	0,15	0,15	0,15	0,15
16	0,25	0,25		0,25	0,25						
17	0,13	0,13	0,13	0,13	0,13	0,06	0,06	0,06	0,06	0,06	0,06
18	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09
19		0,33	0,33	0,33							

los conjuntos difusos extendidos $\tilde{N}^e_i$ son:

$\tilde{N}^e_1 = 0{,}33/\{A,B\} + 0{,}33/\{C,D\} + 0{,}17/\{E\} + 0{,}17/\{G\}$
$\tilde{N}^e_2 = 0{,}33/\{F,H\} + 0{,}66/\{I,J\}$
$\tilde{N}^e_3 = 1/\{F,H\}$
....
$\tilde{N}^e_{18} = 0{,}18/\{A,B\} + 0{,}18/\{C,D\} + 0{,}09/\{E\} + 0{,}18/\{F,H\} + 0{,}09/\{G\} + 0{,}18/\{I,J\} + 0{,}09/\{K\}$
$\tilde{N}^e_{19} = 0{,}33/\{A,B\} + 0{,}66/\{C,D\}$.

Para esta iteración si $\alpha = 0{,}6$, entonces $w = 0{,}36$.

Iteración 18

El nuevo universo M^{18} de la iteración 18 es el siguiente:

$M^{18} = \{ \{A,B\},\{C,D,E,G,K\},\{F,H\}, \{I,J\}\}$

Los conjuntos difusos extendidos $\tilde{N}^e_i$ son:

$\tilde{N}^e_1 = 0{,}34/\{A,B\}+0{,}68/\{C,D,E,G,K\}$
$\tilde{N}^e_2 = 0{,}33/\{C,D,E,G,K\}+0{,}66/\{I,J\}$
$\tilde{N}^e_3 = 0{,}5/\{C,D,E,G,K\}+0{,}5/\{I,J\}$
....
$\tilde{N}^e_{18} = 0{,}18/\{A,B\}+0{,}45/\{C,D,E,G,K\} + 0{,}18/\{F,G\}+0{,}18/\{I,J\}$
$\tilde{N}^e_{19} = 0{,}33/\{A,B\} + 0{,}66/\{C,D,E,G,K\}$.

Evaluación de Iteración 18

Se desea realizar una evaluación sobre la clasificación que se establece en la iteración 18, y fijamos un $\alpha = 0{,}6$ para el conjunto de nivel $N^e_{i\alpha}$, entonces, al realizar el cálculo de w, se deduce que: $w = 0{,}47$, por la ecuación descrita en (1).

Iteración 19

En la Iteración 19 el nuevo universo M^{19} es el siguiente:

$M^{19} = \{\{A,B,C,D,E,G,K\}, \{F,G\},\{I,J\}\}$,

entonces los conjuntos difusos $\tilde{N}^e_i$ son:

$\tilde{N}^e_1 = 1/\{A,B,C,D,E,G,K\}$
$\tilde{N}^e_2 = 0{,}33/\{A,B,C,D,E,G,K\}+0{,}66/\{I,J\}$
$\tilde{N}^e_3 = 0{,}5/\{A,B,C,D,E,G,K\}+0{,}5/\{I,J\}$
$\tilde{N}^e_4 = 0{,}34/\{A,B,C,D,E,G,K\}+0{,}34/\{F,H\}+0{,}34/\{I,J\}$
$\tilde{N}^e_5 = 1/\{I,J\}$
$\tilde{N}^e_6 = 0{,}33/\{A,B,C,D,E,G,K\}+0{,}66/\{F,H\}$
$\tilde{N}^e_7 = 1/\{A,B,C,D,E,G,K\}$
$\tilde{N}^e_8 = 1/\{A,B,C,D,E,G,K\}$
$\tilde{N}^e_9 = 0{,}33/\{A,B,C,D,E,G,K\}+0{,}66/\{F,H\}$
$\tilde{N}^e_{10} = 0{,}66/\{A,B,C,D,E,G,K\} + 0{,}4/\{I,J\}$
$\tilde{N}^e_{11} = 0{,}5/\{A,B,C,D,E,G,K\}+0{,}25/\{F,G\}+0{,}25/\{I,J\}$
$\tilde{N}^e_{12} = 0{,}8/\{A,B,C,D,E,G,K\}+0{,}2/\{F,H\}$
$\tilde{N}^e_{13} = 0{,}4/\{A,B,C,D,E,G,K\}+0{,}3/\{F,H\}+0{,}3/\{I,J\}$
$\tilde{N}^e_{14} = 1/\{A,B,C,D,E,G,K\}$
$\tilde{N}^e_{15} = 0{,}4/\{A,B,C,D,E,G,K\}+0{,}3/\{F,H\}+0{,}3/\{I,J\}$
$\tilde{N}^e_{16} = 1/\{A,B,C,D,E,G,K\}$
$\tilde{N}^e_{17} = 0{,}77/\{A,B,C,D,E,G,K\}+0{,}12/\{F,G\}+0{,}12/\{I,J\}$
$\tilde{N}^e_{18} = 0{,}63/\{A,B,C,D,E,G,K\}+0{,}18/\{F,G\}+0{,}18/\{I,J\}$
$\tilde{N}^e_{19} = 1/\{A,B,C,D,E,G,K\}$.

Evaluación de Iteración 19

Para la Iteración 19, si se evalúa la clasificación que se establece, con $\alpha = 0{,}6$ y calculamos w con la ecuación (1), se tiene $w = 0{,}73$.

Esto quiere decir que el 73% de los Ni están en una única clase de la partición, en al menos el 60% de su asignación total en la relación.

Es importante recalcar que los valores de α y w, son valores que se establecen de acuerdo con las necesidades y requerimientos de cada problema en particular, en este caso el análisis requería tomar los valores de $\alpha = 0{,}6$ y $w = 0{,}7$; de esta forma, la iteración 19 cumple por encima de los requerimientos de α y w.

Si los valores de α y w son muy pequeños, entonces, el número de grupos aumenta y si es muy grande disminuye, y puede correrse el riesgo de tener un solo grupo, que para efectos prácticos, es inconveniente, por lo que es importante tomar valores adecuados de α y w.

ALGORITMO DE DISTRIBUCION PARA UNA RELACION DISFUNCIONAL

Definición 3. Sea M1 un elemento del universo resultante del proceso de clasificación, (i.e. $M1 \in M^n$), y sea N1 el conjunto de elementos de N relacionados con los elementos de M1, donde el número de elementos de N1 y M1 son n1 y m1 respectivamente.

Si existen arcos $\{N1_p, M1_q\}$ con $p \in \{1...n1\}$ y $q \in \{1...m\}-\{1...m1\}$ tales que, la relación R definida por:

$$R = \{ (N1_i, M1_j) / i = 1...n1,\ j = 1...m1\} - \{(N1_p, M1_q) / p \in 1...n1\},\ q \in \{1...m1\} \}.$$

sea una relación disfuncional[6], entonces, se puede establecer un algoritmo eficiente (Θ (n1)), para resolver la distribución de dicha relación.

Algoritmo

Sea R1 un rectángulo[1], que relaciona los elementos de N1 y M1, además, la matriz asignada para los elementos de N1 y M1 es de tamaño sxt.

1. Suponga que el tamaño de las matrices es de 3 x 5 y que m1=5, entonces, la matriz del elemento 1 es la siguiente:

m_1	m_4	m_2	m_5	m_3	m_1
m_2	m_5	m_3	m_1	m_4	m_2
m_3	m_1	m_4	m_2	m_5	m_3

Las otras 4 matrices se generan por permutaciones, donde todos cambian de posición, y es posible solo 5 veces ya que, para una coordenada (i,j) los 5 elementos pueden estar solo una vez.

2. Se asignan elementos del conjunto Ni de tal forma que la suma de sus distribuciones coincida con los valores sxt de una tabla.

3. Se repite el paso 2, hasta que todos los elementos de Ni y todas las tablas estén asignadas.

Ejemplo 4. En el Cuadro 3 se muestran los datos de 5 objetos con 9 variables.

CUADRO 3. Datos del Ejemplo 4

	A	B	C	D	E
1	1	1	1	1	1
2	6	6	6	6	6
3	5	5	5	5	5
4	5	5	5	5	5
5	6	6	6	6	6
6	2	2	2	2	2
7	2	2	2	2	2
8	2	2	2	2	2
9	1	1	1	1	1

La solución que se obtiene al aplicar el algoritmo anterior es la siguiente:

Objeto A

5	4	7	2	3
5	4	7	2	1
6	2	3	5	4
6	2	3	5	9
3	5	4	8	2
3	5	4	8	2

Objeto B

2	3	5	4	8
2	3	5	4	8
5	4	7	2	3
5	4	7	2	1
6	2	3	5	4
6	2	3	5	9

Objeto C

4	6	2	3	5
4	6	2	3	5
2	3	5	4	8
2	3	5	9	8
5	4	7	2	3
5	4	7	2	1

Objeto D

3	5	4	7	2
3	5	4	7	2
4	6	2	3	5
4	6	2	1	5
2	3	5	4	8
2	3	5	9	8

Objeto E

6	2	3	5	4
6	2	3	5	9
3	5	4	7	2
3	5	4	7	2
4	8	2	3	5
4	8	2	1	5

Variable 1

				A
			D	B
			E	C

Variable 2

B	E	C	A	D
B	E	C	A	D
C	A	D	B	E
C	A	D	B	E
D	B	E	C	A
D	B	E	C	A

Variable 3

D	B	E	C	A
D	B	E	C	
E	C	A	D	B
E	C	A		
A	D	B	E	C
A	D	B		

Variable 4

C	A	D	B	E
C	A	D	B	
D	B	E	C	A
D	B	E		
E	C	A	D	B
E	C	A		

Variable 5

A	D	B	E	C
A	D	B	E	C
B	E	C	A	D
B	E	C	A	D
C	A	D	B	E
C	A	D	B	E

Variable 6

E	C			
E	C			
A	D			
A	D			
B				
B				

Variable 7

		A	D	
		A	D	
		B	E	
		B	E	
		C		
		C		

Variable 8

				B
				B
				C
				C
	E		A	D
	E		A	D

Variable 9

				E
			C	A
			D	B

Se puede observar en la solución del ejemplo anterior que la distribución que se establece es muy deseable, primero por la facilidad con que se realiza la distribución, y segundo, la distribución en la mayoría de los casos se realiza en el menor número posible de columnas (días) como puede apreciarse en los Variables 6 y 7.

CONCLUSIONES

Es importante resaltar el control difuso que se establece en cada iteración del algoritmo de clasificación jerárquica aglomerativo. Un problema de la clasificación propuesta por el método clásico de clasificación jerárquica, es que crea una jerarquía indexada, haciendo cálculos innecesarios dependiendo de la aplicación, además, no establece una partición determinada, sino que el usuario tiene que realizar el corte en el árbol jerárquico[2].

Este mismo enfoque (de control difuso), se puede adaptar a otros métodos de clasificación iterativa de objetos. Los parámetros α y w son establecidos por el experto de acuerdo con la calidad de la clasificación que se desea.

En la construcción de la relación disfuncional, al excluir algunos arcos de la relación, se genera una dependencia de relaciones disfuncionales, al servir estos arcos de enlace entre ambas relaciones; este problema de los enlaces es un tema para estudio de futuras investigaciones.

BIBLIOGRAFIA

1. A. Jaoua, M. Beaudry y J. Desharnais, *Difunctional Relations: A formal Tool For Program Specification*, Département d´Informatique Faculté des Sciences et de Génie, Université Laval, Québec, Canada, 1991.

2. E. Piza, La Clasificación Automática Jerárquica Aglomerativa, *Revista Ciencias Económicas*, Vol. VII, No. 1, pp. 95-111, 1987.

3. G. Brassard y P. Bratley, *Algorithmics: Theory and Practice*, Prentice Hall Int, New Jersey,1988.

4. H. J. Zimmermann, *Fuzzy Set Theory and Its Applications*, Kluwer-Nijhoff Publishing, Boston, 1985.

5. J. Espinoza y J. Trejos, Clasificación por Particiones, *Ciencia y Tecnología*, Vol. XIII, No. 1 y 2, pp. 129-154, 1989.

6. N. Le Thanh, Dependences *Difonctionnelles (Iso-Dependances) et la Methode de Decomposition Rectangulaire*, Papport de synthèse, Université de NICE-CNRS, France, 1992.

7. R. Slowinski y J. Stefanowski, Rough Classification in Incomplete Information System, *Mathl. Comput. Modelling*, Vol. 12, No. 10/11, pp. 1347-1357, 1989.

8. Z. Pawlak, Rough Classification, *Int. J. Man-Mach. Stud.*, Vol. 20, pp. 469-483, 1984.