Herramientas de clasificación difusa

Manual del usuario

Miguel E. Equihua Zamora
Xalapa, Ver., México
Febrero/2000

1. Introducción

1.1 Antecedentes de la teoría de conjuntos difusos

La teoría de conjuntos difusos fue propuesta inicialmente por Zadeh (1965) como una manera de representar y manejar incertidumbre no estadística (Bezdek 1987). Hay un gran cuerpo de aplicaciones de este enfoque teórico en una amplia gama de áreas, pero principalmente en el reconocimiento de patrones, el apoyo a la toma de decisiones y la inteligencia artificial. Este enfoque teórico ha sido controvertido en varias ocasiones (por ejemplo Tribus 1979).

Un rama de aplicaciones bien desarrollada es el análisis de cúmulos, que ha sido también discutido en la literatura ecológica (Dayong 1988, McBratney y Moore 1985). Roberts (1986) presenta un ejemplo de ordenación con base en la teoría de conjuntos difusos. En el área del manejo de los recursos naturales también ha encontrado aplicación este enfoque teórico (Ayyub y McCuen 1987, Wenger y Rong 1987). Una revisión general de posibles aplicaciones en ecología y una introducción al tema lo constituye el artículo de Bosserman y Ragade (1982).

Con objeto de facilitar la exposición subsecuente, es necesario introducir algunas nociones elementales de la teoría de conjuntos difusos lo que haré a continuación. En la teoría ordinaria de conjuntos, un elemento es miembro de un conjunto particular o no lo es en absoluto. Se puede asociar a cada elemento una variable indicadora del estatus de la afiliación en un conjunto. Esta variable indicadora tomará el valor de ´1´ si el elemento

es miembro del conjunto y el valor de ´0´en otro caso. En la teoría de conjuntos difusos esta idea la variable indicadora se extiende permitiendo que tome valores continuos en el intervalo [0, 1]. De esta manera, los elementos tienen efectivamente grados proporcionales de afiliación y por tanto, los límites del conjunto, que no precisamente definidos ya, se tornan borrosos o difusos. La variable indicadora puede ser representada por funciones adecuadas que pueden ser definidas subjetivamente o estimadas objetivamente (pueden inclusive ser expresadas en forma paramétrica en la forma que se suele emplear con las probabilidades). En algunas ocasiones es útil considerar que la función de afiliación es equivalente a una medida de probabilidad; pero debe notarse que este no es el caso en la mayoría de los casos, a pesar de que los valores de afiliación están constreñidos a caer dentro del intervalo [0, 1]. Una de las diferencias es que los valores de afiliación no se requiere que sumen ´1´ en el universo de medición, como es el caso de las probabilidades (Kauffmann 1975 y Bezdek 1981). Las probabilidades están relacionadas con la incertidumbre de observar un resultado particular y por tanto con el proceso de observación. Los valores de afiliación están relacionados a la definición del objeto mismo, es decir, pueden ser considerados como representantes del grado con el cual un objeto corresponde con la "descripción semántica" del conjunto (Bezdek 1987). Por ejemplo, un árbol viejo puede ser definido como uno que tiene *cerca* de 100 años o más. Un conjunto difuso correspondiente con esta definición podría ser:

Arbol viejo = { 85/0.80, 90/0.90, 95/0.95, 100 o más/1.00 }

en donde los valores después de la diagonal son valores de afiliación. Se sigue del ejemplo que cualquier árbol tendrá, según su edad, un grado de compatibilidad con el

conjunto definido arriba. Debe observarse que en este ejemplo no estamos suponiendo que la edad se mide con error ni ninguna clase de variaciones aleatorias. Lo que sí se está implicando es que la definición de "árbol viejo" es vaga. La teoría de conjuntos difusos busca precisamente capacitarnos para enfrentar situaciones en las que la incertidumbre deriva de definiciones vagas de los objetos.

1.2 Conjuntos difusos para definir sistemas ecológicos

De acuerdo con Bosserman y Ragade (1982), los sistemas entre más complejos a menudo tienen mayor número de variables de estado relevantes (aquellas necesarias para definir al sistema), debido a que hay un número más grande de descripciones alternativas del sistema. En consecuencia, muchos conceptos y definiciones en ecología son imprecisas porque los ecosistemas son grandes y de organización difusa. Con base en estas ideas parece atractivo utilizar en ecología un enfoque teórico que sea apropiado para enfrentarse con objetos definidos con imprecision.

La clasificación de comunidades ecológicas o de unidades ambientales en general es equivalente al concepto de partición de conjuntos. En la teoría de conjuntos ordinarios, una partición es la división de un conjunto original en subconjuntos que son mutuamente excluyentes y que no están vacíos (a estos subconjuntos se les denomina como "particiones duras" en la mayoría de la literatura de conjuntos borrosos). Este es el modo tradicional de construcción de los sistemas de clasificación tradicionales. En la teoría de conjuntos difusos una partición no consiste necesariamente de subconjuntos mutuamente

excluyentes; de hecho, una verdadera partición de conjuntos difusos debe tener al menos un par de subconjuntos con traslape (Bezdek 1981). Esta propiedad me parece muy interesante para la representación de sistemas ecológicos como las comunidades, las unidades de paisaje y los ecosistemas. Con la teoría de conjuntos difusos se puede lograr un acercamiento que a la vez incorpore conceptos de comunidades identificables (más o menos discretas) y gradientes. Actualmente es muy claro y ha sido ampliamente documentado que la composición específica de los sistemas ecológicos varia en forma más o menos continua a lo largo de gradientes ambientales; sin embargo, a pesar de ello en muchos casos es posible reconocer cierta estructura de comunidad. Esta estructura de los sistemas ecológicos es importante para explicar muchos fenómenos (Roughgarden y Diamond 1986).

Se puede apreciar fácilmente que los sistemas ecológicos, especialmente a nivel de comunidades y escalas similares, se ajustan mal a un sistema de particiones duras y por lo tanto es natural que los conceptos continuistas, al evitar estas restricciones, producen en general mejores resultados. No obstante, debe también tenerse presente que las técnicas de análisis de gradientes comúnmente empleadas han sido diseñadas para representar la muestra mediante la maximización de las distancias entre unidades de muestreo, lo que tiende a ocultar la estructura de distancias dentro de los grupos que podrían estar presentes. (Kruscl 1977). A pesar de esto, a menudo se pueden identificar grupos en los análisis de gradientes de muestras de comunidades ecológicas. Aún en el caso de que se acepte que no existen ecosistemas o comunidades como tales, a menudo pueden trazarse límites convencionalmente aceptables que satisfagan propósitos prácticos diversos (Greig-Smith 1983). Esto apunta nuevamente a que un enfoque de particiones

borrosas puede ser apropiado para describir sistemas ecológicos complejos, porque es entonces posible lograr una mezcla entre continuidad de variación e identificación de unidades, más o menos, discretas. Este tipo de sistemas pueden entonces ser descritos en un sentido "semántico" preciso.

1.3 Cúmulos con traslape en ecología

El objetivo primario de toda técnica de análisis de cúmulos es la partición de una muestra dada en grupos "homogéneos", relativamente hablando. El término homogéneo significa que todos los puntos en un mismo grupo están cerca unos de otros y alejados de los puntos de los restantes grupos. En el análisis de cúmulos clásico se supone implícitamente que existen subconjuntos disjuntos en la población. Sin embargo, de acuerdo con Dubois y Prade (1980) la separación de los cúmulos es una noción difusa en sí misma y por lo tanto la representación de los grupos por medio de conjuntos difusos es apropiada en situaciones en las que se espera que ocurra un traslape substancial o la mezcla de los grupos.

La idea de sistemas ecológicos deberían ser investigados con métodos que permiten el traslape entre los grupos ha estado presente en la literatura ecológica desde la década de los cincuentas (véase por ejemplo Fager 1957). Andre (1984) desarrolló un algoritmo para la identificación de lo que Fager había llamado " grupos recurrentes" con traslape, que son grupos de especies que tienden a presentarse juntas. Estas especies pueden considerarse como miembros frecuentes del ambiente mutuo y por lo tanto

agruparlas juntas es razonable.

Otro enfoque al análisis de cúmulos borrosos fue desarrollado por Dunn (1974a), Bezdek (1974, 1981 y 1987) y Bezdek *et al.* (1981a, 1981b). Este enfoque parece apropiado para el análisis de datos ecológicos. El método produce verdaderos subconjuntos difusos. El método es conocido como *k*-medias borroso (o también Isodata difuso). La técnica está basada en minimizar la función de mínimos cuadrados y abarca toda una familia de técnicas de clasificación numérica. El procedimiento Isodata difuso produce un agrupamiento basado en la identificación de casos "prototípicos", usualmente denominados como los "centroides". Los cúmulos son construidos minimizando las distancias entre esos centroides y cada una de las observaciones. Esta técnica ya ha sido utilizada en el análisis de datos ecológicos (véase por ejemplo Bezdek 1987, Equihua 1990, 1991).

2. El algoritmo básico de *k*-medias difuso

El algoritmo de cómputo de la clasificación numérica de *k*-medias está basado en minimizar la suma de cuadrados dentro de grupos

$$J_m(\mathbf{U},\mathbf{V},\mathbf{A}) = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m} \left\|\mathbf{x}_i - \boldsymbol{\mu}_j\right\|_{\mathbf{A}}^2 \qquad (1)$$

en donde

$$\left\|\mathbf{x}_i - \boldsymbol{\mu}_j\right\|_{\mathbf{A}}^2 = (\mathbf{x}_i - \mathbf{v}_j) = d_{ij\mathbf{A}}^2$$

es una medida de distancia calculada como una métrica inducida por una norma. **A** es la matriz que induce la norma (cualquier matriz de dimensión *p*, dada por el número de atributos en la muestra). **U** es la matriz de afiliaciones y **V** es la matriz de "centros de los cúmulos", $\mathbf{v}_j$. La constante *n* es el tamaño de muestra y *c* es el número de cúmulos. El vector $\boldsymbol{x}_i$ son las mediciones de los atributos observados en el individuo *i*. El término *m* es el parámetro de borrosidad. Si *m* = 1 entonces el algoritmo produce una partición dura y aumenta lo difuso de la clasificación conforme crece el valor de *m*. Aunque *m* puede tomar cualquier valor en el intervalo [1, ∞), se ha encontrado en forma meramente empírica, que en la mayoría de los casos valores cercanos a 2 producen los mejores resultados (Bezdek 1981, McBratney y Moore 1985 y Granath 1984).

Para minimizar la ecuación (1) se requiere satisfacer las siguientes dos expresiones:

$$\mathbf{v}_j = \frac{\sum_{i=1}^{n} u_{ij}^2 \mathbf{x}_i}{\sum_{i=1}^{n} u_{ij}^2} \quad , \quad 1 \le j \le c \tag{2}$$

$$u_{ij} = \frac{d_{ij\mathbf{A}}^{-2}}{\sum_{k=1}^{c} d_{ik\mathbf{A}}^{-2}} \quad , \quad 1 \le j \le c, 1 \le k \le c, 1 \le i \le n \tag{3}$$

Estas dos ecuaciones determinan las condiciones necesarias para que **U** y **V** se asocien con un mínimo local o un punto silla de la ecuación (1). La condición que se tiene cuando $\mathbf{x}_i = \mathbf{v}_j$ y que implica $d_{ij} = 0$ corresponde a una singularidad del problema, es decir, requiere un tratamiento especial en el cómputo por tratarse de una condición límite. En los casos en los que esto ocurre el individuo *i* no debe tener afiliación en ningún subgrupo más que en aquel en el cual se cumple d_{ij}=0. La otra restricción algorítmica que es importante mencionar es la de que se fuerza a que los valores de afiliación cumplan: $\sum u_{ij} = 1$ para cada individuo *i*.

Las matrices de inducción de norma más comúnmente empleadas son la idéntica, que induce una norma Euclideana, la matriz diagonal (formada con las inversas de las varianzas de cada atributo) que induce una norma diagonal y la inversa de la matriz de varianzas y covarianzas, que induce la norma de Mahalanobis. Las normas que se empleen se asocian con diferentes propiedades geométricas y estadísticas de los datos. En particular las normas indicadas arriba se relacionan con las siguientes propiedades de los datos y de los cúmulos (Bezdek 1981):

- Euclideana: es apropiada cuando los atributos son estadísticamente independientes y aproximadamente igualmente variables, la forma de los cúmulos es esférica.
- Diagonal: esta norma es útil cuando los atributos son estadísticamente independientes y marcadamente desigualmente variables, los cúmulos son de forma hiperelipsoidal
- Mahalanobis: esta norma encuentra aplicación en situaciones similares a las de la diagonal pero en circunstancias en donde los atributos tienen dependencia estadística.

A continuación describiré la estructura general del algoritmo difuso de *k*-medias (Bezdek 1981):

1. Seleccionar un número de cúmulos, $c > 1$; escoger cualquier métrica normalizada y fijar $m > 1$ (usualmente cerca de 2).
2. Inicializar **U** con una partición difusa arbitraria.
3. Calcular los *c* centros de los cúmulos difusos, $\mathbf{v}_j$, (ecuación 2).
4. Actualizar **U** usando la ecuación (3) y los nuevos centros de los cúmulos. Si ocurre una singularidad, asignar el valor de 1 a la afiliación correspondiente en el primer cúmulo para el que se cumpla $d_{ij}=0$, asignar valores de afiliación de 0 en el resto de los grupos.
5. Calcular el cambio entre $\mathbf{U}^{[r]}$ y $\mathbf{U}^{[r+1]}$. Si el cambio es menor que un valor seleccionado como criterio de convergencia detener el cómputo, en otro caso repetir el proceso a partir del paso 3.

Se ha demostrado que este algoritmo converge a un mínimo local o a un punto silla en cada secuencia de iteraciones y también que esta convergencia se alcanza a una tasa lineal (Hathaway y Bezdek 1986). Estos autores han demostrado además que hay un "dominio de atracción", del cual no puede escapar la secuencia de iteraciones. Esto implica que la oscilaciones entre dos distintos minimizadores en una secuencia de iteración no se espera, porque no es posible que ocurra analíticamente. Dadas las propiedades de convergencia a un mínimo local de este algoritmo de clasificación, es claro que diferentes

valores de afiliación iniciales pueden desembocar en diferentes minimizadores. La estrategia de inicialización adoptada en el programa que desarrollé, está basada en la división del primer componente principal de la muestra, aunque también es posible proporcionar un conjunto inicial externo.

3. Ejemplo

3.1 Contaminación del Agua

El ejemplo que se presenta es la clasificación difusa de los datos de calidad de agua en las 27 estaciones hidrográficas de la CNA en el estado de Veracruz. Los indicadores empleados son los valores para 1993 de sólidos suspendidos totales (SST93), la cantidad de nitratos (NO393) y la demanda bioquímica de oxígeno (DBO, 93). La mejor partición del conjunto de estaciones se obtuvo con tres subconjuntos. Los ríos representados en la muestra se enumeran a continuación y corresponden con los números de los sitios reportados en los resultados del análisis de clasificación difusa.

1 Río Tuxpan
2 Río Cazones
3 Río Tecolutla
4 Río Nautla
5 Río Nautla
6 Río Misantla
7 Río Actopan
8 Río Actopan
9 Río Actopan
10 Rio La Antigua
11 Río La Antigua
12 Río Jamapa
13 Río Jamapa
14 Río Jamapa
15 Río Blanco
16 Río Blanco
17 Río Blanco
18 Río Blanco
19 Río Papaloapan
20 Río Papaloapan
21 Río Papaloapan
22 Laguna Catemaco
23 Río Coatzacoalcos
24 Río Coatzacoalcos
25 Río Coatzacoalcos
26 Río Coatzacoalcos
27 Río Tonalá

En los resultados, se aprecia que los grupos 1 y 3 se caracterizan por ser estaciones con una alta cantidad de sólidos suspendidos pero el grupo 3 además agrupó a los sitios que tienen las mayores cargas de desechos orgánicos. Por otra parte, las estaciones del grupo 2 son las que se encuentran en mejores condiciones, relativamente hablando.

A continuación se muestra la clasificación difusa de los datos descritos arriba.

```
File used: \veracruz\agua93.prn
Number of individuals (rows): 27
Number of features (columns): 3
Number of clusters: 3
Analysis on the original coordinates (variables)
Method: fuzzy c-means
Fuzziness parameter:    2.000
Norm: Diagonal
Convergence criterion: 0.00010000

Partition coefficient:   0.5008  (3 clusters)

         CENTROIDS

          SST93       N0393      DBO93

  1 |   125.9846     0.8977     2.6948
  2 |    67.0575     0.4838     2.6152
  3 |   110.8696     0.6529    23.2307

         MEMBERSHIPS
               grupo 1   grupo 2    grupo 3
    1   |      0.0836     0.8719     0.0445
    2   |      0.5463     0.2662     0.1874
    3   |      0.6754     0.2250     0.0996
    4   |      0.1277     0.7918     0.0805
    5   |      0.0334     0.9595     0.0071
    6   |      0.0739     0.8971     0.0290
    7   |      0.2437     0.6668     0.0895
    8   |      0.4417     0.3577     0.2006
    9   |      0.0755     0.9003     0.0242
   10   |      0.7891     0.1407     0.0702
   11   |      0.2308     0.7357     0.0334
   12   |      0.9529     0.0376     0.0095
   13   |      0.8668     0.1142     0.0190
   14   |      0.8841     0.0866     0.0292
   15   |      0.5079     0.2160     0.2761
   16   |      0.1008     0.0900     0.8092
```

17	0.0516	0.0567	0.8918
18	0.2422	0.7110	0.0467
19	0.2325	0.7110	0.0565
20	0.1743	0.7976	0.0282
21	0.0206	0.9736	0.0058
22	0.1801	0.6172	0.2027
23	0.0449	0.9431	0.0121
24	0.0276	0.9660	0.0064
25	0.1219	0.8537	0.0244
26	0.4147	0.4987	0.0866
27	0.2976	0.6294	0.0730

4. Referencias

Ayyub B. M. y McCuen R.H., 1987. Quality and uncertainity assesment of wildlife habitat with fuzzy sets. *Journal of Water Resources Planning and Management*, 113 (1): 95-109.

Bezdek J.C., 1981. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York, 256 pp.

Bezdek J.C., 1987. Some non-standard clustering algorithms. *In:* Legendre P. y Legendre L. (eds.). *Developments in numerical ecology (NATO ASI Series Vol. G14)*, SpringerVerlag, Berlin, Heilderberg.

Bosserman R.W. y Ragade R.K., 1982. Ecosystems analysis using fuzzy set theory. *Ecological Modelling*, 16:191-208.

Roberts D.W., 1986. Ordination on the basis of fuzzy set theory. *Vegetatio*, 66:123-131.

McBratney A.B. y Moore A.W., 1985. Application of fuzzy sets to climatic classification. *Agricultural and Forest Meteorology*, 35:165-185.

Kauffmann A., 1975. *Introduction to the theory of fuzzy subsets (Vol. I fundamental theoretical elements).* Academic Press, EUA, 416pp.

Wenger R.B. y Rong Y., 1987. Two fuzzy set models for comprehensive environmental decision-making. *Jorunal of Environmental Management*, 25:167-180.