RNPS No. 2142
ISSN 2072-6287
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

Algoritmos de agrupamiento difuso, índices de validación: un estado del arte

Joan Sosa-García, Sandro Vega-Pons
y José Ruiz-Shulcloper

RT_049 **abril 2012**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

REPORTE TÉCNICO
Reconocimiento de Patrones

Algortmos de agrupamiento difuso, índices de validación: un estado del arte

Joan Sosa-García, Sandro Vega-Pons y
José Ruiz-Shulcloper

RT_048 **abril 2012**

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Algoritmos de agrupamiento difuso, índices de validación: un estado del arte

Joan Sosa-García[1], Sandro Vega-Pons[1] y José Ruiz-Shulcloper[2]

[1] Dpto. Minería de Datos, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV),
La Habana, Cuba
{jsosa,svega}@cenatav.co.cu

[2] Dpto. Reconocimiento de Patrones, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV),
La Habana, Cuba
jshulcloper@cenatav.co.cu

RT_049, Serie Azul, CENATAV
Aceptado: 9 de abril de 2012

Resumen. La clasificación no supervisada o agrupamiento es una técnica esencial en cualquier campo de investigación que involucre el análisis o procesamiento de datos, tales como la segmentación de imágenes, minería de datos, recuperación de documentos, etc. Los algoritmos de agrupamiento difuso surgen como una nueva alternativa para solucionar los problemas de clasificación no supervisada. Estos algoritmos brindan distintos puntos de vista para el análisis de los datos, así como una novedosa forma de descubrir la estructuración subyacente de los mismos. En el presente trabajo se abordarán algunos aspectos de la teoría de conjuntos difusos. Se analizarán los principales algoritmos de agrupamiento difuso existentes en la literatura, así como algunos de los índices de validación sobre los resultados de estos algoritmos. Adicionalmente, se presentará una taxonomía de los distintos algoritmos de agrupamiento difuso y, de igual manera, una para los índices de validación.

Palabras clave: conjuntos difusos, agrupamiento difuso, índices de validación.

Abstract. Unsupervised classification or clustering is an essential technique in any field of research involving data analysis or processing such as image segmentation, data mining, document retrieval, etc. Fuzzy clustering algorithms have emerged as a new alternative to solve the problem of clustering. They provide different perspectives for data analysis, as well as a new way to find out their underlying structure. In this paper we examine some aspects of the fuzzy set theory. We analyze the main fuzzy clustering algorithms in the literature as well as some of the validation indexes used to evaluate the results of these algorithms. Besides, we present a fuzzy clustering algorithms taxonomy and another for validation indexes.

Keywords: fuzzy set, fuzzy clustering, validity index.

1 Introducción

Cada día, las personas se encuentran una gran cantidad de información almacenada o representada como datos, para el posterior análisis y manejo de la misma. Una de las principales formas para el

análisis de los datos, es clasificarlos o agruparlos en un conjunto de categorías o grupos. Para el entendimiento de un nuevo objeto o el análisis de un nuevo fenómeno, las personas intentan, en la mayoría de los casos, observar las características que puedan describirlos y más aún, intentan compararlos con objetos o fenómenos ya conocidos, basados en alguna similitud o disimilitud (generalizadas como proximidad). Básicamente los sistemas de clasificación se pueden dividir en supervisados y no supervisados. La clasificación supervisada consiste en dada una colección de objetos previamente clasificados o etiquetados, llamado conjunto de entrenamiento, y un nuevo objeto aún no etiquetado, clasificar dicho objeto. Típicamente el conjunto de entrenamiento es utilizado para aprender las descripciones de las clases que son utilizadas para etiquetar el nuevo objeto.

Por otra parte, la clasificación no supervisada o agrupamiento se basa en estructurar un conjunto de datos en un número de grupos, donde los objetos dentro de un mismo grupo muestran un cierto grado de proximidad o similitud superior a los objetos que se encuentran en grupos distintos. El agrupamiento duro particional asigna cada elemento del conjunto a uno y solo uno de los grupos, asumiendo así los límites bien definidos entre los grupos. Este modelo a menudo no se corresponde con la descripción real de los datos, donde los límites entre los grupos pueden ser difusos y se requiere una descripción más matizada del objeto al grupo específico. De aquí surge la idea del agrupamiento difuso como alternativa para modelar lo más real posible los problemas en cuestión. En particular el agrupamiento difuso, acepta el hecho de que los grupos o clases en los datos no suelen estar completamente separados (ver Figura 1.1) y por lo tanto, se asigna un grado de pertenencia que por lo general suele estar entre 0 y 1, de cada elemento del conjunto de datos a cada uno de los grupos. Las técnicas más comunes de agrupamiento difuso se proponen minimizar una función objetivo cuyos principales parámetros son los grados de pertenencia de cada elemento a todos los grupos. Dichos parámetros determinan la localización, así como la forma de los grupos. Aunque la extensión determinista (agrupamiento duro particional) a agrupamiento difuso parece ser un concepto evidente, resulta que para la obtención de los grados de pertenencia, es necesario introducir un difusificador (*Fuzzifier*). Por lo general, el difusificador se utiliza simplemente para controlar en qué medida se solapan los grupos y suele ser muy útil e interesante para el problema en particular que se desea resolver.

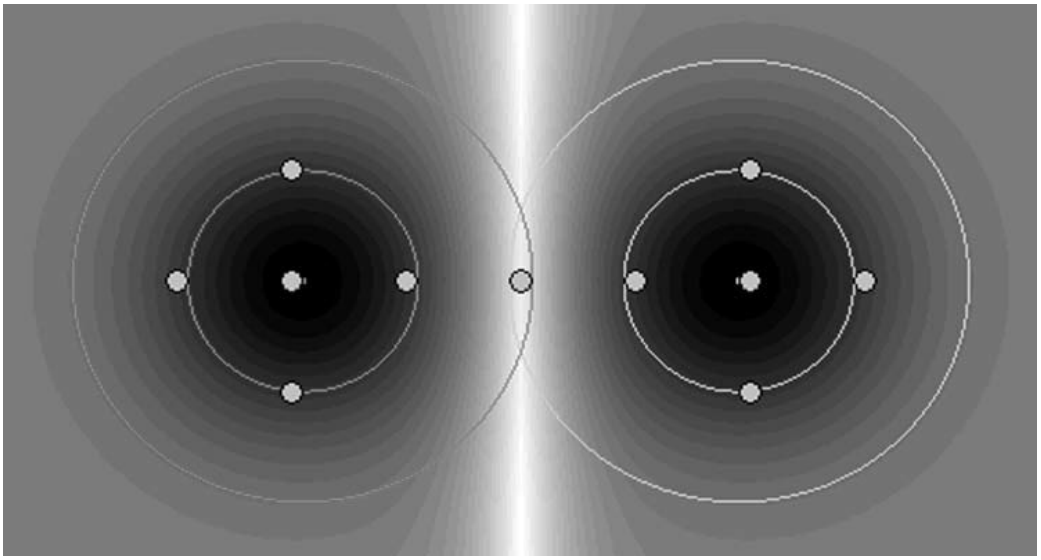

Fig. 1.1. Agrupamiento difuso de 2 grupos. Las zonas oscuras indican altos grados de pertenencia. En este caso cada elemento pertenece a los 2 grupos con un grado de pertenencia.

De esta manera, numerosos problemas en las ciencias de la vida son mejor abordados por la toma de decisiones en un ambiente difuso[1-7]. Bezdek [8] desarrolló una familia de algoritmos de agrupamiento, basados en la extensión difusa del criterio de error de los mínimos cuadrados y demostró la convergencia de los algoritmos a un mínimo local. Algoritmos relacionados, teniendo en cuenta

diferencias en la forma de los grupos, han sido propuestos por Bezdek y Dunn [9], Bezdek [10] y Gustafson y Kessel [11].

El desarrollo de algoritmos de agrupamientos difusos surge como una alternativa al proceso de agrupamiento. Esta variante difusa puede ser muy importante, ya que mediante esta técnica se puede simular con mayor exactitud los procesos y situaciones de la vida real, haciendo más clara y comprensible las soluciones a los problemas reales. Un ejemplo de su utilización es en la clasificación de noticias en la web. En este caso es más recomendable utilizar los algoritmos de agrupamiento difuso por la misma naturaleza de las noticias y su pertenencia a las distintas categorías, debido a que una misma noticia puede ser deportiva, cultural, económica, etc. Esta técnica ha sido desarrollada en la actualidad y una gran variedad de algoritmos se encuentran presentes en la literatura, cada uno con sus peculiaridades y su forma particular de abordar el tema. Existen dos familias de algoritmos de agrupamiento: los restringidos (en los que el número de agrupamientos a obtener es conocido o es un parámetro a determinar) y los libres en los que se desconoce este número.

Existen 3 dificultades principales encontradas durante el agrupamiento difuso sobre datos reales, en el caso del agrupamiento restringido:

- El número de grupos no siempre puede ser definido a priori y uno tiene que encontrar un criterio de validación (índice de validación) con el fin de descubrir el número óptimo de grupos en el conjunto de datos. En el caso de los agrupamientos libres el problema no es ese número sino la determinación del criterio de agrupamiento a aplicar.
- En el caso de los algoritmos basados en centroides, la ubicación de los mismos no es necesariamente conocida a priori y por lo tanto se tiene que hacer una distribución inicial de los centroides que puede influenciar para bien o para mal en el algoritmo. Es decir, existe una dependencia con la selección y ubicación de los mismos.
- La presencia de una gran variabilidad en las formas de los grupos y las variaciones en las densidades de los grupos.

En el presente trabajo se abordarán las principales técnicas de agrupamiento difuso empleadas en la actualidad, al igual que cómo se pueden enfrentar las dificultades mencionadas anteriormente. Este trabajo además, tiene como objetivo hacer un análisis crítico y valorativo sobre las principales técnicas de agrupamiento difuso, así como también las distintas medidas de evaluación para este tipo de estructuraciones.

El trabajo se divide en 5 secciones. La sección 2 aborda la Teoría de Conjuntos Difusos, para luego definir el problema del agrupamiento difuso. En la sección 3 se presentan las principales técnicas de agrupamiento difuso así como algunos de los algoritmos de agrupamiento difuso más utilizados, que implementan dichas técnicas. En la sección 4 se analizan los diferentes índices de validación para las estructuraciones difusas y por último, en la sección 5 se exponen las conclusiones del trabajo.

2 Teoría de conjuntos difusos

A continuación se define una serie de símbolos que serán utilizados posteriormente. Si X es un conjunto y x es un elemento de ese conjunto, se escribe de la forma $x \in X$. Las *conjunciones lógicas*: *implica*, *es implicado por*, *sí y solo sí* se denotan $\Longrightarrow, \Longleftarrow, \Longleftrightarrow$ respectivamente. El álgebra del *conjunto potencia* $P(X)$ *de* X (el conjunto de todos los subconjuntos (duros) de X) es formulada en términos de relaciones y operaciones familiares. Sean $A, B \in P(X)$:

Contiene: $A \subset B \Leftrightarrow \forall x \in X\ (x \in A \implies x \in B)$.

Igualdad: $A = B \Leftrightarrow A \subset B\ y\ B \subset A$.

Complemento: $\tilde{A} = \{x \in X \mid x \notin A\}$.

Intersección: $A \cap B = \{ x \in X | x \in A \, y \, x \in B \}$.

Unión: $A \cup B = \{ x \in X | x \in A \, ó \, x \in B \}$.

Se indica que A es subconjunto de $P(X)$ escribiendo $A \subset P(X)$, de la misma forma A es un elemento de $P(X)$, entonces $A \subset X \Longrightarrow A \in P(X)$. Existen otros símbolos, los cuales también se utilizarán como son: $\forall - para\ todo$, $\exists - existe$ y $| - tal\ que$. El conjunto que no contiene elementos es el conjunto vacío, denotado por $\emptyset$.

El quíntuplo de las operaciones primitivas es indicado de la forma $(\subset, =, \tilde{}, \cap, \cup)$. Diferentes estructuras algebraicas se pueden construir por la combinación de estas operaciones aplicadas a los elementos de $P(X)$, o más general aún, a los elementos de cualquier familia $\mathfrak{F}(X)$ de subconjuntos de X.

Se asume que X es un conjunto finito, siendo n su cardinalidad (el número de elementos que pertenecen a X), indicado de la forma $|X| = n$. Por consiguiente, $|P(X)| = 2^n$ y $P(X)$ satisface las siguientes propiedades:

$$\begin{gathered} \emptyset \subset P(X). \\ A \in P(X) \Longrightarrow \tilde{A} \in P(X). \\ A, B \in P(X) \Longrightarrow A \cup B \in P(X). \end{gathered} \tag{2.1}$$

Desde que los conjuntos difusos ($Fuzzy\ Sets$) son modelados matemáticamente mediante funciones, es intuitivo brindar otra caracterización de $(\subset, =, \tilde{}, \cap, \cup)$ en términos de funciones. Sea $A \in P(X)$, la función $u_A: X \longrightarrow \{0,1\}$ definida por

$$u_A(x) = \begin{cases} 1, & x \in A \\ 0, & otro\ caso \end{cases}. \tag{2.2}$$

Es la función característica (o indicador) del subconjunto duro $A \subset X$. Para cada $A \in P(X)$ existe una única u_A. Denotamos por $P(F_X)$ al conjunto de todas las funciones características sobre X. Las operaciones y relaciones de la Teoría de Conjuntos son equivalentes a las siguientes operaciones de la Teoría de Funciones:

Contiene: $u_A \leq u_B \Leftrightarrow u_A(x) \leq u_B(x), \ \forall x \in X$.

Igualdad: $u_A = u_B \Leftrightarrow u_A(x) = u_B(x), \ \forall x \in X$.

Complemento: $\tilde{u}_A(x) = 1 - u_A(x), \ \forall x \in X$.

Intersección: $u_{A \cap B}(x) = (u_A \wedge u_B)(x) = \min\{u_A(x), u_B(x)\}, \ \forall x \in X$.

Unión: $u_{A \cup B}(x) = (u_A \vee u_B)(x) = ma\,x\{u_A(x), u_B(x)\}, \ \forall x \in X$.

El conjunto vacío es la función constante $\mathbb{O} \in P(F), \mathbb{O}(x) = 0 \ \forall x \in X$, el conjunto entero X es la función constante $\mathbb{I} \in P(F), \mathbb{I}(x) = 1 \ \forall x \in X$.

Siguiendo la idea original de Zadeh [12], se define formalmente un Conjunto Difuso A de X expresado por una función de pertenencia (función característica) $u_A: X \longrightarrow [0,1]$, donde el valor de $u_A(x')$ representa el grado de pertenencia de x' en A. Valor de $u_A(x')$ cercano a 1, eleva el grado de pertenencia de x' al conjunto difuso A. Se denota $P_f(F)$ como el conjunto de todos los subconjuntos difusos de X (equivalente al conjunto de todas las funciones u definidas sobre X con valores en $[0,1]$). Las operaciones definidas anteriormente son válidas para los conjuntos difusos.

Cardinalidad de un conjunto difuso: La cardinalidad de un conjunto difuso, al igual que en el caso de los conjuntos duros, se define como la suma de los grados de pertenencia de todos los elementos al conjunto y se expresa de la siguiente forma:

Sea A un conjunto difuso definido sobre X

$$|A| = \sum_{x \in X} u_A(x)\,, \tag{2.3}$$

donde $u_A: X \longrightarrow [0,1]$.

Además de la unión e intersección, se pueden definir otras formas de construir combinaciones de conjuntos difusos [12]. A continuación se mencionarán algunas.

Producto algebraico: El producto algebraico entre A y B se denota por AB y es definido en términos de la función de pertenencia de A y B por:

$$u_{AB}(x) = u_A(x)u_B(x)\,. \tag{2.4}$$

Claramente,

$$AB \subset A \cap B\,. \tag{2.5}$$

Suma algebraica: La suma algebraica de A y B es denotada por $A + B$ y se define mediante:

$$u_{A+B}(x) = u_A(x) + u_B(x)\,, \tag{2.6}$$

siempre y cuando la suma $u_A(x) + u_B(x) \leq 1$. De esta manera, a diferencia del producto algebraico, la suma algebraica es significativa solamente cuando $u_A(x) + u_B(x) \leq 1, \forall\, x \in X$.

2.1 Relaciones difusas [12]

El concepto de una relación (la cual es una generalización de una función) tiene una extensión natural a los conjuntos difusos y desempeña un importante papel en la teoría de conjuntos y sus aplicaciones. Al igual que en el caso de conjuntos duros, para los cuales existen relaciones binarias, ternarias,…, z-arias, para los conjuntos difusos se pueden definir este tipo de relaciones lo que en este caso se llamarían Relaciones Difusas.

Ordinariamente, una relación binaria es definida como un conjunto de pares ordenados [13], un ejemplo: el conjunto de todos los pares ordenados de números reales x y y tal que $x \leq y$. En el contexto de los conjuntos difusos, una relación difusa binaria en X es un conjunto difuso en el espacio $X \times X$. Por ejemplo, la relación denotada por $x >> y$, donde $x, y \in \mathbb{R}^1$, puede ser estimado como un conjunto difuso A en $\mathbb{R}^2$, con la función de pertenencia de A, $\mu_A(x, y)$, teniendo los siguientes valores representativos: $\mu_A(10, 5) = 0$; $\mu_A(100, 10) = .7$; $\mu_A(100, 1) = 1$; etc.

Más general, se puede definir una relación difusa $z - aria$ en X como un conjunto difuso A en el espacio $X \times X \times \ldots \times X$. Para estas relaciones, la función de pertenencia es de la forma $\mu_A(x_1, x_2, \ldots, x_z)$, donde $x_i \in X, \forall\, i = 1, \ldots, z$.

En el caso de relaciones difusas binarias, la composición de dos relaciones difusas A y B es denotado por $B\ o\ A$ y se define como una relación difusa en X donde tiene como función de pertenencia:

$$\mu_{B\ o\ A}(x, y) = Sup_v\ Min[\mu_A(x, v), \mu_B(v, y)]\,. \tag{2.7}$$

Una relación (dura) R sobre X se define como una función $R: X \times X \longrightarrow \{0,1\}$ donde $x, y \in X$, se dice que están relacionados si $R(x, y) = 1$. Una relación R en X se dice que es una relación de equivalencia sí y solo sí $\forall x, y, z \in X$ se cumple:

$$\begin{aligned} R(x,x) &= 1, reflexiva. \\ R(x,y) &= R(y,x), simétrica. \end{aligned} \tag{2.8}$$

$R(x,z) = R(y,z) = 1 \Longrightarrow R(x,y) = 1, transitiva$.

En Zadeh [14] se define una relación de similitud S en X si y solo si S es una relación difusa y $\forall\, x, y\, , z \in X$, se cumple:

$$\begin{aligned} S(x,x) &= 1, reflexiva. \\ S(x,y) &= S(y,x), simétrica. \\ S(x,y) &\geq \bigvee_{z \in X} \big(S(x,z) \wedge S(y,z)\big), transitividad\, . \end{aligned} \tag{2.9}$$

Donde ∨ y ∧ denotan máximo y mínimo respectivamente. Como se puede apreciar la relación de similitud S es una generalización de la relación de equivalencia.

Las relaciones difusas no han sido utilizadas solamente en el ámbito de los algoritmos de agrupamiento difuso, sino que también han sido aplicadas a la combinación de los resultados de los algoritmos de agrupamiento jerárquico. En esta combinación, la idea básica es combinar las distintas jerarquías producidas por algoritmos de agrupamiento jerárquico [15].

2.2 Espacio de las particiones difusas

Ruspini [16] introdujo una $c - partición\ difusa\ \mu = (\mu_1, \mu_2, \ldots, \mu_c\ )$ asumiendo que $\mu_i(x)$ es una función con valores en el intervalo [0,1], tal que $\mu_1(x) + \mu_2(x) + \ldots + \mu_c(x) = 1$.

Consideremos un conjunto finito de datos $X = \{x_1, x_2, \ldots, x_n\}$ de $\mathbb{R}^p$. Se denota $\mu_i(x_j)$ por μ_{ij}, $\forall i, j, i = 1, \ldots, c,\ j = 1, \ldots, n$ y $r(x_j, x_k)$ por $r_{jk}, \forall j, k = 1, \ldots, n$. Sea V_{cn} el conjunto de todas las matrices de dimensión $c \times n$ y sea u_{ij} el $ij - ésimo$ elemento de $U \in V_{cn}$. Siguiendo Bezdek [17], se define:

$$M_c = \left\{ U \in V_{cn} |\, \mu_{ij} \in \{0,1\} \forall i, j;\ \sum_{i=1}^{c} \mu_{ij} = 1\ \forall j;\ 0 < \sum_{j=1}^{n} \mu_{ij}\ \forall i \right\}. \tag{2.10}$$

Entonces M_c es exactamente el espacio de las $c - particiones\ duras$ (no degeneradas) para el conjunto finito de datos X. Se define $A \leq B$ si y solo si $a_{ij} \leq b_{ij}, \forall i, j$ donde $A = [a_{ij}]$ y $B = \left[b_{ij}\right] \in V_{nn}$. Se define $RoR = [r'_{ij}] \in V_{nn}$ con $r'_{ij} = \bigvee_{k=1}^{n} \left(r_{ik} \wedge r_{kj}\right)$. Sea

$$R_n = \left\{\, R \in V_{nn} | r_{ij} \in \{0,1\} \forall i, j; I \leq R; R = R^T; R = RoR \,\right\}. \tag{2.11}$$

Entonces R_n es el conjunto de todas las relaciones de equivalencias sobre X. Para cualquier $U = \left[\mu_{ij}\right] \in M_c$, sea la matriz de relación $R = \left[r_{jk}\right]$ en V_{nn} definida por

$$r_{jk} = \begin{cases} 1, & si\ \mu_{ij} = \mu_{ik} = 1\ para\ algún\ i, \\ 0, & en\ otro\ caso \end{cases}. \tag{2.12}$$

Por lo que R es obviamente una relación de equivalencia correspondiente a las $c - particiones\ duras\ U$, ya que satisface las condiciones de reflexividad, simetría y transitividad. Es

decir, para toda $U \in M_c$ existe una matriz de relación R en R_n tal que R es una relación de equivalencia correspondiente a U.

Para una extensión difusa de M_c y R_n, sea:

$$M_{fc} = \left\{ V \in V_{cn} \mid \mu_{ij} \in [0,1]\ \forall i,j;\ \sum_{i=1}^{c} \mu_{ij} = 1\ \forall j;\ \sum_{j=1}^{n} \mu_{ij} > 0\ \forall i \right\}. \tag{2.13}$$

$$R_{fn} = \left\{ R \in V_{nn} \mid r_{ij}\ \in [0,1]\ \forall i,j; I \leq R; R = R^T; R \geq RoR \right\}. \tag{2.14}$$

Entonces M_{fc} es el espacio de las $c - particiones\ difusas$ (no degeneradas) de X y R_{fn} es el conjunto de todas las relaciones de similitud en X. Bezdek y Harris [18] mostraron que $M_c \subset M_{co} \subset M_{fc}$ y además que M_{fc} es la envoltura convexa[1] M_{co}, donde M_{co} es el conjunto de las matrices obtenidas relajando la última condición de M_c a $\sum_{j=1}^{n} \mu_{ij} \geq 0\ \forall i$.

2.3 Definición del problema de agrupamiento difuso

Sea $X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^p$ el conjunto de todos los elementos. De forma intuitiva, el principal objetivo de un algoritmo de agrupamiento difuso es brindar como resultado una $c - partición\ difusa$ (M_{fc}), donde se encuentre contemplado el grado de pertenencia de cada elemento del conjunto hacia todos los grupos difusos dados como resultado del algoritmo. Vale destacar, que existen otros tipos de particiones difusas [19], las que pueden relajar algunas de las condiciones que se cumplen para M_{fc}, aunque en el presente trabajo nos centraremos fundamentalmente en las $c - particiones\ difusas$ (M_{fc}) definidas anteriormente.

3 Agrupamiento difuso

Los algoritmos de agrupamiento se han estudiado durante mucho tiempo y hoy en día son utilizados en un gran número de aplicaciones. Se evidencia su utilización en los sistemas de acceso a la información, de minería de datos o de visión por computadoras. Diferentes algoritmos han sido desarrollados utilizando distintos enfoques y teniendo en cuenta supuestos subyacentes de los datos y sobre el conjunto final de grupos. Fuzzy C-means, algoritmos basados en Fuzzy C-means [20], Metis, Aprendizaje Participativo (PL-A) [21], Agrupamiento Difuso Jerárquico [22, 23], Agrupamiento Difuso basado en Computación Evolutiva [24, 25], etc., son algunos de los algoritmos y técnicas de agrupamiento conocidas. Los algoritmos de agrupamiento, de forma general, se pueden clasificar siguiendo diferentes criterios.

---

[1] En Matemática se define la envoltura convexa de un conjunto de puntos X de dimensión n como la intersección de todos los conjuntos convexos que contienen a X. Dados k puntos $x_1, \dots, x_k$ su envoltura convexa C viene dada por la expresión:

$$C(X) = \left\{ \sum_{i=1}^{k} \alpha_i x_i \,\middle|\, x_i \in X, \alpha_i \in \mathbb{R}, \alpha_i \geq 0, \sum_{i=1}^{k} \alpha_i = 1 \right\}.$$

En el caso particular de puntos en un plano, si no todos los puntos están alineados, entonces su envoltura convexa corresponde a un polígono convexo cuyos vértices son algunos de los puntos del conjunto inicial de puntos. Una forma intuitiva de ver la envoltura convexa de un conjunto de puntos en el plano, es imaginar una banda elástica estirada que los encierra a todos. Cuando se libere la banda elástica tomará la forma de la envoltura convexa.

Uno de los criterios es la *dirección* del proceso de agrupamiento (para los algoritmos jerárquicos). En este caso, los métodos son divididos en aglomerativos y divisivos. Los algoritmos aglomerativos construyen los grupos mediante la unión de aquellos grupos que son más similares. Esto corresponde a una técnica *bottom-up*. Los algoritmos divisivos por su parte siguen una técnica *top-down*.

Otro criterio corresponde a los grados de pertenencia de los elementos a los grupos. En este caso se pueden dividir en duros y difusos. En el agrupamiento duro, los grados de pertenencia de un elemento a un grupo son *booleanos* o $\{0, 1\}$ donde el valor 0 indica que el elemento no pertenece al grupo y 1 que sí pertenece. En el caso del agrupamiento difuso, como ha sido definido en la sección 2, los grados de pertenencia se toman en un intervalo, que por lo general suele ser $[0, 1]$.

En el caso particular de los algoritmos de agrupamiento difuso se han desarrollado muchos métodos, mediante el uso de diferentes técnicas de agrupamiento. Los principales algoritmos son resumidos en la siguiente taxonomía.

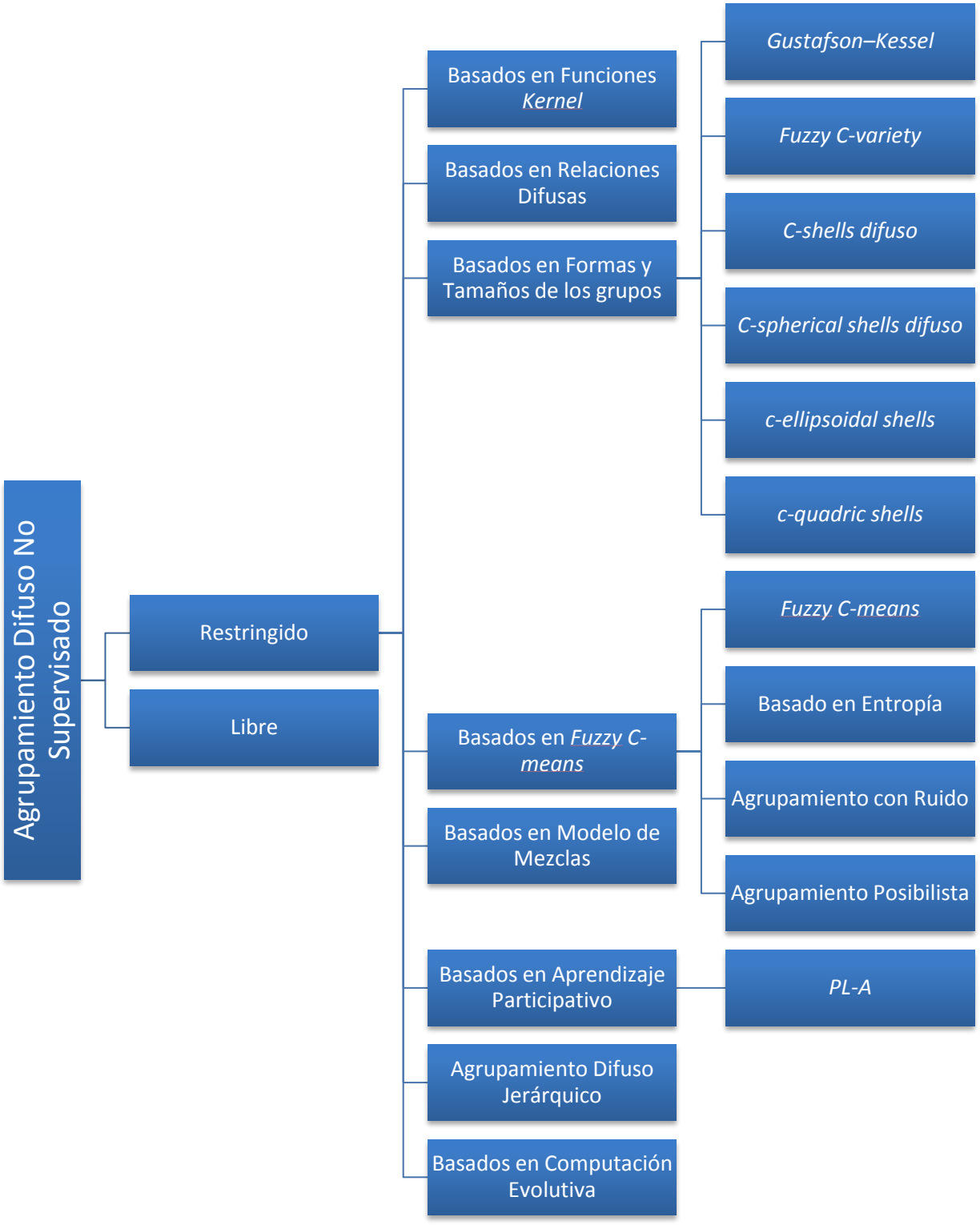

Fig. 3.1. Taxonomía de los principales algoritmos de agrupamiento difuso restringidos.

A continuación se analizan algunos de estos métodos.

3.1 Agrupamiento mediante relaciones difusas

Los agrupamientos difusos, basados en relaciones difusas, fueron propuestos por vez primera por Tamura [26]. Donde propone un procedimiento de $N - pasos$ utilizando la composición de relaciones difusas partiendo de una relación difusa R sobre X, reflexiva y simétrica. Se demostró que existe un t tal que $I \leq R \leq R^2 \leq \cdots \leq R^t = R^{t+1} = \cdots = R^\infty$ cuando X es un conjunto finito. Entonces R^t es utilizado para definir una relación de equivalencia R_λ mediante la regla: $R_\lambda(x, y) = 1 \Leftrightarrow \lambda \leq R^t(x, y)$. Efectivamente, R^t es una relación de similitud. Consecuentemente se puede particionar el conjunto de elementos en grupos mediante la relación de equivalencia R_λ. Para todo $0 \leq \lambda_k \leq \cdots \leq \lambda_2 \leq \lambda_1 \leq 1$, se puede obtener el correspondiente $k - nivel$ de la jerarquía de grupos, $D_i = \{grupos\ equivalentes\ \ de\ R_{\lambda_i}\ en\ X\},\ i = 1, \ldots, k$ donde D_i refina D_j para $i < j$. Esta jerarquía de grupos es conocida como métodos jerárquicos. Dunn [27] propuso un método para computar R^t basado en el algoritmo de $Spanning\ Tree$ de Prim. A continuación se muestra un ejemplo:

$$R = \begin{bmatrix} 1 & & & \\ .4 & 1 & & \\ .8 & .6 & 1 & \\ .3 & 0 & 0 & 1 \end{bmatrix}$$

$$R^2 = \begin{bmatrix} 1 & & & \\ .6 & 1 & & \\ .8 & .6 & 1 & \\ .3 & .3 & .3 & 1 \end{bmatrix} = R^3 = \begin{bmatrix} 1 & & & \\ .6 & 1 & & \\ .8 & .6 & 1 & \\ .3 & .3 & .3 & 1 \end{bmatrix}$$

Consecuentemente,

$$\lambda = .29 \Longrightarrow \{1, 2, 3, 4\}.$$

$$\lambda = .59 \Longrightarrow \{1, 2, 3\} \cup \{4\}.$$

$$\lambda = .79 \Longrightarrow \{1, 3\} \cup \{2\} \cup \{4\}.$$

$$\lambda = .81 \Longrightarrow \{1\} \cup \{2\} \cup \{3\} \cup \{4\}.$$

Una de las principales críticas a estos métodos, es el alto costo computacional producto de la cantidad de operaciones que se deben realizar con las matrices como la multiplicación, entre otras. En [28] se propone un algoritmo nuevo para calcular eficientemente R^t, basado en las propiedades de las relaciones difusas y de los $\lambda - corte$. Este algoritmo posee un costo computacional $O(n^2)$ y espacial $O(1)$. Estos costos evidencian una mejora sobre los algoritmos anteriores que estaban basados en métodos cuadrados y *Spanning Tree*, los cuales poseían unos costos de $O(n^3 \log_2 n)$ y $O(n^2)$ computacional y espacial, respectivamente.

Esto es un tipo de agrupamiento basado en la Teoría de Conjuntos Difusos. Sin embargo, estos métodos son eventualmente métodos noveles y nada más se aplican algoritmos jerárquicos aglomerativos (particionales), que no conllevan a particiones difusas. Por esta razón no se ha investigado mucho y no se han obtenido grandes resultados.

3.2 K-means ($Hard\ C - means$) [29]

El algoritmo $c - means$ agrupa los objetos de un conjunto dado $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^p$ en c conjuntos disjuntos G_i $(i = 1, 2, 3, \ldots, c)$ donde cada uno de estos es llamado grupo. En cada grupo se

determina el centro del grupo o centroide. A continuación se describe un simple procedimiento, el cual es utilizado comúnmente.

3.2.1 Algoritmo KM:

KM1 [Generar los valores iniciales]: Generar c centroides iniciales v_i $(i = 1, 2, 3, \dots, c)$.

KM2 [Ubicar al centroide más cercano]: Ubicar cada uno de los objetos x_k $(k = 1, 2, 3, \dots, n)$, al grupo donde se encuentre más cerca del centroide.

$$x_k \in G_i \Leftrightarrow i = arg\ min_{1 \le j \le c}(d_{jk})^2,$$

donde $(d_{jk})^2 = \|x_k - v_j\|^2$.

KM3 [Actualizar los centroides]: Calcular el nuevo centroide de cada grupo.

$$v_i = \frac{1}{|G_i|} \sum_{x_k \in G_i} x_k \,,$$

donde $|G_i|$ es el número de elementos en $G_i, i = 1, 2, 3, \dots, c$.

KM4 [Comprobar la convergencia]: Si los grupos son convergentes, parar; sino ir para el paso **KM2**.

Una de las posibles formas de comprobar la convergencia pudiera ser:

- Los objetos no cambian de grupo.
- Los centroides no cambian.
- Un número finito de iteraciones.

3.2.2 Formulación de optimización

Sea $G = (G_1, \dots, G_c)$ y $v = (v_1, v_2, \dots, v_c) \in \mathbb{R}^{cp}, con\ v_i \in \mathbb{R}^p$. Los conjuntos $G_1, \dots, G_c$ son disjuntos y su unión es el conjunto de todos los objetos:

$$\bigcup_{i=1}^{c} G_i = X, \;\; G_i \cap G_j = \emptyset \,, \forall i, j, i \neq j \,. \tag{3.1}$$

Esto se considera una partición del conjunto X. Consideremos la siguiente función objetivo:

$$J_{cm}(G, v) = \sum_{i=1}^{c} \sum_{x_k \in G_i} (d_{ik})^2 \,. \tag{3.2}$$

La función $J_{cm}(G, v)$ es una de las primeras funciones objetivo de la cual derivan una gran cantidad de funciones. Para la minimización de esta función, se realiza un proceso iterativo, como fue descrito anteriormente. Este algoritmo solamente produce particiones del conjunto de elementos, donde cada elemento pertenece a uno y solo uno de los grupos. Variantes de la función J_{cm} son utilizadas con el objetivo de obtener $c - particiones\ difusas$.

3.3 *Fuzzy C-means* [8]

En general las técnicas de agrupamiento difuso (*Fuzzy Clustering*) se basan en encontrar el mínimo de una función objetivo que determina los prototipos o centroides de los grupos buscados. El número de grupos suele ser un parámetro conocido c. Sea

$$X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^p \,,$$

el conjunto de todos los elementos. Sea

$$v = (v_1, v_2, \dots, v_c) \in \mathbb{R}^{cp}, \qquad con\ v_i \in \mathbb{R}^p\ ,$$

donde v_i es el centroide del grupo i y llamemos u_{ik} al grado de pertenencia del elemento x_k al grupo i. Una de las funciones objetivo más utilizadas y de la cual derivan una gran variedad de funciones es $J_2: M_{co} \times \mathbb{R}^{cp}$ definida por [30]:

$$J_2(U, v) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^2 (d_{ik})^2\ , \tag{3.3}$$

donde $U \in M_{fc}$ es una $c - partición\ difusa$ de X y $d_{ik} = \|x_k - v_i\|^2$ ($\|\ .\ \|$ es cualquier producto interno inducido por una norma sobre $\mathbb{R}^p$).

Esta función objetivo fue introducida por Dunn en [30], donde también realiza un análisis de la siguiente función:

$$J_1(U, v) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} (d_{ik})^2\ . \tag{3.4}$$

Ambas funciones surgen con el propósito de obtener soluciones difusas aunque demuestra que para esta última el mínimo sólo puede ser alcanzado cuando $u_{ik} \in \{0, 1\}$.

Como se puede apreciar, en ambas funciones la única diferencia es el exponente que afecta los grados de pertenencia, uno de los motivos que lleva a Bezdek [8] a generalizar las mismas de la siguiente manera:

$$J_m(U, v) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m (d_{ik})^2\ , \tag{3.5}$$

donde m es el exponente de ponderación: $m \in [1, \infty)$.

Dicho parámetro regula el grado difuso de la $c - partición\ difusa$. Debido a que cada término de J_m es proporcional a $(d_{ik})^2$, J_m es el error cuadrático del criterio de agrupamiento y las soluciones de:

$$min_{M_{fc} \times \mathbb{R}^{cp}} \{J_m(U, v)\}\ , \tag{3.6}$$

son los puntos estacionarios de J_m de menor error cuadrático. Se obtiene una familia infinita de algoritmos de agrupamiento difuso (una para cada $m \in [1, \infty)$), mediante las condiciones necesarias para las soluciones de (3.6). El teorema básico se menciona a continuación.

Teorema I [17]: Asumir que $\|\ .\ \|$ es un producto interno inducido: fijar $m \in [1, \infty)$, sea X el conjunto de los elementos, donde existan al menos $c < n$ elementos distintos y se define $\forall k$ los conjuntos:

$$I_k = \{i | 1 \le i \le c; d_{ik} = \|x_k - v_i\| = 0\}.$$

$$\tilde{I}_k = \{1, 2, \dots, c\} - I_k\ ,$$

entonces $(U, v) \in M_{fc} \times \mathbb{R}^{cp}$ puede ser un mínimo global de J_m solamente si

$$I_k = \emptyset \Rightarrow u_{ik} = \frac{1}{\left[\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{2/(m-1)}\right]}\ . \tag{3.7a1}$$

O

$$I_k \neq \emptyset \Rightarrow u_{ik} = 0 \; \forall i \in \tilde{I}_k \; y \; \sum_{i \in I_k} u_{ik} = 1 \, . \tag{3.7a2}$$

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m} \quad \forall i \, . \tag{3.7b}$$

El algoritmo Fuzzy c-Means fue propuesto por Dunn [30] y generalizado por Bezdek [8].

3.3.1 Algoritmo Fuzzy C-Means (FCM):

FCM1: Fijar c, $2 \leq c \leq n$; seleccionar cualquier producto interno inducido por una norma sobre $\mathbb{R}^p$ y fijar m, $1 \leq m < \infty$. Inicializar $U^{(0)} \in M_{fc}$. Entonces para cada l, $l = 0, 1, 2, \ldots$:

FCM2: Calcular los c centroides $\{v_i^{(l)}\}$ utilizando (3.7b) y $U^{(l)}$.

FCM 3: Actualizar $U^{(l)}$ utilizando (3.7a) y $\left\{v_i^{(l)}\right\}$.

FCM 4: Comparar $U^{(l)}$ y $U^{(l+1)}$ mediante una norma de matrices: si $\left\|U^{(l+1)} - U^{(l)}\right\| \leq \varepsilon_L$ parar, en otro caso saltar a **FCM2**.

3.3.2 Conclusiones y observaciones

La función J_m está generalmente aceptada, aunque presenta una contradicción, como fue planteada en [31]: a medida que se aumente el valor del parámetro m, como los valores de las funciones de pertenencia se encuentran en el intervalo $[0, 1]$, el valor de la función objetivo es cada vez menor lo que se traduce en que para el mismo número de grupos, una partición más difusa tiene un menor error que una menos difusa.

Flores-Sintas [32] además refleja que no existe solamente el problema del exponente m, además el algoritmo no tiene un buen comportamiento cuando los grupos no se encuentran bien alejados y tienen tamaños desproporcionados. Otro aspecto a tener en cuenta sería la obtención de los valores de los grados de pertenencia. Para la deducción de éstos, dados por el FCM, Dunn [30] los obtiene aplicando los Multiplicadores de Lagrange e imponiendo la condición de que la suma de los grados de pertenencia de un elemento a todos los grupos sea 1. Para este enfoque, Flores-Sintas [33] muestra que esta condición es necesaria pero no suficiente, es decir, la expresión de los grados de pertenencia dados por el FCM son magnitudes que sumadas dan 1 pero no tienen por qué ser probabilidades. Además como explica Flores-Sintas [32] los grados de pertenencia obtenidos con el algoritmo FCM son grados de participación, mientras que en la formulación de la teoría de conjuntos difusos de Zadeh, el grado de pertenencia de un elemento en un conjunto difuso no depende del grado de pertenencia del mismo elemento en otro conjunto difuso, definido sobre el mismo universo. Por tanto, para que los grados de pertenencia definidos por el algoritmo FCM tengan sentido según la teoría desarrollada por Zadeh, no se les puede imponer la condición de que la suma de todos los grados de pertenencia de un elemento en todos los grupos sea 1. El grado de pertenencia de un elemento en un conjunto difuso debe ser independiente del grado de pertenencia del propio elemento en el resto de los conjuntos difusos. De esta manera se puede concluir que uno de los principales problemas en el algoritmo *Fuzzy C-means* surge producto de utilizar las particiones de Ruspini, por lo que en otros algoritmos se utiliza otro tipo de particiones como es el caso del Agrupamiento Posibilista que será analizado más adelante.

3.4 *Fuzzy C-means* basado en entropía

Primero generalizaremos el algoritmo FCM a FC(J, U), el cual recibe dos argumentos, una función objetivo J y una matriz de restricciones U.

3.4.1 Algortimo FC(J, M)

Generalización del algoritmo Fuzzy C-means con argumentos

FC1: Generar c centroides iniciales $\bar{v}_i$ $(i = 1, 2, 3, \dots, c)$.

FC2: Calcular: $\bar{U} = arg\ min_{U \in M_{fc}} J(U, \bar{v})$.

FC3: Calcular: $\bar{v} = arg\ min_{v} J(\bar{U}, v)$.

FC4: Si $\bar{U}$ o $\bar{v}$ son convergentes, parar; sino ir para el paso **FC2**.

De este modo, FC(J_1, M_c) es empleada para el algoritmo K-means y FC(J_m, M_{fc}) para *Fuzzy C-means*. Consideraremos otras formas de difusificar el algoritmo K-means.

Como se analizó anteriormente los métodos de Bezdek y Dunn introducen la no linealidad mediante los términos $({u_{ki}}^m)$ en el algoritmo K-means, a continuación analizaremos el uso de otras formas de no linealidad.

Los métodos de Bezdek y Dunn poseen otras características: como es la de suavizar la solución dura. Además, la solución difusa aproxima la dura en el sentido de que la solución difusa converge a la solución dura cuando $m \longrightarrow 1$. A grandes rasgos, se puede decir que la solución difusificada *regulariza* (aproxima) la solución dura.

Una regularización típica se realiza mediante la adición de una función de regularización. En el contexto actual, consideramos

$$J'(U, v) = J_{cm}(U, v) + kK(u), \qquad (k > 0), \tag{3.8}$$

en donde $K(u)$ es una función de regularización no lineal y k es un parámetro de regularización.

Analizaremos 2 funciones de regularización: una es una función de entropía y la otra es una función cuadrática

$$K(u) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log u_{ik}\,. \tag{3.9}$$

$$K(u) = \frac{1}{2} \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^2\,. \tag{3.10}$$

Nótese que ambas funciones son estrictamente convexas[2] (cóncava hacia arriba) y por lo tanto son capaces de difusificar la matriz de pertenencia U. Cuando utilizamos la primera forma, el algoritmo es nombrado regularización por entropía o método basado en entropía. La primera formulación de esta idea es llamada método de entropía máxima por Li y Mukaidono [34]; luego Miyamoto y Mukaidono [35] la han reformulado utilizando la idea de la regularización.

Luego, la siguiente función objetivo es utilizada para el **método basado en entropía**:

---

[2] Una función real f definida en un intervalo (o en cualquier subconjunto convexo de algún espacio vectorial) se llama **función convexa** o **cóncava hacia arriba**, si para dos puntos cualesquiera x e y con x, y que pertenecen a su dominio de definición C y cualquier $t \in [0, 1]$, se cumple: $f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$. En otras palabras, una función es convexa si y sólo si su epigrafo (el conjunto de puntos situados en o sobre el grafo) es un conjunto convexo.

Una **función estrictamente convexa** es aquella en donde se cumple que: $f(tx + (1 - t)y) < tf(x) + (1 - t)f(y)$, para cualquier $t \in (0, 1)$ y $x \neq y$.

$$J_{efc}(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}(d_{ik})^2 + k\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}\log u_{ik}\,, \qquad k>0\,. \tag{3.11}$$

De esta manera, el método basado en entropía utiliza el algoritmo FC(J_{efc}, M_{fc}). Las soluciones en los pasos FC2 y FC3 son como sigue:

$$u_{ki} = \frac{exp\left(-\frac{d_{ik}}{k}\right)}{\sum_{j=1}^{c} exp\left(-\frac{d_{jk}}{k}\right)}, \tag{3.12}$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}x_k}{\sum_{k=1}^{n} u_{ik}}\,. \tag{3.13}$$

Como se puede apreciar cuando $k \to 0$, la solución converge a la solución dura.

El principio de la entropía máxima ha sido estudiado en diferentes campos [36]. El método de Li y Mukaidono es un método de máxima entropía con la restricción de igualdad:

$$max\ -\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}\log u_{ik}, \tag{3.14}$$

sujeto a: $\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}(d_{ik})^2 = L\,.$

Es fácil ver que esta formulación es equivalente a la regularización por entropía cuando el valor *L* no es dado, pero puede ser cambiado como un parámetro. Particularmente, un *L* desconocido puede ser reemplazado por la regularización del parámetro k, mientras k es interpretado como el multiplicador de Lagrange en el método de máxima entropía. La idea es utilizar el concepto de la regularización, el cual es más adecuado para el agrupamiento difuso.

3.4.2 Adición de un término cuadrático

Ahora consideremos el segundo método, añadiendo el término cuadrático. De tal manera, la siguiente función objetivo es considerada:

$$J_{qfc}(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}(d_{ik})^2 + \frac{1}{2}k\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^2\,, \qquad k>0\,, \tag{3.15}$$

y es utilizado el algoritmo FC(J_{qfc}, M_{fc}).

La solución de FC3 es la misma que en el caso del método basado en entropía. En contraste, la solución de FC2 no tiene una forma simple como la anterior. Está dada por un algoritmo, el cual necesita un proceso de derivación complicado. Debido a su complejidad no será abordado en este trabajo.

Para abordar este tema en detalle se puede consultar la siguiente literatura: [37].

3.5 Agrupamiento posibilista PCM (*Possibilistic Clustering, Possibilistic C-means*)

Aunque a menudo es deseable, el carácter *relativo* de los grados de pertenencia probabilísticos, puede ser engañoso [38]. Altos valores de pertenencia del elemento en más de un grupo, puede dar la impresión de que el elemento es típico de ellos, pero esto no siempre es así. Consideremos por ejemplo el caso sencillo de 2 grupos que se muestra a continuación:

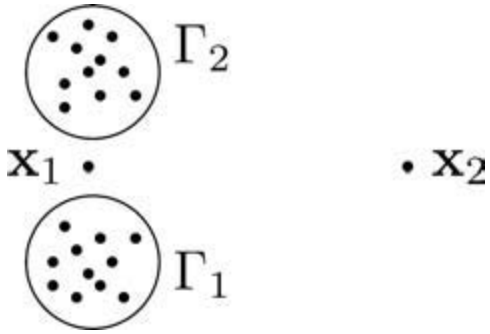

Fig. 3.2. Puntos en el plano.

El elemento x_1 tiene la misma distancia a ambos grupos y por lo tanto se le asigna un grado de pertenencia alrededor de los 0.5, esto se puede aceptar. Sin embargo, los mismos grados son asignados al elemento x_2, aunque este elemento se encuentra más lejos de estos 2 grupos y se debe considerar menos típico. Debido a la normalización, la suma de los grados de pertenencia debe ser 1. En consecuencia x_2 recibe altos grados de pertenencia a ambos grupos. Para una correcta interpretación de los grados de pertenencia se debe tener en cuenta que son más bien los grados de participación que de tipicidad, ya que el peso constante de 1 dado a un elemento debe ser distribuido sobre todos los grupos. Al eliminar la normalización en la restricción del FCM, se intenta lograr una asignación más intuitiva de los grados de pertenencia y evitar los efectos indeseables de la normalización. Se puede definir otro tipo de particiones, que se llaman particiones posibilistas:

$$M_{pc} = \left\{ V \in V_{cn} \mid \mu \in [0,1] \colon \sum_{j=1}^{n} \mu_{ij} > 0 \ \forall i \right\}. \tag{3.16}$$

La función objetivo J_m que minimiza las distancias al cuadrado entre los grupos y los elementos asignados, no sería adecuada para el agrupamiento difuso posibilista. La eliminación de la restricción de la normalización lleva a que el mínimo se alcance cuando $u_{ij} = 0 \ \forall i, j$, o lo que es lo mismo los elementos no están asignados a ningún grupo y todos los grupos están vacíos. Con el objetivo de evitar esta solución trivial (que también está prohibida por la restricción anterior), se introduce un término de penalización, lo que obliga a los grados de pertenencia a alejarse del 0. La función objetivo queda de la siguiente manera:

$$J_{pm}(U, v) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m (d_{ik})^2 + \sum_{i=1}^{c} n_i \sum_{k=1}^{n} (1 - u_{ik})^m \,, \tag{3.17}$$

donde $n_i > 0 \ \forall i$ [19]. El primer término lleva a la reducción de las distancias ponderadas, mientras que el segundo suprime la solución trivial, ya que esta suma premia valores altos de pertenencia (cercanos a 1), que hacen que la expresión $(1 - u_{ik})^m$ sea aproximadamente 0. En paralelo con el primer término, valores altos de pertenencia pueden ser esperados, especialmente para aquellos elementos que están cerca de sus grupos, ya que con un alto grado de pertenencia la distancia ponderada a un grupo más cercano es menor que a grupos más lejanos. Las constantes n_i para los grupos, son

utilizadas como valores de referencia que indican a qué distancia de un grupo, un elemento debe recibir mayores grados de pertenencia. Estas consideraciones marcan la diferencia con respecto a los enfoques de agrupamiento probabilístico. A diferencia de los agrupamientos probabilísticos donde cada elemento tiene un peso constante de 1, los posibilistas tienen que aprender los pesos de los elementos.

La fórmula para actualizar los grados de pertenencia se obtiene de igualar la derivada de J_{pm} a cero [19], análogo al algoritmo FCM, con el objetivo de buscar el mínimo global de la función:

$$u_{ij} = \frac{1}{1 + \left(\frac{{d_{ij}}^2}{n_i}\right)^{\frac{1}{m-1}}}. \tag{3.18}$$

Primero que todo, la ecuación de actualización muestra que los grados de pertenencia de un elemento x_j a un grupo i depende solamente de su distancia ${d_{ij}}^2$ al grupo. Pequeñas distancias se corresponden con altos grados de pertenencia mientras que grandes distancias resultan en bajos grados de pertenencia.

3.6 Algoritmo de Gustafson–Kessel (*Gustafson–Kessel*)

El algoritmo de Gustafson–Kessel [11] sustituye la distancia euclidiana por la distancia de Mahalanobis específica por grupo, a fin de adaptarse a distintos tamaños y formas de los grupos. Para el grupo i, la distancia de Mahalanobis se define como:

$$\left(d_{ij}\right)^2 = \left(x_j - v_i\right)^T S_i^{-1} \left(x_j - v_i\right), \tag{3.19}$$

donde S_i es la matriz de covarianza del grupo. Utilizando la distancia euclidiana como en los algoritmos anteriores es equivalente a asumir que $S_i = I, \forall i$, es decir, todos los grupos tienen la misma covarianza que es igual a la matriz identidad. Por lo tanto, sólo pueden identificar grupos esféricos como se analizó en FCM, pero no pueden identificar grupos que tienen diferentes formas y tamaños.

El algoritmo de Gustafson–Kessel modela cada grupo tanto por su centroide como por su matriz de covarianza. Así los centroides de los grupos son tuplas $V_i = (v_i, S_i)$ y tanto v_i como S_i van a ser determinadas. Las características de la matriz S_i (dimensión $p \times p$, donde p es el número de características) definida positiva, representa la forma del grupo i. Restricciones específicas pueden ser tomadas en cuenta, por ejemplo restringiendo la forma de los grupos a ser paralelos a los ejes, considerando solamente matrices diagonales. Los tamaños de los grupos también pueden ser controlados, si se conocen con antelación, mediante las constantes $\rho_i > 0$ exigiendo que $det(S_i) = \rho_i$. Por lo general, se asume que los grupos tienen el mismo tamaño y se pone $det(S_i) = 1$.

La función objetivo entonces es idéntica a la del FCM, utilizando como distancia la definida en este caso. Las actualizaciones de los centroides se mantiene igual y en el caso de los grados de pertenencia se utiliza la misma del algoritmo FCM, pero sustituyendo en el caso correspondiente la distancia euclidiana por la distancia específica para cada grupo. Las ecuaciones de actualización para las matrices de covarianza son:

$$S_i = \frac{{S_i}^*}{\sqrt[p]{det\ ({S_i}^*)}}, \qquad donde\ {S_i}^* = \frac{\sum_{j=1}^{n} u_{ij} \left(x_j - v_i\right)\left(x_j - v_i\right)^T}{\sum_{j=1}^{n} u_{ij}}. \tag{3.20}$$

Ellas son definidas como la covarianza de los elementos asignados al grupo i, modificadas para incorporar la información de las asignaciones difusas.

El algoritmo de Gustafson–Kessel intenta extraer mucha más información de los datos que los algoritmos basados en la distancia euclidiana. Es más sensible a la inicialización, por tanto se recomienda para la inicialización el uso de unas pocas iteraciones de los algoritmos FCM o PCM en dependencia del tipo de partición. En comparación con FCM y PCM, este algoritmo exhibe mayor costo computacional debido a las inversiones de las matrices.

3.7 Fuzzy Shell Clustering

Una de las áreas de aplicación de los algoritmos de agrupamiento difusos es el análisis y reconocimiento de imágenes. Variantes del FCM y el PCM han sido propuestas para la detección de líneas, círculos o elipses sobre el conjunto de elementos, que corresponden a subestructuras más complejas de los datos. Los algoritmos *Shell Clustering* [39] extraen prototipos que tienen una naturaleza diferente a los elementos. Estos algoritmos necesitan modificar la definición de distancia entre un elemento y un prototipo y sustituir la distancia euclidiana por otras distancias. Por ejemplo el algoritmo c-variedades difusas (*Fuzzy C-variety, FCV*) fue desarrollado para el reconocimiento de líneas, planos o hiperplanos (ver Figura 3.3). Cada grupo es un subespacio afín, caracterizado por un *punto* y un conjunto de vectores unitarios ortogonales $C_i = (v_i, e_{i1}, e_{i2}, \dots, e_{iq})$, donde q es la dimensión del subespacio afín. La distancia entre un elemento x_j y el grupo i es definida entonces como:

$$\left(d_{ij}\right)^2 = \left\|x_j - v_i\right\|^2 - \sum_{l=1}^{q} (x_j - v_i)^T e_{il} \,. \tag{3.21}$$

Otras variantes similares de FCM y de PCM incluyen el algoritmo *c-elliptotypes difuso adaptativo*, (AFCE) que asigna segmentos de líneas disjuntos a diferentes grupos (ver Figura 3.4). Los contornos de círculos pueden ser detectados por el algoritmo *c-shells difuso* y el *c-spherical shells difuso*. El algoritmo *c-ellipsoidal shells* difuso es capaz de reconocer objetos con límites en forma de círculo (elipses). El algoritmo *c-quadric shells* difuso (FCQS) es además capaz de reconocer hipérbolas, parábolas o grupos lineales. Su flexibilidad se puede observar en las Figuras 3.5 y 3.6. Las técnicas de agrupamiento *Shell* también se han extendido a las estructuras no lisa (*non-smooth*) como rectángulos y otros polígonos. Las Figuras 3.7 y 3.8 ilustran los resultados obtenidos con el algoritmo *c-rectangular* difuso (FCRS) y el algoritmo *c-2-rectangular shells* difuso (FC2RS), respectivamente. Se puede abordar con mayor profundidad en este tema, en la siguiente literatura: [40, 41].

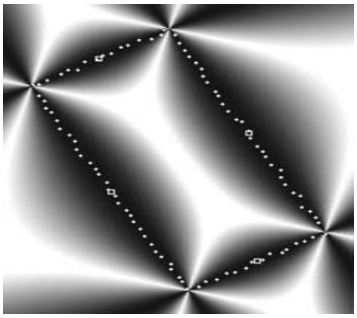

Fig. 3.3. Análisis FCV.

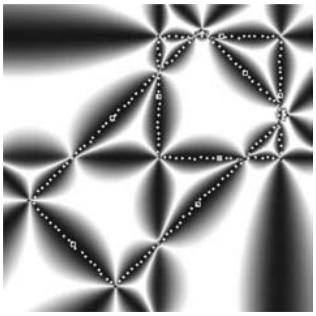

Fig. 3.4. Análisis AFCE.

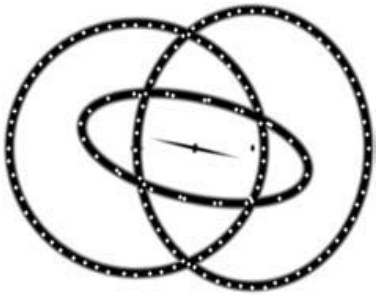

Fig. 3.5. Análisis FCQS.

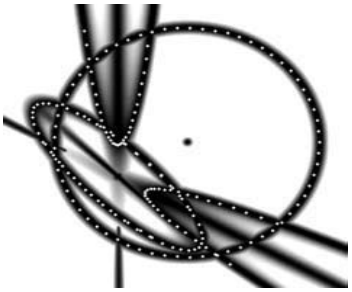

Fig. 3.6. Análisis FCQS.

Fig. 3.7. Análisis FCRS.

Fig. 3.8. Análisis FC2RS.

3.8 Agrupamiento difuso basado en *Kernel*

Las variantes *Kernel* de los algoritmos de agrupamiento difuso, además de modificar la función de distancia, permiten manejar datos que no sean vectoriales, tales como secuencias, árboles o grafos, sin necesidad de modificar completamente los algoritmos. De manera más general, los métodos basados en *Kernel* pueden ser aplicados independientemente de la naturaleza de los datos, sin necesidad de adaptar el algoritmo. A continuación analizaremos los algoritmos basados en *Kernel*. Los elementos pueden ser vectoriales o no, por lo que los denotaremos σ_j en lugar de x_j.

3.8.1 Principios

Los métodos kernel están basados en una transformación de la representación de los datos ϕ: $\chi \rightarrow \mathcal{F}$ donde χ denota el espacio de entrada y $\mathcal{F}$ es llamado el espacio de características. $\mathcal{F}$ usualmente es de grandes dimensiones o incluso infinita y sólo está limitado a ser un espacio de Hilbert, es decir, a disponer de un producto escalar. El segundo principio de los métodos kernel es que los elementos no se manejan directamente en el espacio de características, lo que podría llevar a altos costos

computacionales debido a la dimensión del espacio, ellos solamente son manejados mediante su producto escalar que se calcula utilizando la representación inicial. Con tal objetivo, se utiliza una función kernel: K: $\chi * \chi \rightarrow \mathbb{R}$, tal que $\forall \sigma_i, \sigma_j \in \chi, \langle \phi(\sigma_i), \phi(\sigma_j) \rangle = K(\sigma_i, \sigma_j)$. Así, no es necesario conocer explícitamente la función ϕ. El producto escalar en el espacio de características solamente depende de la representación inicial.

Los métodos *kernel* son algoritmos escritos solamente en términos del producto escalar entre los elementos. El enriquecimiento de la representación de los datos viene dado por el uso del producto escalar sobre la base de una transformación implícita de los datos, en lugar de utilizar solamente la distancia euclidiana. La posibilidad de aplicar el algoritmo para datos no vectoriales solamente depende de la disponibilidad de una función $K: \chi * \chi \rightarrow \mathbb{R}$ teniendo las propiedades de un producto escalar [42].

3.8.2 Agrupamiento difuso basado en Kernel

La variante Kernel de los agrupamientos difusos [43] consiste en transponer la función objetivo al espacio de características, es decir, aplicándole la transformación de los datos ϕ. Los centroides de los grupos entonces pertenencen al espacio de características, por lo tanto ellos se denotan $v_i^{\phi}, \forall i$ ($v_i^{\phi} \in \mathcal{F}$) y se buscan en forma de combinaciones lineales de las transformaciones de los datos, como:

$$v_i^{\phi} = \sum_{j=1}^{n} a_{ij} \phi(\sigma_j) \,. \tag{3.22}$$

Esta formulación es coherente con la solución obtenida con el FCM. La optimización debe entonces proporcionar los valores de a_{ij}, junto con los grados de pertenencia. Debido a la forma anterior de los centroides, la distancia euclidiana entre un elemento y el centroide en el espacio de características puede ser computada mediante [43]:

$$d_{\phi_{ij}}^2 = \left\| \phi(\sigma_j) - v_i^{\phi} \right\|^2 = K_{jj} - 2 \sum_{r=1}^{n} a_{ir} K_{jr} + \sum_{r,s=1}^{n} a_{ir} a_{is} K_{rs} \,. \tag{3.23}$$

Donde denotamos $K_{ij} = K(\sigma_i, \sigma_j) = \langle \phi(\sigma_i), \phi(\sigma_j) \rangle$. Luego, la función objetivo queda de la siguiente forma:

$$J_m^{\phi} = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij})^m \left(K_{jj} - 2 \sum_{r=1}^{n} a_{ir} K_{jr} + \sum_{r,s=1}^{n} a_{ir} a_{is} K_{rs} \right). \tag{3.24}$$

Las condiciones de minimización dan lugar a las siguientes ecuaciones de actualización:

$$u_{ij} = \frac{1}{\sum_{l=1}^{c} \left(\frac{d_{\phi_{ij}}^2}{d_{\phi_{lj}}^2} \right)^{\frac{1}{m-1}}}, \quad a_{ij} = \frac{(u_{ij})^m}{\sum_{r=1}^{n} (u_{ir})^m}, \quad v_i^{\phi} = \frac{\sum_{j=1}^{n} (u_{ij})^m \phi(\sigma_j)}{\sum_{j=1}^{n} (u_{ij})^m}. \tag{3.25}$$

Como se puede apreciar, las ecuaciones de actualización al igual que la función objetivo, se pueden expresar únicamente en términos de la función *Kernel*, es decir, en términos del producto escalar. La ecuación anterior acerca de los grados de pertenencia muestra que tiene la misma forma que en el algoritmo FCM, sustituyendo la distancia euclidiana por la distancia en el espacio de características, como se definió $d_{\phi_{ij}}^2$. La expresión de los centroides es comparable también al caso del FCM, como la media ponderada de los datos. La diferencia es que los centros de los grupos pertenecen al espacio de

características y no tienen una representación explícita, solamente son conocidos los coeficientes de peso.

Existen otras variantes para la *kernelization* del FCM, como por ejemplo una variante propuesta por Zhang y Chen en [44, 45]. Este último sólo considera el *Kernel Gaussiano* $K(\sigma_i, \sigma_j) = e^{\left(\frac{-d(\sigma_i,\sigma_j)^2}{\sigma^2}\right)}$ y explota sus propiedades para simplificar el algoritmo. Más preciso, utiliza la hipótesis de que el centroide del grupo se puede buscar explícitamente en el espacio de entrada ($v_i \in \chi$) y considera su transformación al espacio de características $\phi(v_i)$. La función objetivo entonces se convierte en:

$$J_m^{\phi} = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m \left\|\phi(v_i) - \phi(\sigma_j)\right\|^2 = 2\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m \left(1 - e^{\left(\frac{-d(v_i,\sigma_j)^2}{\sigma^2}\right)}\right), \tag{3.26}$$

aprovechando el hecho de que el *Kernel Gaussiano* lleva a $d_{\phi}^2(\sigma_i, \sigma_j) = K(\sigma_i, \sigma_i) + K(\sigma_j, \sigma_j) - 2K(\sigma_i, \sigma_j) = 2\left(1 - K(\sigma_i, \sigma_j)\right)$.

Se puede apreciar que para la aplicación de un método *kernel*, se necesita la selección de una función *Kernel* y sus parámetros, lo cual puede ser difícil o complicado. Uno de los principales motivos que llevan a Huang, et al. [46] a proponer el algoritmo Multiple Kernel Fuzzy C-means (MKFC) basado en el FCM, donde se combinan un conjunto de funciones *kernel* y además se ajustan automáticamente los pesos asociados a cada una de estas funciones.

El problema de la selección de una función *Kernel* y sus parámetros, puede ser similar al problema de la selección de características (*feature selection*) y la elección de la representación de los datos, en el caso de los métodos no basados en funciones *Kernel*. En [47] se realiza un análisis comparativo de los algoritmos de agrupamiento difuso basados en funciones *kernel* y los algoritmos de agrupamiento difuso, donde se evidencian algunas de estas ideas.

3.9 Agrupamiento con ruido (*Noise Clustering, NC*)

El algoritmo de agrupamiento con ruido fue propuesto inicialmente por Davé en [48] y se extendió posteriormente [49, 50]. Consiste en añadir además del parámetro c acerca de la cantidad de grupos deseados, un nuevo grupo ruidoso, con el objetivo de agrupar los elementos que están mal representados por grupos normales, tales como elementos ruidosos o valores atípicos. El centroide del grupo ruidoso se considera que se encuentra a una distancia constante δ, de todos los elementos del conjunto. Esto significa que todos los elementos tienen a priori la misma *probabilidad* de pertenecer al grupo ruidoso. Durante el proceso de optimización, esta *probabilidad* se adapta como una función de probabilidad según la cual, los elementos pertenecen a grupos normales. El grupo ruidoso, es entonces, introducido en la función objetivo, como cualquier otro grupo:

$$J_{mnc}(U, v) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m (d_{ik})^2 + \sum_{k=1}^{n}\delta^2\left(1 - \sum_{i=1}^{c}u_{ik}\right)^m . \tag{3.27}$$

El término añadido es similar a los términos de la primera suma: la distancia a los centroides de los grupos es sustituida por δ y el grado de pertenencia a los grupos se define como el complemento a 1 de la suma de todos los grados de pertenencia a los grupos *normales*. Esto en particular, implica que los elementos atípicos tienen bajos grados de pertenencia a los grupos estándar y alto grado de pertenencia al grupo ruidoso, lo que hace que se reduzca la influencia sobre los grupos *normales*. Como el PCM, el enfoque de agrupamiento ruidoso relaja la restricción de normalización, donde los grados de pertenencia a los grupos normales deben sumar 1.

Si comparamos PCM con el presente algoritmo (NC) se puede observar que los algoritmos son idénticos en el caso de un solo grupo, con δ^2 correspondiente con n [49, 50]. En el caso de $c > 1$, la diferencia es que PCM considera un n_i por grupo, mientras que en el NC se define un simple parámetro. Esto significa que PCM posee la ventaja de tener un grupo ruidoso por cada uno de los grupos *normales*, mientras que NC solamente posee uno. Como consecuencia, los grados de pertenencia sobre el grupo ruidoso son diferentes para los 2 métodos: en el caso del PCM, para cada grupo ruidoso es el complemento a 1 de la pertenencia al grupo *normal* asociado. En el agrupamiento ruidoso, como solamente se tiene un grupo ruidoso el grado de pertenencia a ese grupo es el complemento a la suma de todos los grados de pertenencia.

Otra diferencia entre PCM y NC viene dada por el hecho de que la función de costo del PCM puede descomponerse en c términos independientes (uno por cada grupo), mientras que en el NC no se puede hacer una descomposición. Esta descomposición es una de las razones por la cual PCM lleva a coincidir grupos. Así, en [51] interpretan NC como un FCM robusto, mientras que PCM se comporta como c algoritmos NC independientes.

La función objetivo anterior requiere del establecimiento del parámetro δ. En el algoritmo NC inicial es:

$$\delta^2 = \lambda \frac{1}{cn} \sum_{k=1}^{n} \sum_{i=1}^{c} (d_{ik})^2 \,. \tag{3.28}$$

Es decir, su valor al cuadrado es una parte de la media de los cuadrados de las distancias entre los elementos y los centroides. Con el parámetro definido por el usuario λ, para determinar la proporción: mientras más pequeño sea el valor, mayor será la proporción de elementos que serán considerados elementos atípicos.

Los agrupamientos ruidosos han sido generalizados para permitir la definición de varios δ y definir una escala de ruido por cada grupo [49, 50].

3.10 Modelo de mezcla de densidades

A continuación analizaremos el modelo de mezcla de densidades que es frecuentemente empleado para la clasificación supervisada y la no supervisada [52, 53]. Para este propósito, utilizaremos en esta sección términos de las probabilidades y las estadísticas.

Aunque las funciones de probabilidad de densidad para la mayoría de las distribuciones estándar son unimodal, se nota que el agrupamiento debería manejar distribuciones multimodal. Por el momento, supongamos que se cuentan con muchos datos y el histograma tiene dos modos de valores máximos. Aparentemente, el histograma está representado por la mezcla de dos densidades de distribuciones unimodal. Por lo tanto la probabilidad de densidad es:

$$p(x) = \alpha_1 p_1(x) + \alpha_2 p_2(x) \,, \tag{3.29}$$

donde α_1 y α_2 son números no negativos tal que: $\alpha_1 + \alpha_2 = 1$. Por lo general ambas distribuciones son normal:

$$p_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_i)^2}{2{\sigma_i}^2}}, i = 1, 2 \,. \tag{3.30}$$

Supóngase además que se tiene una buena estimación de los parámetros α_i, μ_i, σ_i, $i = 1, 2$. Luego de tener una buena aproximación de la distribución de la mezcla se puede resolver el problema del agrupamiento utilizando la fórmula de Bayes para la probabilidad a posteriori.

Supongamos $P(X|C_i)$ y $P(C_i)$ $(i = 1, 2, 3, \dots c)$ sea la probabilidad condicional del evento X dado que la clase C_i ocurre y la probabilidad a priori de la clase C_i respectivamente. Suponemos que necesariamente ocurre uno de los C_i $(i = 1, 2, 3, \dots c)$. La fórmula de Bayes se utiliza para determinar la clase de X:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{\sum_{j=1}^{c} P\left(X\middle|C_j\right)P(C_j)}. \tag{3.31}$$

Luego aplicamos la fórmula anterior para el ejemplo en cuestión:

$$P(X) = P(a < x < b). \tag{3.32}$$

$$P(C_i) = \alpha_i. \tag{3.33}$$

$$P(X|C_i) = \int_a^b p_i(x)dx\,, (i = 1, 2)\,. \tag{3.34}$$

Entonces se tiene

$$P(C_i|X) = \frac{\alpha_i \int_a^b p_i(x)dx}{\sum_{j=1}^{c} \alpha_j \int_a^b p_j(x)dx}. \tag{3.35}$$

Supongamos que tenemos la observación y. Tomando dos números a y b tal que $a < y < b$, tenemos la probabilidad de la clase C_i dado X, mediante la fórmula anterior. Tomando el límite cuando $a \to y$ y b$\to y$, tenemos la probabilidad de la clase C_i dado el objeto y:

$$P(C_i|y) = \frac{\alpha_i p_i(y)}{\sum_{j=1}^{2} \alpha_j p_j(y)}. \tag{3.36}$$

Esto nos da la probabilidad de asignar una observación a cada una de las clases, por lo que el problema del agrupamiento es resuelto, pero en lugar de grados de pertenencia, se tiene la probabilidad de pertenencia a una clase.

La fórmula (3.36) es inmediatamente generalizada al caso de c clases. El problema es cómo obtener buenas estimaciones de los parámetros. Por consiguiente, el algoritmo EM [53] pudiera ser utilizado con este propósito.

4 Índices de validación de agrupamientos difusos

La $c - partición\ difusa$ proporcionada por un algoritmo de agrupamiento difuso tiene como objetivo identificar la estructura presente en el conjunto de elementos iniciales. En estas particiones, al igual que en el caso de las $c - particiones\ duras$, los objetos asignados al mismo grupo son más similares entre sí que con otros objetos que pertenecen a grupos distintos. Sin embargo, aunque el entorno es difuso, el objetivo de la clasificación es generar una $c - partición\ difusa$ bien definida que sea lo más parecida posible a la estructura natural de los datos. Por lo tanto, una pregunta difícil es ¿cómo una partición se adapta a la estructura desconocida de los datos? Este problema requiere de un análisis para la validación de los grupos, utilizando algún criterio para la validación. Este problema es conocido en la literatura como: problema de la validación de los grupos (*cluster validity*).

Dado que la mayoría de los métodos difusos presumen que el número c de grupos es conocido, un criterio de validación para encontrar el c óptimo que pueda describir completamente la estructura de los datos, se convierte en el tema más estudiado en la validación de grupos.

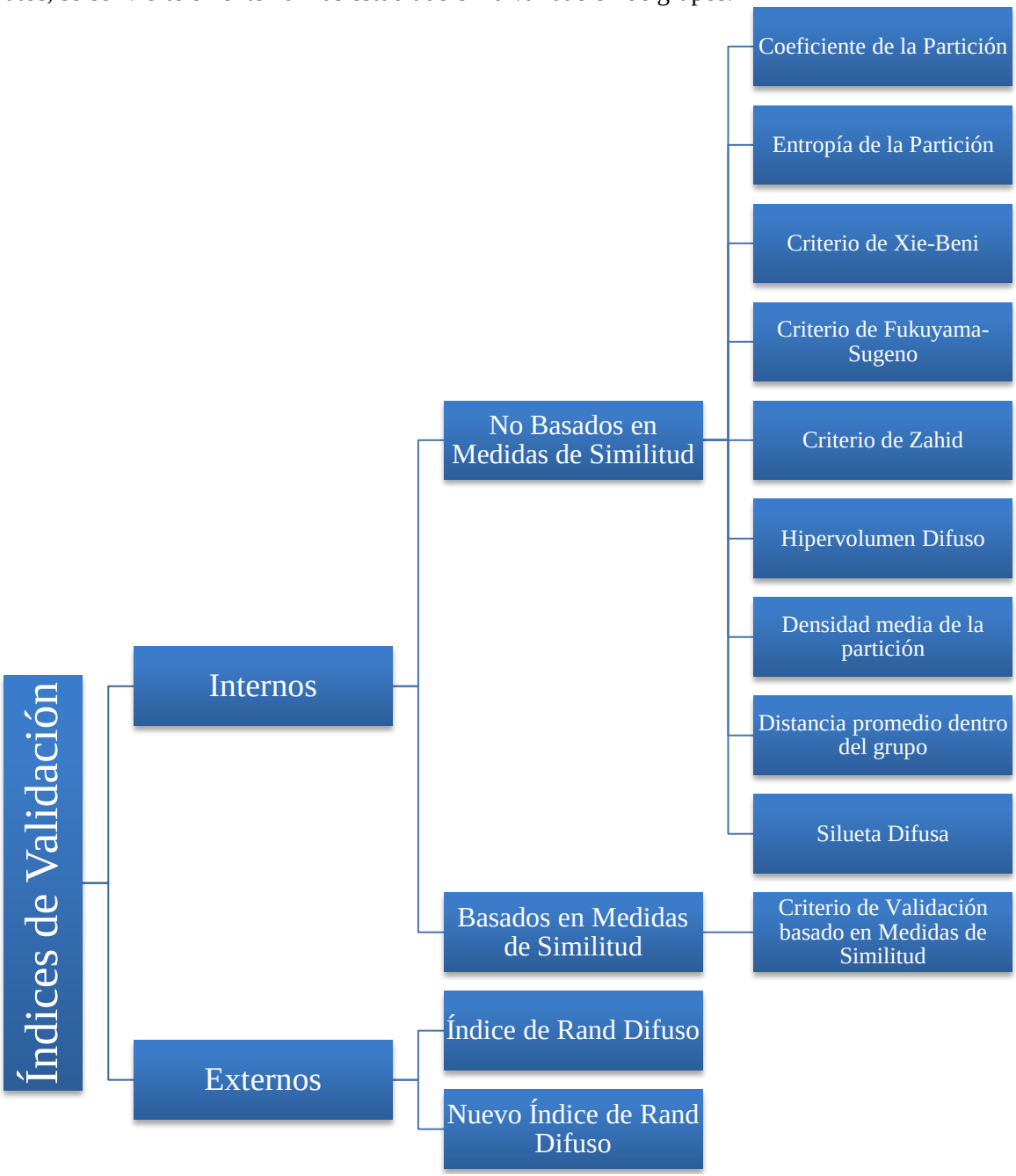

Fig. 4.1. Taxonomía de los principales índices de validación de agrupamientos difusos.

En la literatura podemos encontrar una gran variedad de índices de validación para estructuraciones difusas. Estos índices se pueden dividir en internos y externos. Los índices internos se basan solamente en la información de la propia estructuración difusa, no necesitan información externa. Por su parte, los índices externos utilizan como patrón para compararse una estructuración difusa específica, la cual es obtenida a partir de una información previa de los datos, donde muchas veces este patrón es visto como la estructuración *real* (*ground-truth*). Se puede destacar que se han desarrollado muchos índices

internos, a diferencia de los externos que han sido poco abordados. En este trabajo se introduce una taxonomía con los diferentes índices de validación existentes, la cual se representa en la Figura 4.1.

4.1 Índices internos

4.1.1 Coeficiente de la partición (Partition Coefficient)

El primer criterio de validación de agrupamientos difusos, asociado con el FCM, fue introducido por Bezdek [54, 55] y definido de la siguiente forma:

$$PC(U) = \sum_{i=1}^{c}\sum_{k=1}^{n} {u_{ik}}^2 , \tag{4.1}$$

donde U es una $c - partición\ difusa$, resultado de un algoritmo de agrupamiento difuso. Este criterio de validación se llama Coeficiente de la Partición (*Partition Coefficient*).

4.1.2 Entropía de la partición (Partition Entropy)

Bezdek también definió otro criterio de validación, llamado Entropía de la Partición (*Partition Entropy*):

$$PE(U) = -\frac{\sum_{i=1}^{c}\sum_{k=1}^{n}[u_{ik}\, log_2(u_{ik})]}{n}. \tag{4.2}$$

Ambos criterios de validación comparten las siguientes propiedades:

$$\frac{1}{c} \le PC \le 1 \Leftrightarrow 0 \le PE \le \log_2 c,$$

$$PC = 1 \Leftrightarrow PE = 0 \Leftrightarrow U\ c - partición\ dura. \tag{4.3}$$

$$PC = \frac{1}{c} \Leftrightarrow PE = \log_2 c \Leftrightarrow U\ c - partición\ difusa\ .$$

Estas propiedades muestran que si PC es cercana a 1 y PE a 0, U se parece más a una $c - particion\ dura$. Esta situación ambigua corresponde a la última propiedad donde cada objeto es asignado a cada uno de los c grupos con grado de pertenencia $1/c$. Por lo tanto, un agrupamiento válido se obtiene mediante la maximización de PC (o minimizando a PE) $para\ c = 2, 3, \dots, c_{max}$. Adicionalmente, para la evaluación de las $c - particiones\ difusas$, PC y PE utilizan solamente los grados de pertenencia u_{ik}, es decir, la propiedad de la matriz difusa U, independiente de la estructura de los datos.

Para tener en cuenta al mismo tiempo las propiedades de los grados de pertenencia difusa y la estructura de los datos en sí, han sido propuestos otros criterios de validación por Xie-Beni [56] y Fukuyama-Sugeno [57].

4.1.3 El criterio de validación de Xie-Beni se define mediante:

$$F_{XB}(U,V:X) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n}[{u_{ik}}^{m}\|x_k - v_i\|^2]}{n(\min_{i,j}\|v_i - v_j\|)} = \frac{J_m(U,V:X)}{nSep(V)}\ . \tag{4.4}$$

Donde $J_m$es la función objetivo del algoritmo FCM y $Sep(V)$ es una medida de separación entre los centroides de los grupos. Esta medida es considerada una medida de la compacidad. Un valor pequeño de F_{XB}, significa que estamos en presencia de una $c - partición\ difusa$ donde los grupos son compactos y bien separados. Por consiguiente, la mejor $c - partición\ difusa$ se obtiene mediante la minimización de F_{XB} con respecto a $c = 2, 3, \dots, c_{max}$.

4.1.4 El criterio de validación de Fukuyama-Sugeno está definido por:

$$F_{FS}(U,V:X) = \sum_{i=1}^{c}\sum_{k=1}^{n}[{u_{ik}}^{m}\|x_k - v_i\|^2] - \sum_{i=1}^{c}\sum_{k=1}^{n}[{u_{ik}}^{m}\|v_i - \bar{v}\|^2]. \tag{4.5}$$

$$= J_m(U,V:X) - J_m(K,V:X)\ .$$

Donde $\bar{v}$ es la media de todo el conjunto de centroides ($\bar{v} = \sum_{i=1}^{c} v_i/c$), J_m es la medida de compacidad y K_m es la medida de separación entre los centroides de los grupos y la media $\bar{v}$. El mínimo de F_{FS} corresponde a una $c - partición\ difusa$ con grupos compactos y bien separados.

4.1.5 El criterio de validación de Zahid [58]

El siguiente criterio de validación combina información acerca de la separación difusa y la compacidad difusa y toma en consideración las propiedades de la matriz de grados de pertenencia (U) y la estructura de los datos. La compacidad difusa, mide la variación manifestada por la concentración de los objetos que pertenecen a un mismo grupo alrededor del centroide del grupo. La separación difusa representa la separación (distanciamiento) entre los centroides de los grupos.

La proporción $separación\ difusa/compacidad\ difusa$ es medida utilizando dos funciones. La primera función, denotada SC_1, calcula el cociente considerando las propiedades geométricas de la estructura de los datos y los grados de pertenencia. Por otro lado, la segunda función llamada SC_2, utiliza solamente las propiedades de la matriz difusa U.

$$SC = SC_1(U,V:X) - SC_2(U). \tag{4.6}$$

$$SC_1(U,V:X) = \frac{[\sum_{i=1}^{c}\|v_i - \bar{v}\|^2]/c}{\sum_{i=1}^{c}(\sum_{k=1}^{n}(u_{ik})^m\|x_k - v_i\|^2/n_i)}, \tag{4.7}$$

donde n_i es la cardinalidad del grupo i y se define de la siguiente forma:

$$n_i = \sum_{k=1}^{n} u_{ik}. \tag{4.8}$$

$$SC_2(U) = \frac{\sum_{i=1}^{c-1}\sum_{r=i+1}^{c}\left(\sum_{k=1}^{n}\left(\min\left(u_{ik},u_{jk}\right)^2\right)/\sum_{k=1}^{n}\min\left(u_{ik},u_{jk}\right)\right)}{\sum_{k=1}^{n}\left(\underset{1\le i\le c}{max}\, u_{ik}\right)^2 \Big/ \sum_{k=1}^{n}\underset{1\le i\le c}{max}\, u_{ik}}. \tag{4.9}$$

4.1.6 Hipervolumen difuso [59]

Inspirado en el principio de que $buenos$ grupos no son difusos, a pesar de que estamos en un ambiente difuso, Gath y Gea propusieron un índice de validación basado en el criterio de hipervolumen y la densidad de grupos difusos. Específicamente, el Hipervolumen difuso (FVH por si siglas en inglés) viene dado por:

$$FVH = \sum_{i=1}^{c}[\det(F_i)]^{1/2}\,. \tag{4.10}$$

Donde F_i es la matriz de covarianza difusa del $i-ésimo$ grupo, definida mediante:

$$F_i = \frac{\sum_{j=1}^{n} h\left(i\middle|x_j\right)\left(x_j - v_i\right)\left(x_j - v_i\right)^T}{\sum_{j=1}^{n} h\left(i\middle|x_j\right)}. \tag{4.11}$$

Bajo la perspectiva de la estimación de máxima verosimilitud, $h\left(i\middle|x_j\right)$ es la probabilidad (posterior) de seleccionar el $i-ésimo$ grupo dado el objeto x_j. Cuando es seleccionado el valor de $m = 2$ (exponente de ponderación FCM), esta probabilidad se acerca al grado de pertenencia del objeto x_j al grupo i [59]. De acuerdo a esto, la matriz de covarianza puede ser escrita de la siguiente forma:

$$F_i = \frac{\sum_{j=1}^{n} u_{ij}\left(x_j - v_i\right)\left(x_j - v_i\right)^T}{\sum_{j=1}^{n} u_{ij}}. \tag{4.12}$$

Una $c-partición\ difusa$ se espera que tenga valor bajo de FVH si es una partición compacta.

4.1.7 Densidad media de la partición (Average Partition Density, APD) [59]

$$APD = \frac{1}{c}\sum_{i=1}^{c}\frac{R_i}{[\det(F_i)]^{1/2}}, \tag{4.13}$$

con R_i:

$$R_i = \sum_{j} u_{ij}\,, \forall j\ tal\ que\ \left(x_j - v_i\right)^T F_i^{-1}\left(x_j - v_i\right) < 1\,, \tag{4.14}$$

donde R_i es la suma de los valores de pertenencia de aquellos elementos que se encuentran dentro de una hiperesfera. Debido a que grupos difusos compactos, proporcionan valores pequeños de $[\det(F_i)]^{1/2}$ y grandes valores de R_i, es fácil concluir que buenas particiones difusas se caracterizan por tener valores grandes de APD.

4.1.8 Distancia promedio dentro del grupo (Average within-cluster distance, AWCD) [60]

Otro índice de validación difuso es la Distancia Promedio Dentro del Grupo, dado por:

$$AWCD = \frac{1}{c}\sum_{i=1}^{c}\frac{\sum_{j=1}^{n} u_{ij}^{m}\left\|x_j - v_i\right\|_A^2}{\sum_{j=1}^{n} u_{ij}^{m}}, \tag{4.15}$$

donde $\|.\|_A$ y m se refieren a la misma norma y exponente de ponderación utilizado en el algoritmo de agrupamiento difuso, respectivamente. Esta medida es el valor medio de las *distancias dentro del grupo* calculado sobre todos los grupos. La *distancia dentro del grupo* de un grupo en específico, a su vez, viene dada por la media ponderada de las distancias entre todos los objetos y el centroide del grupo, con cada distancia pesada por el valor de pertenencia de cada objeto al correspondiente grupo.

4.1.9 Extensión difusa del índice de la silueta (Fuzzy silhouette)

Silueta dura (*Crisp silhouette, CS*) [61, 62]

En el contexto de las particiones duras producidas por un algoritmo basado en prototipos (centroides, ejemplo K-means), este tipo de algoritmos se basan en la idea de que si el objeto o_j pertenece al grupo p, entonces se encuentra más cerca del prototipo p (v_p) que del resto de los centroides. En el caso más general del contexto de las particiones difusas, por otro lado, esto significa que el grado de pertenencia del objeto o_j al $p-ésimo$ grupo es más grande que los grados de pertenencia de este mismo objeto a cualquiera de los otros grupos.

Sea la distancia promedio del objeto o_j a todos los objetos que pertenecen al grupo p, denotada por a_{pj}. Adicionalmente, sea la distancia promedio de este objeto a todos los objetos que pertenecen a otro grupo q, $q \neq p$, llamada d_{qj}. Finalmente, sea b_{pj} el mínimo de todos los d_{qj} computados sobre $q = 1, 2, \ldots, c, q \neq p$, que representa la disimilitud del objeto o_j a su más cercano grupo. Entonces la silueta del objeto o_j se define como:

$$s_j = \frac{b_{pj} - a_{pj}}{\max\{a_{pj}, b_{pj}\}}, \tag{4.16}$$

donde el denominador es usado solamente como término de normalización. Claramente, el más alto s_j, corresponde a la mejor asignación del objeto o_j al grupo p. En el caso en que el grupo p está constituido por el elemento o_j solamente, entonces la silueta de este elemento se define como $s_j = 0$. Esto evita que la Silueta dura, que se define como el promedio de los s_j, $j = 1, 2, \ldots, N$,

$$CS = \frac{1}{N}\sum_{j=1}^{N} s_j, \tag{4.17}$$

encuentre la solución trivial $c = N$, donde cada elemento del conjunto de datos se encuentra solo formando un grupo. De esta forma, la mejor partición se encuentra cuando CS se maximiza, lo cual implica que se está minimizando la distancia $intra - cluster$ (a_{pj}) mientras se maximiza la distancia $inter - cluster(b_{pj})$.

Silueta difusa [63]

En el caso de la Silueta dura, no se hace uso de la matriz de pertenencia. La matriz de la partición difusa (matriz de pertenencia), $P = \left[u_{ij}\right]_{cxN}$ es utilizada solamente para imponer una partición dura sobre los elementos $\tilde{P} = \left[\tilde{u}_{ij}\right]_{cxN}$, para la cual la medida CS puede ser aplicada. Específicamente, $\tilde{P}$ es tal que, $\tilde{u}_{ij} = 1$ si $i = arg_{max\,l}\{u_{lj}\}$ y $\tilde{u}_{ij} = 0$ en otro caso. Por consiguiente CS no puede ser capaz de

discriminar entre grupos solapados (incluso si estos grupos tienen sus propias regiones con mayor densidad de datos) ya que deja de lado la información contenida en la matriz de la partición difusa P sobre los grados de pertenencia, para los cuales los grupos se superponen unos con otros. Esta información puede ser utilizada para revelar las regiones con alta densidad de datos, haciendo hincapié en la importancia de los objetos concentrados en la vecindad de los centroides, mientras se reduce la importancia de los objetos que se encuentran en áreas solapadas. Para ello, se define, un criterio generalizado de la Silueta, llamado Silueta difusa (*Fuzzy Silhouette*):

$$FS = \frac{\sum_{j=1}^{N}\left(u_{pj} - u_{qj}\right)^{\alpha} s_j}{\sum_{j=1}^{N}\left(u_{pj} - u_{qj}\right)^{\alpha}}, \tag{4.18}$$

donde s_j es la silueta del objeto o_j de acuerdo al criterio de la Silueta dura, u_{pj} y u_{qj} son el primero y segundo valores más altos de la $j-ésima$ columna de la matriz de pertenencia, respectivamente y $\alpha \geq 0$ es un coeficiente de ponderación. Claramente, cuando el exponente α se aproxima a 0 por arriba, la medida FS se asemeja a la medida CS. Por el contrario, al aumentar α, FS se aleja de CS al disminuir la importancia relativa de los objetos en áreas de solapamiento. En consecuencia, un aumento de α tiende a enfatizar el efecto de revelar regiones pequeñas con mayor densidad de datos (subgrupos), si es que existen. Este efecto puede ser particularmente útil, por ejemplo, cuando se trata de conjunto de datos contaminados por ruidos.

Se pueden apreciar diferencias con respecto a la Silueta dura. La ecuación de FS se diferencia de CS, ya que FS es un promedio pesado de los valores de CS. El peso de cada término está determinado por la diferencia entre los dos más grandes grados pertenencia del correspondiente objeto (el más grande restándole el segundo más grande). De esta manera, un objeto cerca de la vecindad de un centroide tiene más importancia que uno localizado en un área de solapamiento de grupos (donde los grados de pertenencia de los objetos para dos o más grupos son similares).

Una pregunta que puede surgir es, por qué no definir la silueta difusa como $\frac{1}{N}\sum_{j=1}^{N}\left(u_{pj} - u_{qj}\right)^{\alpha}$, debido a que una buena partición difusa se espera que sea de tal manera que cada objeto tiene un alto grado de pertenencia a un grupo difuso en específico y bajos grados de pertenencia al resto de los grupos. El problema es que este requisito puede ser fácilmente alcanzado por un algoritmo de agrupamiento difuso, incluso cuando el número de grupos difusos c, no es consecuente con la distribución espacial de los datos. Por ejemplo, cuando el número de grupos difusos es inadecuadamente pequeño, el algoritmo puede asignar a un elemento dado, un alto grado de pertenencia, incluso si este objeto no se encuentra cerca del centroide del grupo correspondiente. Este problema se lleva a cabo principalmente porque, a diferencia de FS, el criterio que acabamos de describir se basa únicamente en la matriz de la partición difusa, además carece de una conexión directa con la información geométrica contenida en los propios elementos y en los centroides de los grupos. Esta es la principal crítica contra los criterios de validación conocidos, basados únicamente en la matriz de la partición difusa, tales como el Coeficiente de la Partición (PC), la Entropia de la Partición (PE), etc.

4.1.10 Medidas de similitud para conjuntos difusos [64]

El concepto de similitud es interpretado de diferentes formas, en dependencia del contexto donde se utilice. La interpretación de similitud en el lenguaje cotidiano es $tener\ características\ en\ común$. Definamos la similitud para conjuntos difusos como el grado en que los conjuntos difusos son iguales. Esta definición está relacionada a los conceptos representados por los conjuntos difusos.

Una medida de similitud pudiera ser definida como una función [64, 65] $S: F(X) * F(X) \rightarrow [0,1]$ y cumple las siguientes propiedades:

$F(X)$ es el conjunto de todos los conjuntos difusos definidos sobre X y $A, B, C \in F(X)$.

I. $S(A,B) = S(B,A)$.

II. $S(A,A^C) = 0$.

III. $S(A,A) = 1$.

IV. $monotonía: A \subseteq B \subseteq C, entonces\ S(A,B) \geq S(A,C)\ y\ S(B,C) \geq S(A,C)$.

Diferentes medidas de similitud han sido propuestas para conjuntos difusos, un estudio de algunas de estas medidas puede ser encontrado en [66, 67]. En general, se pueden dividir en dos grandes grupos [68]:

I. Medidas de Similitud Geométricas.
II. Medidas de Similitud basadas en la Teoría de Conjuntos.

Medidas de similitud geométricas

El modo más *sencillo* de calcular la similitud entre conjuntos difusos es basado en su distancia. Este cálculo se realiza en dos etapas: en la primera parte se obtiene la distancia entre los conjuntos difusos y en la segunda se transformaría esta distancia calculada en una similitud.

Varias medidas de distancias han sido presentadas en la literatura. Entre las más empleadas se encuentran:

I. Distancia de Hamming

$$d_H(A,B) = \sum_{i=1}^{n} |A(x_i) - B(x_i)| .$$

II. Distancia de Hamming Normalizada

$$d_{nH}(A,B) = \frac{1}{n} \sum_{i=1}^{n} |A(x_i) - B(x_i)| .$$

III. Distancia Euclidiana

$$d_E = \sqrt{\sum_{i=1}^{n} \left(A(x_i) - B(x_i)\right)^2} .$$

IV. Distancia de Minkowski

$$d_r(A,B) = \left[\sum_{i=1}^{n} |A(x_i) - B(x_i)|^r \right]^{\frac{1}{r}}, \qquad r \geq 1 .$$

La relación entre la noción de similitud y distancia ha sido expresada de distintas formas. Si d es una medida de distancia entre los conjuntos difusos A y B sobre un universo X, en [69-71] se han presentado las siguientes medidas de similitud, respectivamente:

I. Koczy

$$S(A,B) = \frac{1}{1 + d(A,B)}.$$

II. Williams y Steele

$$S(A,B) = e^{-\alpha d(A,B)}.$$

donde α es un parámetro.

III. Sanitini

$$S(A,B) = 1 - d_r(A,B),\ \ r = 1, 2, \dots, \infty.$$

Medidas de similitud basadas en la teoría de conjuntos

Las medidas de similitud basadas en la Teoría de Conjuntos, hacen hincapié en la intuición de que el grado de similitud debe tener en cuenta tanto el solapamiento de los conjuntos dados como la cantidad de diferencias simétricas. Este tipo de medidas se basan en las operaciones de la Teoría de Conjuntos como son la unión y la intersección.

Si la unión y la intersección son modeladas por max y min y se define $\boxdot$ como:

$$A \boxdot B = max[min\big(A(x), 1 - B(x)\big), min(B(x), 1 - A(x))]. \tag{4.19}$$

Las siguientes son medidas de similitud basadas en la Teoría de Conjunto [72]:

I. $S(A,B) = 1 - |A \boxdot B|$.

II. $S(A,B) = \frac{|A \cap B|}{|A \cup B|}$ reescribiendo esta función en términos de las funciones de pertenencia dadas:

$$S(A,B) = \frac{\sum_{j=1}^{n} min(u_{Aj}, u_{Bj})}{\sum_{j=1}^{n} max(u_{Aj}, u_{Bj})}.$$

III. $S(A,B) = sup_{x \in X}\, A \cap B(x) = max_{x \in X}\big(min\big(u_{Aj}, u_{Bj}\big)\big)$.

A partir de estas medidas de similitud se pueden conformar distintos índices de validación que midan, al igual que los índices definidos en esta sección, ciertos criterios sobre las particiones difusas como son compacidad, separación, etc., pero en este caso basados en medidas de similitud. Un ejemplo de este tipo de medidas de evaluación se puede encontrar en [68].

4.2 Índices externos

4.2.1 Extensión difusa del índice de Rand y otros índices relacionados

El Índice de Rand [73] es muy sencillo e intuitivo, maneja dos matrices de particiones duras (R y Q) sobre el mismo conjunto de datos. La partición de referencia R, codifica las etiquetas de las clases, es decir, la partición de los datos en k clases conocidas. La partición Q, por su parte, particiona los datos en v categorías o grupos, y es la que va a ser evaluada.

Teniendo en cuenta las observaciones anteriores, el índice de Rand se define como:

$$IR = \frac{a + d}{a + b + c + d}, \tag{4.20}$$

donde:

- a: Es el número de pares de objetos que pertenecen a la misma clase en R y al mismo grupo en Q.
- b: Es el número de pares de objetos que pertenecen a la misma clase en R y a diferentes grupos en Q.
- c: Es el número de pares de objetos que pertenecen a diferentes clases en R y al mismo grupo en Q.
- d: Es el número de pares de objetos que pertenecen a diferentes clases en R y a diferentes grupos en Q.

Formulación del índice de Rand basado en la teoría de conjuntos

Para la posterior extensión difusa del índice de Rand, primero se va a definir el mismo basado en la Teoría de Conjuntos, lo que permitirá un mejor entendimiento de los análisis posteriores.

Sean V, X, Y y Z los siguientes conjuntos duros:

- V: Conjunto de pares de objetos que pertenecen a la misma clase en R.
- X: Conjunto de pares de objetos que pertenecen a diferentes clases en R.
- Y: Conjunto de pares de objetos que pertenecen al mismo grupo en Q.
- Z: Conjunto de pares de objetos que pertenecen a diferentes grupos en Q.

De las definiciones anteriores, se puede deducir que los términos individuales del índice de Rand se pueden escribir de la siguiente forma:

$$a = |V \cap Y|.$$

$$b = |V \cap Z|.$$

$$c = |X \cap Y|.$$

$$d = |X \cap Z|\,.$$

Donde $|.|$ y $\cap$ son la cardinalidad y la intersección de conjuntos, respectivamente. Luego, se puede escribir en función de las clases y grupos, cada uno de los conjuntos V, X, Y y Z de la siguiente manera:

$$V = \bigcup_{i=1}^{k} V_i.$$

$$X = \bigcup_{\substack{i_1,\, i_2=1 \\ (i_1 \neq i_2)}}^{k} X_{i_1 i_2.}$$

$$Y = \bigcup_{l=1}^{v} Y_l.$$

$$Z = \bigcup_{\substack{l_1,\, l_2=1 \\ (l_1 \neq l_2)}}^{v} Z_{l_1 l_2}.$$

Donde ∪ representa la unión para conjuntos:

- V_i: Conjunto de pares de objetos que pertenecen a la $i-$ésima clase en R.
- $X_{i_1 i_2}$: Conjunto de pares de objetos que pertenecen a diferentes clases i_1 e i_2 en R, es decir, un objeto pertenece a la clase i_1 y el otro a la clase i_2. $(i_1 \neq i_2)$
- Y_l: Conjunto de pares de objetos que pertenecen al $l-$ésimo grupo en Q.
- $Z_{l_1 l_2}$: Conjunto de pares de objetos que pertenecen a diferentes grupos l_1 y l_2 en Q, es decir, un objeto pertenece al grupo l_1 y el otro al grupo l_2. $(l_1 \neq l_2)$

Entonces, la alternativa equivalente al índice de Rand se logra simplemente sustituyendo las variables a, b, c y d.

Índice de rand difuso [74]

Sea Q una matriz de partición difusa (matriz de pertenencia U) resultado de algún algoritmo difuso de agrupamiento. Entonces, una extensión difusa del índice de Rand para la evaluación de la exactitud de Q con respecto a la partición R puede ser obtenida simplemente redefiniendo los conjuntos duros V, V_i, X, $X_{i_1 i_2}$, Y, Y_l, Z y $Z_{l_1 l_2}$, de tal modo que $|.|$, ∪ y ∩ se convierten en la cardinalidad, unión e intersección de conjuntos difusos respectivamente. Nótese que, puesto que las clases se conocen, entonces R puede ser tratada como una matriz de partición dura. En aras de la comodidad, esta matriz también será tratada como una matriz de partición difusa.

A continuación se redefinirán los conjuntos duros anteriores, como conjuntos difusos:

- V_i: Conjunto difuso de los pares de objetos que pertenecen a la $i-$ésima clase en R. El grado de pertenencia de cada par de objetos (o_1, o_2) está dado por el valor verdadero de la siguiente proposición: *el objeto* o_1 *pertenece a la i-ésima clase y el objeto* o_2 *pertenece a la i-ésima clase*, que es, $r_{io_1}\, t\, r_{io_2}$, donde t es una norma triangular ($t-norma$ [64], por ejemplo: mínimo o producto) utilizada como conjunción para describir el "y" de la proposición.
- $X_{i_1\, i_2}$: Conjunto difuso de los pares de objetos que pertenecen a diferentes clases i_1 y i_2 en R. El grado de pertenencia de cada par de objetos (o_1, o_2) está dado por el valor verdadero de la siguiente proposición: *el objeto* o_1 *pertenece a la clase* i_1 *y el objeto* o_2 *pertenece a la clase* i_2, que es, $r_{i_1 o_1}\, t\, r_{i_2 o_2}$.
- Y_l: Conjunto difuso de los pares de objetos que pertenecen al $l-$ésimo grupo de Q. El grado de pertenencia de cada par de objetos (o_1, o_2) está dado por el valor verdadero de la siguiente proposición: *el objeto* o_1 *pertenece al l-ésimo grupo y el objeto* o_2 *pertenece al l-ésimo grupo*, que es, $q_{lo_1}\, t\, q_{lo_2}$.
- $Z_{l_1\, l_2}$: Conjunto difuso de los pares de objetos que pertenecen a diferentes grupos l_1 y l_2 en Q. El grado de pertenencia de cada par de objetos (o_1, o_2) está dado por el valor verdadero de la siguiente proposición: *el objeto* o_1 *pertenece al grupo* l_1 *y el objeto* o_2 *pertenece al grupo* l_2, que es, $q_{l_1 o_1}\, t\, q_{l_2 o_2}$.
- V: Conjunto difuso de pares de objetos que pertenecen a la misma clase en R. El grado de pertenencia de cada par de objetos (o_1, o_2) está dado por el valor verdadero de la siguiente proposición: *(el objeto* o_1 *pertenece a la clase* 1 *y el objeto* o_2 *pertenece a la clase* 1*)* o *(el objeto* o_1 *pertenece a la clase* 2 *y el objeto* o_2 *pertenece a la clase* 2*)*, etc, que es:

$$V(o_1, o_2) = \left(r_{1o_1}\ t\ r_{1o_2}\right) s \ldots s \left(r_{ko_1}\ t\ r_{ko_2}\right) \triangleq \underset{i=1}{\overset{k}{s}} \left(r_{io_1}\ t\ r_{io_2}\right),$$

donde s es una $co-norma$ triangular ($s-norma$ [64], ejemplo: máximo).

- X: Conjunto difuso de pares de objetos que pertenecen a diferentes clases en R. El grado de pertenencia de cada par de objetos (o_1, o_2) está dado por el valor verdadero de la disyunción de las siguientes proposiciones: *el objeto o_1 pertenece a la clase i_1 y el objeto o_2 pertenece a la clase i_2*, para $i_1, i_2 = 1, \ldots, k \ con\ i_1 \neq i_2$, que es:

$$X(o_1, o_2) = \underset{(i_1 \neq i_2)}{\overset{k}{\underset{i_1,\ i_2 = 1}{s}}} \left(r_{i_1 o_1}\ t\ r_{i_2\, o_2}\right).$$

- Y: Conjunto difuso de pares de objetos que pertenecen al mismo grupo en Q. El grado de pertenencia de cada par de objetos (o_1, o_2) está dado por el valor verdadero de la siguiente proposición: *(el objeto o_1 pertenece al grupo* 1 *y el objeto o_2 pertenece al grupo* 1*)* o *(el objeto o_1 pertenece al grupo* 2 *y el objeto o_2 pertenece al grupo* 2*)*, etc, que es:

$$Y(o_1, o_2) = \underset{l=1}{\overset{v}{s}} \left(q_{lo_1}\ t\ q_{lo_2}\right).$$

- Z: Conjunto difuso de pares de objetos que pertenecen a diferentes grupos en Q. El grado de pertenencia de cada par de objetos (o_1, o_2) está dado por el valor verdadero de la disyunción de las siguientes proposiciones: *el objeto o_1 pertenece al grupo l_1 y el objeto o_2 pertenece al grupo l_2*, para $l_1, l_2 = 1, \ldots, v \ con\ l_1 \neq l_2$, que es:

$$Z(o_1, o_2) = \underset{(l_1 \neq l_2)}{\overset{v}{\underset{l_1,\ l_2 = 1}{s}}} \left(q_{l_1 o_1}\ t\ q_{l_2\, o_2}\right).$$

Ahora, dado que la intersección de dos conjuntos difusos es generalmente computada utilizando una $t-norma$ como conjunción y la cardinalidad de un conjunto difuso está dada por la suma de sus valores de pertenencia, se infiere directamente de las definiciones anteriores que:

$$|V \cap Y| = \sum_{o_1=1}^{o_2-1} \sum_{o_2=2}^{N} V(o_1, o_2)\ t\ Y(o_1, o_2).$$

$$|V \cap Z| = \sum_{o_1=1}^{o_2-1} \sum_{o_2=2}^{N} V(o_1, o_2)\ t\ Z(o_1, o_2).$$

$$|X \cap Y| = \sum_{o_1=1}^{o_2-1} \sum_{o_2=2}^{N} X(o_1, o_2)\ t\ Y(o_1, o_2).$$

$$|X \cap Z| = \sum_{o_1=1}^{o_2-1} \sum_{o_2=2}^{N} X(o_1, o_2)\ t\ Z(o_1, o_2)\ .$$

Se puede apreciar que $V(o_1, o_2)$ y $V(o_2, o_1)$ se refiere al mismo elemento del conjunto difuso V, ya que (o_1, o_2) y (o_2, o_1) es el mismo par de objetos. Ocurre lo mismo con los conjuntos difusos X, Y y Z. Por esta razón, en las sumatorias se cumple que: $o_1 < o_2$.

Debido a la construcción de esta extensión del índice, el índice de Rand original es un caso particular de este índice.

Dado que la cardinalidad de un conjunto difuso es siempre un número no negativo, entonces se deduce que el índice de Rand difuso mantiene la propiedad fundamental de la normalidad, es decir, $IR \in [0, 1]$. Sin embargo, $Q = R$ (la partición a evaluar coincide exactamente con la partición de referencia) no es una condición necesaria y suficiente para que $IR = 1$. Sigue siendo necesaria, pero la suficiencia se garantiza solamente si la partición de referencia R es una partición dura. Esto tiene poco impacto en la utilidad práctica del índice de Rand Difuso puesto que la partición de referencia debe ser siempre dura.

El índice de Rand difuso no está asociado a ningún caso particular o algoritmo de agrupamiento difuso. En efecto, el único requerimiento es que el algoritmo produzca como salida una partición difusa Q de los datos, cuyos elementos: $q_{ij} \in [0, 1]$.

Índices relacionados

Existe una gran variedad de índices que utilizan los mismos términos que el índice de Rand (a, b, c, d). Para estos índices también es válida la extensión descrita anteriormente. Entre estos índices externos se encuentran: índice de Rand Ajustado [75], Coeficiente de Jaccard [76, 77], índice de Fowlkes‑Mallows [78], entre otros.

4.2.2 Un nuevo índice de Rand [79]

A continuación nos centraremos en el índice de Rand, visto como una función de distancia. Gracias a la transformación afín $D_R = 1 - R$, todos los resultados pueden ser transformados directamente, según la concepción inicial del índice de Rand como una medida de similitud.

Dado una partición difusa U de X, cada elemento $x \in X$ puede ser caracterizado mediante su vector de pertenencia:

$$M(x) = (u_{1x}, u_{2x}, \dots, u_{cx}) \in [0,1]^c,$$

donde u_{ik} es el grado de pertenencia del elemento k al $i - ésimo$ grupo. A continuación se define una relación de equivalencia difusa sobre X en términos de una medida de similitud sobre los vectores de pertenencia asociados a cada elemento.

$$E_p(x, x') = 1 - \|M(x) - M(x')\|\ ,$$

donde $\|.\|$ es una distancia sobre $[0, 1]^c$. El único requisito que debe cumplir esta distancia es que sus valores se encuentren en el intervalo $[0, 1]$.

Dado dos particiones difusas P y Q, la idea es generalizar el concepto de concordancia como sigue. Se considera que x y x' son concordantes en la medida en que P y Q coinciden en su grado de concordancia. El grado de concordancia se define como:

$$1 - \left|E_p(x, x') - E_Q(x, x')\right| \in [0, 1] .$$

Análogamente, el grado de discordancia es:

$$\left|E_p(x, x') - E_Q(x, x') .\right|$$

La medida de distancia sobre particiones difusas es definida entonces por la suma normalizada de los grados de discordancia:

$$d(P, Q) = \frac{\sum_{(x,x'),\ x,x' \in X} \left|E_p(x,x') - E_Q(x,x')\right|}{n(n-1)/2} .$$

De igual modo,

$$1 - d(P, Q) .$$

corresponde al grado de concordancia normalizado y además es una generalización directa del índice de Rand original [73].

5 Conclusiones

Los algoritmos de agrupamiento difuso son una alternativa a la clasificación no supervisada o agrupamiento. La gran cantidad de aplicaciones de los métodos de clasificación no supervisada en diferentes campos de investigación, hace que el estudio de los métodos de agrupamiento difuso sea de gran interés, como una nueva variante para enfrentar *con mayor exactitud* los problemas de clasificación.

El tema del agrupamiento difuso como se pudo observar, tiene un gran impacto en la actualidad, debido principalmente a su capacidad de revelar estructuraciones de los datos más próximas a la realidad en muchos casos. Además, se puede decir que por la filosofía que aplican estos métodos, de descubrir conjuntos difusos, permite la existencia de un margen de error en cuanto a la cantidad de conjuntos difusos descubiertos con relación a la cantidad exacta de grupos que subyacen de los datos. A lo largo del trabajo se aprecia que la gran mayoría de los algoritmos de agrupamiento difuso poseen parámetros en su propia definición así como en algunos casos una fase de inicialización, lo que complica el proceso de agrupamiento y puede conllevar a posibles errores que de cierta forma influirían en el agrupamiento final. En algunos casos es necesario introducir el grado difuso de la partición (m), la medida de distancia a utilizar, etc., por lo que se recomienda al igual que en el agrupamiento clásico, un estudio más profundo del problema a resolver que posibilite mejores resultados. Otro aspecto a destacar es que una gran cantidad de métodos difusos están basados en funciones objetivo, por lo que en estos casos la optimización puede ser un proceso que influya grandemente en el resultado final.

En nuestra opinión no se ha estudiado con gran profundidad la difusificación de particiones duras, aprovechando las propiedades de la matriz producida por un algoritmo de agrupamiento particional, conjuntamente con las propiedades de los datos, las similitudes entre ellos, etc.

En el caso de los índices de validación para particiones difusas, se puede concluir que se han desarrollado una gran cantidad de índices internos, pero no pasa lo mismo con los índices externos que han sido poco abordados y como muestra de esto es su reciente aparición. Desde nuestro punto de vista se hace necesario el desarrollo de nuevos índices externos, que permitan la comparación entre

particiones difusas, ayudando a fundamentar las experimentaciones y como punto de partida para un posible proceso de combinación de particiones difusas. Además intentaremos proponer nuevos índices para la validación de particiones difusas. Otro punto importante a destacar sobre la novedad de los índices de validación externos es que se puede proponer el desarrollo de nuevos índices basados en medidas de similitud para conjuntos difusos o mediante la extensión de índices ya existentes para particiones duras.

Por los problemas anteriormente mencionados, se hace necesario el estudio de métodos de combinación de agrupamientos difusos que a su vez permitan suprimir posibles errores como consecuencia de utilizar un método de agrupamiento difuso en particular. En este sentido centraremos nuestros trabajos futuros como forma de complementar todo lo relacionado con el amplio tema de las particiones difusas.

Referencias bibliográficas

1. Bezdek, J.C., Feature selection for binary data: medical diagnosis with fuzzy sets, in Proceedings of the June 7-10, 1976, national computer conference and exposition1976, ACM: New York, New York. p. 1057-1068.
2. Gath, I. and E. Bar-On, Computerized method for scoring of polygraphic sleep recordings. Computer Programs in Biomedicine, 1980. 11(3): p. 217-223.
3. Larsen, L.E., et al., A test of sleep staging systems in the unrestrained chimpanzee. Brain Research, 1972. 40(2): p. 319-343.
4. Zhao, F., et al., A novel fuzzy clustering algorithm with non local adaptive spatial constraint for image segmentation. Signal Processing, 2011. 91(4): p. 988-999.
5. Di Maio, F. and E. Zio. Ensemble of Unsupervised Fuzzy C-Means classifiers for clustering health status of oil sand pumps. in Proceedings of the European Safety and Reliability Conference - ESREL. 2011. Troyes, France.
6. Liang, D., Z. Yongping, and Z. Xueying, An Approach to Retinal Image Segmentations Using Fuzzy Clustering in Combination with Morphological Filters, in 30th Chinese Control Conference (CCC)2011. p. 3062-3065.
7. Park, S., D.U. An, and H. Yoo, Document clustering using NMF and fuzzy relation, in Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication 2011, ACM: Seoul, Korea. p. 1-5.
8. Bezdek, J.C., Fuzzy Mathematics in Pattern Classification, 1973, Applied Math. Center, Cornell University.
9. Bezdek, J.C. and J.C. Dunn, Optimal Fuzzy Partitions: A Heuristic for Estimating the Parameters in a Mixture of Normal Distributions. IEEE Trans. Comput., 1975. 24(8): p. 835-838.
10. Bezdek, J.C., et al., Detection and characterization of cluster substructure. SIAM Journal of Applied Mathematics, 1981. 40(2): p. 339-372.
11. Gustafson, D.E. and W.C. Kessel, Fuzzy clustering with a fuzzy covariance matrix, in Proc. IEEE CDC1979. p. 761-776.
12. Zadeh, L.A., Fuzzy sets. Information and Control, 1965. 8(3): p. 338-353.
13. Halmos, P.R., Naive set theory / Paul R. Halmos, 1960, Princeton, N. J : Van Nostrand.
14. Zadeh, L.A., Similarity relations and fuzzy orderings. Inf. Sci., 1971. 3(2): p. 177-200.
15. Mirzaei, A. and M. Rahmati, A Novel Hierarchical-Clustering-Combination Scheme Based on Fuzzy-Similarity Relations, in IEEE Transactions on Fuzzy Systems 2010. p. 27 - 39
16. Ruspini, E.H., A new approach to clustering. Information and Control, 1969. 15(1): p. 22-32.
17. Bezdek, J., Pattern Recognition with Fuzzy Objective Function Algorithms 1981: Kluwer Academic Publishers.
18. Bezdek, J.C. and J.D. Harris, Convex decompositions of fuzzy partitions. Journal of Mathematical Analysis and Applications, 1979. 67(2): p. 490-512.

19. Krishnapuram, R. and J.M. Keller, A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems, 1993. 1(2): p. 98 - 110.
20. Jezewski, M. and J. Leski, Fuzzy Clustering Finding Prototypes on Classes Boundary Computer Recognition Systems 4, R. Burduk, et al., Editors. 2011, Springer Berlin / Heidelberg. p. 177-186.
21. Silva, L., F. Gomide, and R. Yager, Participatory learning in fuzzy clustering, in Proc. 14th IEEE Int. Conf.on Fuzzy Systems2005: Reno, USA. p. 857–861.
22. Torra, V., Fuzzy c-means for Fuzzy Hierarchical Clustering, in Proc. 14th IEEE Int. Conf.on Fuzzy Systems2005. p. 646 - 651
23. Ling-Juan, L. and L. Yu-Long, A Hierarchical Fuzzy Clustering Algorithm, in 2010 International Conference on Computer Application and System Modeling (ICCASM 2010)2010 Taiyuan p. V12-248 - V12-251.
24. Velasco, J.R., S. Lopez, and L. Magdalena, Genetic fuzzy clustering for the definition of fuzzy sets, in Proceedings of the Sixth IEEE International Conference on Fuzzy Systems 1997: Barcelona , Spain p. 1665 - 1670.
25. Srivastava, V., B. Tripathi, and V. Pathak, An Evolutionary Fuzzy Clustering with Minkowski Distances Neural Information Processing, B.-L. Lu, L. Zhang, and J. Kwok, Editors. 2011, Springer Berlin / Heidelberg. p. 753-760.
26. Tamura, S., S. Higuchi, and K. Tanaka, Pattern classification based on fuzzy relations. IEEE Transactions on Systems, Man and Cybernetics., 1971. 1(1): p. 61-66.
27. Dunn, J.C., A Graph Theoretic Analysis of Pattern Classification via Tamura's Fuzzy Relation. IEEE Transactions on Systems, Man and Cybernetics., 1974. 4(3): p. 310 - 313.
28. Shi, L. and P. He, A Fast Fuzzy Clustering Algorithm for Large-Scale Datasets, in Advanced Data Mining and Applications, X. Li, S. Wang, and Z. Dong, Editors. 2005, Springer Berlin / Heidelberg. p. 732-732.
29. Macqueen, J., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statist. Probability1967. p. 281-297.
30. Dunn, J.C., A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Cybernetics and Systems, 1973. 3(3): p. 32 — 57.
31. A. Flores-Sintas, J.M. Cadenas, and F. Martin, Partition validity and defuzzification. Fuzzy Sets and Systems, 2000. 112(3): p. 433-447.
32. A. Flores-Sintas, J.M. Cadenas, and F. Martin, A local geometrical properties application to fuzzy clustering. Fuzzy Sets and Systems, 1998. 100(1-3): p. 245-256.
33. A. Flores-Sintas, J.M. Cadenas, and F. Martin, Membership functions in the fuzzy C-means algorithm. Fuzzy Sets and Systems, 1999. 101(1): p. 49-58.
34. R.P. Li and M. Mukaidono, A Maximum Entropy Approach to Fuzzy Clustering, in Proc. Fourth IEEE Int'l Conf. Fuzzy Systems1995. p. 2227-2232.
35. Miyamoto, S. and M. Mukaidono, Fuzzy c-means as a regularization and maximum entropy approach, in Proc. of the 7th International Fuzzy Systems Association World Congress (IFSA 1997)1997: Prague, Czech, June 25-30, 1997. p. 86–92.
36. Wu, N., The Maximum Entropy Method 1997, Berlin: Springer.
37. Miyamoto, S., Introduction to Cluster Analysis: Theory and Applications of Fuzzy Clustering 1999, Tokyo: Morikita-Shuppan.
38. Timm, H., et al., An extension to possibilistic fuzzy cluster analysis. Fuzzy Sets and Systems, 2004. 147(1): p. 3-16.
39. Klawonn, F., R. Kruse, and H. Timm. Fuzzy shell cluster analysis. in Learning, networks and statistics. 1997. Springer.
40. Höppner, F., et al., Fuzzy Cluster Analysis 1999, Chichester, United Kingdom: J.Wiley & Sons.
41. Bezdek, J.C., Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, 2005, Springer Science+Business Media, Inc.
42. Schölkopf, B. and A.J. Smola, Learning with Kernels, 2002, MIT Press.

43. Wu, Z.-d., W.-x. Xie, and J.-p. Yu, Fuzzy C-Means Clustering Algorithm Based on Kernel Method, in Fifth International Conference on Computational Intelligence and Multimedia Applications, ICCIMA 20032003, IEEE Computer Society. p. 49 - 54.
44. Zhang, D.-Q. and S.-C. Chen, Clustering Incomplete Data Using Kernel-Based Fuzzy C-means Algorithm. Neural Process. Lett., 2003. 18(3): p. 155-162.
45. Zhang, D.-Q. and S.-C. Chen, Kernel-based fuzzy and possibilistic c-means, in Proc. of ICANN'032003. p. 122–125.
46. Huang, H., Y. Chuang, and C. Chen, Multiple Kernel Fuzzy Clustering. IEEE Transactions on Fuzzy Systems, 2011. 20(1): p. 120 - 134.
47. Graves, D. and W. Pedrycz, Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study. Fuzzy Sets and Systems, 2010. 161(4): p. 522-543.
48. Davé, R., Characterization and detection of noise in clustering. Pattern Recognition Letters, 1991. 12(11): p. 657-664.
49. Davé, R. and S. Sen, On generalising the noise clustering algorithms, in Proc. of the 7th IFSAWorld Congress, IFSA'971997. p. 205–210.
50. Davé, R. and S. Sen, Generalized noise clustering as a robust fuzzy c-m-estimators model, in Proc. of the 17th Int. Conference of the North American Fuzzy Information Processing Society: NAFIPS'981998. p. 256–260.
51. Davé, R. and R. Krishnapuram, Robust clustering method: A unified view. IEEE Transactions on Fuzzy Systems, 1997. 5(2): p. 270-293.
52. McLachlan, G. and D. Peel, Finite Mixture Models 2000, New York: Wiley.
53. Dempster, A.P., N.M. Laird, and D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1977. 39(1): p. 1-38.
54. Bezdek J, C., Numerical taxonomy with fuzzy sets. Journal of Mathematical Biology, 1974. 1: p. 57-71.
55. Bezdek, J.C., Cluster Validity with Fuzzy Sets. Cybernetics and Systems, 1973. 3(3): p. 58 - 73.
56. Xie, X.L. and G. Beni, A Validity Measure for Fuzzy Clustering. IEEE Trans. Pattern Anal. Mach. Intell., 1991. 13(8): p. 841-847.
57. Fukuyama, Y. and M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method, in Proc. 5th Fuzzy Syst. Symp.1989. p. 247-250.
58. Zahid, N., M. Limouri, and A. Essaid, A new cluster-validity for fuzzy clustering. Pattern Recognition, 1999. 32(7): p. 1089-1097.
59. Gath, I. and A.B. Geva, Unsupervised optimal fuzzy clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1989. 11(7): p. 773-780.
60. Krishnapuram, R. and C.-P. Freg, Fitting an unknown number of lines and planes to image data through compatible cluster merging. Pattern Recognition, 1992. 25(4): p. 385-400.
61. Kaufman, L. and P. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis 1990, New York: Wiley-Interscience.
62. Everitt, B.S., S. Landau, and M. Leese, Cluster Analysis 4ed2001, Paris: Arnold.
63. Campello, R.J.G.B. and E.R. Hruschka, A fuzzy extension of the silhouette width criterion for cluster analysis. Fuzzy Sets and Systems, 2006. 157(21): p. 2858-2875.
64. Beg, I. and S. Ashraf, Similarity Measures for Fuzzy Sets. Appl. and Comput. Math., 2009. 8(2): p. 192-202.
65. Bustince, H., Indicator of inclusion grade for interval-valued fuzzy sets. Application to approximate reasoning based on interval-valued fuzzy sets. International Journal of Approximate Reasoning, 2000. 23(3): p. 137-209.
66. Cross, V.V., An analysis of fuzzy set aggregators and compatibility measures, 1993, Wright State Univ. Dayton.
67. Setnes, M., Fuzzy rule-base simplification using similarity measures, 1995, Dept. Elect. Eng., Contr. Lab., Delft Univ. Technol.

68. Zarandi, M.H.F., E. Neshat, and I.B. Türksen, A new cluster validity index for fuzzy clustering based on similarity measure, in 11th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing2007. p. 127–135.
69. László Kóczy, T. and T. Domonkos, Fuzzy rendszerek 2000: Typotex.
70. Williams, J. and N. Steele, Difference, distance and similarity as a basis for fuzzy decision support based on prototypical decision classes. Fuzzy Sets and Systems, 2002. 131(1): p. 35-46.
71. Santini, S. and R. Jain, Similarity is a Geometer. Multimedia Tools Appl., 1997. 5(3): p. 277-306.
72. Zwick, R., E. Carlstein, and D.V. Budesco, Measures of similarity amongst fuzzy concepts: A comparative analysis. International Journal of Approximate Reasoning, 1987. 1(2): p. 221-242.
73. Rand, W.M., Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association (JASA), 1971. 66: p. 846-850.
74. Campello, R.J.G.B., A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. Pattern Recognition Letters, 2007. 28(7): p. 833-841.
75. Hubert, L. and P. Arabie, Comparing partitions. Journal of Classification, 1985. 2: p. 193-218.
76. Jain, A. and R. Dubes, Algorithms for clustering data 1988: Prentice-Hall, Inc.
77. Halkidi, M., Y. Batistakis, and M. Vazirgiannis, On clustering validation techniques. Journal of Intelligent Information Systems, 2001. 17(2--3): p. 107-145.
78. Fowlkes, E.B. and C.L. Mallows, A method for comparing two hierarchical clusterings. Journal of the American Statistical Association (JASA), 1983. 78: p. 553–569.
79. Hüllermeier, E. and M. Rifqi, A Fuzzy Variant of the Rand Index for Comparing Clustering Structures, in Proceedings of the IFSA-EUSFLAT2009. p. 1294-1298.

RT_049, abril 2012
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. No. 21812 e/218 y 222, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba