Mora F., Walter; Astorga M., Alcides. *Clasificación de datos: implementación de un algoritmo clustering difuso.* **Tecnología en marcha.** Vol. 12. no. especial.

CLASIFICACION DE DATOS: IMPLEMENTACION DE UN ALGORITMO CLUSTERING DIFUSO

Walter Mora F. *
Alcides Astorga M. *

El enfoque tradicional (crisp) de la clasificación de datos por medio de particiones usando técnicas del análisis cluster exige que los subconjuntos obtenidos sean mutuamente excluyentes, lo cual, por el tipo de información que se usa en áreas como Medicina, Agronomía y Meteorología entre otras, es bastante artificial. En este artículo se relacionan algunos conceptos de la lógica difusa con el análisis cluster para obtener por medio de un algoritmo del tipo c-means, el grado de pertenencia de un dato a un subconjunto de la clasificación obtenida. Finalmente, en este trabajo, a modo de ejemplo clasificamos un conjunto de datos usando tanto análisis cluster crisp, como difuso. Para obtener los resultados se implementó el algoritmo ISODATA.

INTRODUCCION

En términos generales el *análisis cluster* consiste en distribuir una cantidad dada de objetos en una serie de conjuntos disjuntos, de forma tal que los elementos que se consideren similares pertenezcan al mismo conjunto. La palabra *clustering* se puede tomar como sinónimo de clasificación o de subconjunto.

La clasificación de objetos, utilizando técnicas del análisis cluster es una disciplina no muy reciente, pero que en los últimos años ha tenido un gran desarrollo teórico debido sobre todo a la amplia gama de aplicaciones que tiene en el campo científico.

Por ejemplo:

- La clasificación de animales y plantas es una actividad tan antigua como la humanidad, donde cada especie pertenece a una serie de subconjuntos que incrementan su tamaño conforme decrece el número de atributos comunes.
- En medicina, una de las actividades más importantes que se desarrolla es la clasificación de enfermedades de acuerdo con una serie de síntomas.
- En la Agricultura, Ciencias Forestales y Meteorología la clasificación climática de ciertas zonas es una actividad de vital importancia para el desarrollo de un país y supervivencia de la población.

Clasificación de datos (patrones, objetos) por medio de los métodos clustering

El modelo teórico con el que funciona el análisis clustering se puede plantear de la siguiente forma:

Dado un conjunto de n datos con sus correspondientes atributos, se busca distribuir esos datos, en c subconjuntos disjuntos, llamados *cluster*. El objetivo es que los objetos pertenecientes a un mismo *cluster* sean bastante similares

* Departamento de Matemática, Instituto Tecnológico de Costa Rica.

El análisis cluster consiste en distribuir una cantidad dada de objetos en una serie de conjuntos disjuntos, de forma tal que los elementos que se consideren similares pertenezcan al mismo conjunto.

entre sí, y los objetos de diferentes *cluster* sean tan diferentes como sea posible.

Varios problemas son inherentes a este tipo de modelos, dentro de los cuales podemos citar:

Determinación de una función de similitud (o disimilitud)

La clasificación de un conjunto de datos se hace de acuerdo con ciertas características que posea la información y que, a su vez, sean significativas para el usuario. Debe realizarse mediante algún criterio, previamente definido, de forma tal que los datos incluidos en un mismo conjunto presenten características muy parecidas, para esto se define una *función de distancia* la cual va a medir la similitud o semejanza, respecto a un criterio, entre elementos de un mismo cluster (y a su vez, la disimilitud entre elementos de diferentes cluster).

El tratamiento que se la ha dado a este problema, a saber, la selección de un criterio para medir la semejanza entre los datos, es abundante, variada y bastante heurística. Por ejemplo, autores como Bezdek, Han-Pao, Zimmermann, entre otros, han orientado sus esfuerzos para aportar diferentes soluciones a este problema.

Determinación del número de cluster

Para un conjunto de datos, sobre todo cuando no se tiene su representación geométrica, uno de los principales problemas es determinar *a priori* el número apropiado de subconjuntos que debe tener la clasificación. Obviamente, el número de cluster en cualquier caso dependerá de la función de similitud dada, y del grado de dispersión de los datos en el caso de que se use un tipo de índice de naturaleza geométrica. Aunque en muchos casos la selección del número de cluster será bas-tante heurística y dependerá del usuario, el tema de la selección de cuántos cluster son necesarios es un problema abierto.

Algoritmos clustering

Existen muchas caracterizaciones de los algoritmos cluster[3], pero, para efectos de este artículo, basta conocer la que divide los algoritmos en los que presuponen un conocimiento aproximado *a priori* del número de subconjuntos en que se tiene que dividir el conjunto de datos y aquellos que construyen en forma heurística el número de subconjuntos requeridos.

De estos algoritmos, los que han recibido mejor aceptación son los del primer tipo, ya que si bien presentan una incertidumbre con respecto a la partición inicial, son relativamente rápidos en los cálculos computacionales y siempre existe la posibilidad de mejorar la partición final variando la partición inicial.

Este tipo de algoritmos plantean el problema de la clasificación de datos de la siguiente forma:

Para un c escogido de antemano (fijo) y para un conjunto de n datos (patrones): denotados $x_1, x_2, x_3, ..., x_n$ contenidos en un espacio S, se busca obtener: $S_1, S_2, S_3, ..., S_c$, subconjuntos de S tal que todo x_i, con $i \in \{1, 2, ..., n\}$ pertenezca a uno de estos subconjuntos y ningún x_i pertenezca a dos subconjuntos diferentes, o sea:

i. $S_1 \cup S_2 \cup .... \cup S_c = S$
ii. $S_i \cap S_j = \varnothing \; \forall \; i \;\; j$

Uno de los algoritmos que se ubican dentro de esta familia lo constituye el ISODATA, del cual nos ocuparemos más adelante.

Algunos aspectos generales del análisis clustering

Cada partición que se haga de S viene descrita por una función indicadora u_i. tal que:

La clasificación de un conjunto de datos se hace de acuerdo con ciertas características que posea la información y que, a su vez, sean significativas para el usuario.

$$u_i: S \rightarrow \{0,1\}$$

donde $u_i(x) = 1$ si $x \in S_i$ y $u_i(x) = 0$, en caso contrario.

A su vez, la función indicadora u_i genera una matriz U de tamaño c x n donde cada entrada u_{ik}, viene dada por:

$$u_{ik} := u_i(x_k).$$

U recibe el nombre de *matriz de pertenencia.*

Además, a la función u_{ik} se le pide que cumpla las siguientes condiciones:

a) $\mu_{ik} \in \{0,1\}\ ;\ 1 \leq i \leq c\ ;\ 1 \leq k \leq n$

b) $\sum_{i=1}^{c} u_{ik} = 1; 1 \leq k \leq n$

c) $0 < \sum_{k=1}^{n} u_{ik} < n; 1 \leq i \leq c$

El punto (b) lo que afirma es que cada elemento del espacio de datos debe pertenecer a un solo cluster y (c) afirma que ningún cluster es vacío. En términos del análisis cluster tradicional la condición (b) es fundamental, aunque (c) se puede obviar.

Ejemplo. Si tenemos S = { x, y, z, w } y se quiere dividir S en tres subconjuntos disjuntos, entonces dos posibles particiones, indicadas por medio de matrices, pueden ser:

		x	y	z	w
	S_1	1	0	0	0
U_1=	S_2	0	1	1	0
	S_3	0	0	0	1

En este caso S_1= { x }; S_2= { y, z }; S_3= {w}

		x	y	z	w
	S_1	0	0	0	0
U_2=	S_2	1	1	1	0
	S_3	0	0	0	1

En este caso S_1= { }; S_2= { x, y, z }; S_3= {w}

ANALISIS CLUSTERING DIFUSO

Un aspecto relevante del enfoque tradicional expuesto, es el hecho de que cada dato debe ser asignado a un solo cluster (subconjunto) de la partición, esto se ve reflejado en la matriz de pertenencia en que la suma de cada columna debe dar 1 (condición b).

Este enfoque, desde el punto de vista teórico, es consistente y válido; sin embargo cuando es implementado en casos prácticos presenta algunas limitaciones.

La clasificación por medio de subconjuntos mutuamente excluyentes impone unas fronteras rígidas (hard), conllevando a que tal distribución, en algunos casos, pueda ser arbitraria, por cuanto podría darse que algún dato aporte información con respecto a las características ubicadas en el cluster S_i, pero es ubicado en el cluster S_j perdiéndose esta información.

Ejemplo. Considere las siguientes figuras:

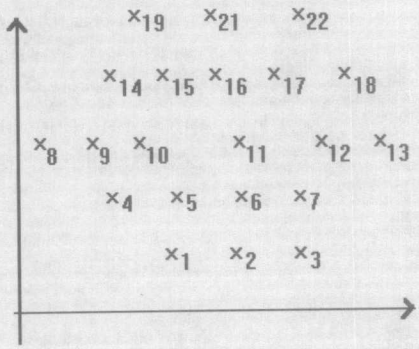

FIGURA 1

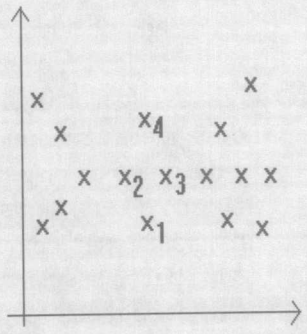

FIGURA 2

El incluir la lógica difusa en el análisis clustering, al menos en lo que nos corresponde, lo que busca es que un dato pueda pertenecer a diferentes clustering dependiendo de la información que aporten sus atributos.

Si en los casos anteriores el conjunto de datos es distribuido en dos clusters S_1 y S_2 y, en un instante del procedimiento de clasificación se tiene que:

Con respecto a la Figura 1:

$S_1 = \{ x_1, x_2, x_3, x_4, x_5, x_6, x_7 \}$
$S_2 = \{x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}\}$

Entonces se tiene que tomar una decisión con respecto a los elementos restantes, sin embargo, por la figura se puede ver que los elementos tienen igual posibilidad de pertenecer a S_1 o S_2.

Con respecto a la Figura 2, una decisión similar se debe tomar en la asignación de los elementos x_1, x_2, x_3 y x_4.

- En el caso de la medicina la presencia de ciertos síntomas no implica total certeza de la enfermedad X, por cuanto puede ser una manifestación, en cierto grado también de la enfermedad Y.

- En el caso de la clasificación de alimentos, tomando en cuenta calorías, calcio, grasas, proteínas, hierro, etc., no está demarcada tan claramente la línea divisora.

A partir de la introducción, por parte de Zadeh, de la *Lógica Difusa*, y de la incorporación de estas herramientas al campo de la clasificación de datos por métodos cluster es que muchos de los algoritmos vigentes han sido adaptados a las nuevas condiciones. Un caso típico es el ISODATA (crisp) para el cual Bezdek demostró que con las nuevas inclusiones difusas, se mantiene su convergencia[13].

El incluir la lógica difusa en el análisis clustering, al menos en lo que nos corresponde, lo que busca es que un dato pueda pertenecer a diferentes clustering dependiendo de la información que aporten sus atributos.

Aspectos generales del análisis cluster difuso

Una partición de un conjunto de datos S es descrita, en este caso, por un conjunto de c funciones continuas $\{\mu_i\}$ i =1,2,.,c que toma valores en [0,1], y donde $\mu_i(x)$ representa el grado de pertenencia de x al *cluster* S_i.

Lo anterior conduce a la determinación de una matriz $\hat{U}$ de tamaño c x n cuyas entradas μ_{ik} vienen dadas por:

$$\mu_{ik} = \mu_i(x_k),$$

y denotan el grado de pertenencia de x_k al cluster S_i y cumple:

a. $\mu_{ik} \in [0,1] \quad 1 \leq i \leq c, \quad 1 \leq k \leq n$

b. $\sum_{i=1}^{c} \mu_{ik} = 1 \qquad 1 \leq k \leq n$

c. $0 < \sum_{k=1}^{n} \mu_{ik} < n \qquad 1 \leq i \leq c$

Algoritmo ISODATA

Como se mencionó, existen diferentes algoritmos para resolver el problema de la clasificación de datos, dentro de este artículo se ha adoptado el enfoque que busca optimizar una función objetivo y se supone que el número de cluster es conocido. Un algoritmo que resuelve el problema cluster de esta manera es el ISODATA.

El funcionamiento de este algoritmo se puede describir de la siguiente forma:

Se eligen un número c de particiones y c centros v_i arbitrarios. Los v_i agrupan los patrones más cerca de ellos en el sentido de la función de similitud, posteriormente se recalculan los v_i como un "*promedio*" de todos los puntos en el cluster S_i. Un método para mejorar la partición inicial es el "*criterio de la variancia*". Este método mide la disimilitud entre los puntos en un

cluster y su centro v_i usando una norma euclídea. El objetivo es minimizar

$$(o)\ z(u,v) = \sum_{i=1}^{c}\sum_{k=1}^{n} U_{ik}\left\|x_k - v_i\right\|^2$$

donde $v = (v_1, v_2, ..., v_n)$.

Planteamiento del problema cluster en términos difusos

Esta metodología se puede enunciar así:

Encontrar c particiones difusas tales que minimicen:

$$(o)\ z(\mu,v) = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m\left\|x_k - v_i\right\|^2$$

Las particiones difusas son subconjuntos difusos caracterizados por las μ_{ik}. Esta metodología es una generalización clara del algoritmo ISODATA usando el criterio de la varianza.

Aplicando técnicas de análisis matemático y diferenciando parcialmente se obtiene que (o) tiene un mínimo local si:

$$(1)\ v_1 = \frac{1}{\sum_{k=1}^{n}(\mu_{ik})^m}\sum_{k=1}^{n}(\mu_{ik})^m x_k$$

$$i = 1,2,\ldots,c$$

$$(2)\ \mu_{ik} = \frac{\dfrac{1}{\left\|x_k - v_1\right\|^{\frac{1}{m-1}}}}{\sum_{j=1}^{c}\dfrac{1}{\left\|x_k - v_j\right\|^{\frac{1}{m-1}}}}$$

$$i=1,2,3,\ldots,c \qquad k=1,2,\ldots,n$$

En términos geométricos podemos decir que μ_{ik} es una ponderación de la atracción que ejerce v_i sobre x_k respecto a las atracciones de los otros v_j; así si x_k está muy cerca de v_i pero suficientemente lejos de los otros v_j entonces $\mu_{ik} \approx 1$.

Uno de los problemas que se presenta en la búsqueda del óptimo, es que las ecuaciones anteriores no pueden ser resueltas analíticamente (o sea obtener una solución exacta en un número finito de pasos), es por esto que en los trabajos realizados hasta la fecha, los métodos usados son algoritmos iterativos, uno de los cuales es ISODATA.

Implementación del Algoritmo ISODATA usando Lógica Difusa

Para resolver el problema (o) sujeto a (1) y (2) se ha aplicado el algoritmo (crisp) ISODATA, con las modificaciones correspondientes para el caso difuso.

La convergencia de ambos algoritmos está sustentada en [11] y [13].

Esquemáticamente el algoritmo (difuso) se puede enunciar así:

Paso 1: Seleccione c y m (2 c n ; 1 < m) e inicialice $\hat{U}^{(0)}$. Seleccione también una norma matricial adecuada. Haga h = 0

Paso 2: Calcule los $\{v_i^{(h)}\}$ usando las entradas de $\hat{U}^{(h)}$ y la condición (1).

Paso 3: Calcule la nueva matriz de pertenencia $\hat{U}^{(h+1)}$ usando $\{v_i^{(h)}\}$ de la condición (2) si x_k $v_i^{(h)}$. En caso contrario haga:

$\mu_{jk}=1$ para j=i
$\mu_{jk}=0$ para j≠i

Paso 4: Calcule = || $\hat{U}^{(h+1)}$ - $\hat{U}^{(h)}$ ||. Si > ε haga h= h+1 y vaya al paso 2.
Si ε finalice.

Aplicación del algoritmo ISODATA difuso a un conjunto de datos[1]

Para efectos del análisis que se realizará en esta sección, se presentan en el

1. Para la obtención de los resultados en esta sección los autores implementaron el algoritmo ISODATA crisp en Turbo Pascal 6.0 e ISODATA difuso en Modula-2.

CUADRO 1: Conjunto de datos

	P1	P2	P3	Q
A	70,0	60,0	75,0	80,0
B	100,0	100,0	100,0	100,0
C	50,0	40,0	30,0	20,0
D	65,0	80,0	80,0	80,0
E	70,0	70,0	70,0	70,0
F	10,0	30,0	45,0	60,0
G	45,0	60,0	100,0	34,0
H	80,0	90,0	45,0	30,0
I	12,0	65,0	90,0	100,0
J	99,0	90,0	92,0	89,0

Cuadro 1 una serie de datos numéricos simulados, los cuales corresponden a un conjunto de 10 estudiantes (denotados A, B, C, D, E, F, G, H, I, J) y a los cuales se les asignan 4 calificaciones (estos atributos se denotarán: P1,P2,P3 y Q).

El objetivo de realizar esta clasificación sobre este conjunto de datos es determinar las ventajas del algoritmo ISODATA difuso con respecto al caso crisp y la conveniencia de complementarlos.

Para hacer la clasificación por promedios se usará ISODATA con la métrica euclídea usual puesto que, por el tipo de datos que se trabajan y con una buena elección inicial de los centros v_i, se obtendrán cluster cuyos elementos (alumnos) tengan un promedio de notas muy similar.

Sin embargo los inconvenientes surgen cuando promedios "parecidos" quedan en distintos clusters. Es aquí donde la clasificación difusa se convierte en una herramienta para discriminar de acuerdo con el grado de pertenencia entre cada cluster "contiguo"; aparte de que nos da el grado de pertenencia de cada dato a todos los clusters S_i.

Para efectos de este trabajo se hacen dos clasificaciones, una con dos clusters y otra con tres clusters.

Clasificación con dos clusters

Sea S={A,B,...,J}.

Una corrida de ISODATA crisp produce el siguiente resultado:

S[1]= {A, B, D, E, G, H, I, J}
S[2]= {C, F}

S[1] y S[2] representan los clusters.

Una corrida de ISODATA difuso produce el siguiente resultado:

Matriz de Pertenencia

	A	B	C	D	E
S_1	0,87	0,86	0,14	0,97	0,88
S_2	0,13	0,14	0,86	0,03	0,12
	F	**G**	**H**	**I**	**J**
S_1	0,15	0,4	0,48	0,54	0,91
S_2	0,85	0,6	0,52	0,46	0,09

Una comparación de los clusters obtenidos en el caso crisp con los datos del Cuadro 1, evidencia que los estudiantes "buenos" (A, B, D y J) y los "no tan buenos"(G, H e I) son ubicados juntos en S[1] y los estudiantes "malos" en S[2].

En este contexto "bueno", "no tan bueno" y "malo" son en realidad conceptos difusos, por lo que:

S_1 = {x ∈ S/ x es un estudiante bueno}
S_2 = {x ∈ S/ x es un mal estudiante}

son dos subconjuntos difusos y la frontera entre ambos conjuntos no es clara.

Si observamos los resultados obtenidos en el caso difuso y los datos del Cuadro 1 se puede notar que verdaderamente G, H e I no se pueden catalogar como estudiantes "buenos" ni "malos", lo cual se puede apreciar en la clasificación difusa y se refleja en la matriz de pertenencia anterior pues, por ejemplo para G se tiene la información "G es un 60 malo y un 40 bueno", que corresponde a la noción de un estudiante "regular".

Clasificación en tres clusters

Una corrida de ISODATA crisp produce la siguiente clasificación:

S[1]= {A,D,E,G,H,I}
S[2]= {C,F}
S[3]= {B, J}

Una corrida en ISODATA difuso produce:

Matriz de pertenencia

	A	B	C	D	E
S1	0,87	0,04	0,12	0,84	0,93
S2	0,05	0,01	0,82	0,03	0,03
S3	0,09	0,94	0,05	0,12	0,04

	F	G	H	I	J
S_1	0,17	0,50	0,43	0,50	0,01
S_2	0,77	0,34	0,34	0,27	0,00
S_3	0,06	0,17	0,23	0,23	0,98

La clasificación hecha por cualquiera de los dos procedimientos, genera, en este caso, conceptos como "muy buenos", "buenos", y "malos".

La clasificación crisp incluye en el conjunto S[2] estudiantes que no son del todo "buenos", lo que conduciría a que en posteriores decisiones, G, H e I sean considerados como estudiantes "buenos".

Observemos que la clasificación usando ISODATA difuso refleja otra vez de una manera más realista la situación que se presenta en el Cuadro 1, por cuanto G,H e I son estudiantes que por sus notas no pueden ser catalogados como "buenos", ni "malos", o sea tienen algo de malo y algo de bueno, en síntesis son estudiantes regulares (por ejemplo G es 0,50 bueno y 0,34 malo).

Otro aspecto importante de notar en la matriz de pertenencia es la información que se obtiene con respecto al grado de pertenencia de un atributo con todos los clusters. Por ejemplo, si bien los alumnos G, H e I se pueden considerar "regulares", H e I tienen un valor de pertenencia de 0,23 en el cluster de los estudiantes "muy buenos" lo cual indica que H e I tienen algunas notas que son "muy buenas", sin embargo, el promedio general los ubica como estudiantes regulares.

CONCLUSIONES

La clasificación de datos mediante análisis clustering, ya sea por medio de métodos crisp o difusos es netamente heurística, por cuanto la selección del número de cluster, inicialización de la matriz de pertenencia, determinación de los centros de masa y el valor de m deben ser determinados por métodos experimentales que toman en cuenta las características de los atributos de los datos.

En todo caso, obviando los problemas mencionados, la clasificación de datos por métodos que incorporen elementos de lógica difusa, toman en cuenta en mejor forma que en el caso crisp, las características principales de los datos, por cuanto aparte de ubicar el elemento x en el cluster S_i nos dice en qué grado el elemento x disfruta de la propiedad que caracteriza a S_i e incluso de la propiedad que caracteriza a otro cluster S_j; lo cual puede ser una información valiosa cuando se trabajen grandes cantidades de datos.

Con respecto al caso difuso una forma de obtener resultados válidos es complementarlo con los métodos crisp. En el caso analizado en este artículo se ha usado la clasificación obtenida por métodos crisp, o sea los valores de la función indicadora, para inicializar la matriz de pertenencia en el algoritmo difuso.

LITERATURA CITADA

1. Bow S. T., *Pattern Recognition*, Marcel Deckker, Inc. New York,1984.

2. Dubois D., Prade H., *Théorie des Possibilités*, Masson, Paris, 1988.

3. Grujter, J. J., McBratney A. B., A modified fuzzy k-means method for predictive clasification,. *Clasification and Related Methods of Data Analysis*, North-Holland,1983, pp 97-105.

4. Hartigan, J.A., *Clustering Algorithms,* John Wiley e Hijos, USA, 1975.

5. Herder G. About Fuzzy Discrimination. COMPSTAT, Vienna 1982, pp.224-29.

6. Libert G., Roubens M., New Experimental Results in Cluster Validity of Fuzzy Clustering, *Clasification and Related Methods of Data Analysis,* North-Holland,1983, pp 205-217.

7. Libert G. y Roubens M., *News Trends in Data Analysis and Aplications*, Editores Jansen J. y Marcotochino J.F., North-Holland, pp. 205-217, 1983.

8. McBratney, A. B. y Moore, A.W., Aplication of Fuzzy Sets to Climate Clasificación, *Agricultural and Forest Meteorology*, Vol 35, 1985, pp. 165-185.

9. Pao Y.H., *Adaptive Pattern Recognition and Neural Networks*, Adison-Wesley, New York ,1989.

10. Peltier M.A. y Dubuisson B., A Human State Detection System Based on a Fuzzy Approach, *International Conference on Fault Diagnosis*, Toulouse, Abril 5-7, 1985.

11. Sabin M. J., Convergence and Consistency of Fuzzy c-means/ ISODATA Algorithms, *IEEE Transactions on Pattern Analyisis and Machine Intelligence,* Vol. Pami 9, No. 5, 1987.

12. Yager R, Approximate reasoning and possibilistic models in classification, *International Journal of Computer and Information Sciences.*, Vol 10, Nº 2, 1981, pp 141-175.

13. Zimmermann H.J., *Fuzzy Set Theory and its Aplications*, Kluwer- Nijhoff Publishing, USA, 1984.