Una Extensión al Agrupamiento Conceptual Jerárquico Basado en Distancias

Ana Funes[1], María José RamírezQuintana[2], Roberto Uzal[3]

[1,3]Universidad Nacional de San Luis, Ejército de los Andes 950, 5700 San Luis, Argentina

[2]DSIC, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, España

[1]afunes@unsl.edu.ar, [3]ruzal@uolsinectis.com.ar

[2]mramirez@dsic.upv.es

Resumen: *El Agrupamiento Conceptual Jerárquico Basado en Distancias, o HDCC por sus siglas del inglés Hierarchical Distance-based Conceptual Clustering, es una aproximación general al agrupamiento conceptual. HDCC extiende el algoritmo jerárquico aglomerativo basado en distancias tradicional, generando descripciones conceptuales a la vez que descubre los grupos.*

Una de las contribuciones más importantes de HDCC es su marco teórico, que brinda un conjunto de herramientas matemáticas y resultados teóricos necesarios para el análisis de consistencia entre distancias y operadores de generalización en el contexto del algoritmo HDCC. En dicho marco se definen tres niveles de consistencia sobre la base de las discrepancias entre las jerarquías de agrupamientos inducidas por la distancia de enlace y las nuevas jerarquías de conceptos y agrupamientos inducidas por el algoritmo HDCC.

Inspirados por el concepto de generalizaciones basadas en distancia propuesto por Estruch (2008), en este trabajo, revisamos y comparamos las condiciones suficientes para operadores de generalización basados en distancias con las definidas en HDCC, extendiendo el marco teórico con un nuevo nivel de consistencia, el nivel de los dendrogramas basados en distancias.

Palabras Claves: *Agrupamiento conceptual, agrupamiento jerárquico, agrupamiento basado en distancias, HDCC.*

Abstract: *Hierarchical Distance-based Conceptual Clustering (HDCC) is a general approach to conceptual clustering. HDCC extends the traditional distance-based agglomerative algorithm by producing on the fly conceptual descriptions of the discovered clusters.*

One of the main contributions of HDCC is its theoretical framework, which provides a set of mathematical tools and theoretical results useful for the analysis of consistency between distances and generalisation operators in the context of HDCC's algorithm. The framework defines three levels of consistency based on the divergences between the clustering hierarchies induced by the linkage distance and the new hierarchies of concepts and clusters induced by HDCC's algorithm.

Inspired by the concept of distance-based generalisation proposed by Estruch (2008), in this work we revise and compare the sufficient conditions for distance-based generalisation operators vs. the properties defined in HDCC and we extend the framework by adding a new level of consistency –the level of distance-based dendrograms.}

Keywords: *Conceptual clustering, hierarchical clustering, distance-based clustering, HDCC.*

INTRODUCCIÓN

Existen dos aproximaciones bien diferenciadas al Aprendizaje Automático. Por un lado, están aquellos métodos basados en medidas de similitud; por otro lado, encontramos los métodos simbólicos basados en modelos.

Para llevar adelante una tarea de aprendizaje

inductivo, los métodos basados en similitud usan funciones de similitud. En particular, esta función de similitud podría ser una distancia. El concepto de similitud es un concepto más amplio que el de distancia, y se encuentra en la base de muchas técnicas inductivas, en las que se espera que elementos similares se comporten de forma similar. El concepto de distancia, no solo formaliza el concepto de similitud entre dos individuos, sino que satisface propiedades adicionales de un espacio métrico las que pueden ser explotadas en beneficio de la tarea de aprendizaje. Tanto para el aprendizaje supervisado como para el no supervisado, existen varios métodos populares basados en similitud, tales como el agrupamiento jerárquico (Berkhin, P, 2006), (Jain, Murty, and Flynn, 1999), el algoritmo k-NN de los k vecinos más cercanos (Cover and Hart, 1967)0, el algoritmo de k-medias (MacQueen, 1967), entre otros.

En las técnicas simbólicas basadas en el modelo, el concepto subyacente es el concepto de generalización. Este es un concepto importante en el Aprendizaje Automático ya que cualquier aprendizaje inductivo involucra algún tipo de generalización. Estas técnicas, a diferencia de las basadas en medidas de similitudes y distancias, se apoyan en la idea de que aquellos conceptos (generalizaciones o patrones) descubiertos a partir de datos antiguos pueden ser empleados para describir nuevos datos que sean cubiertos por las generalizaciones o patrones extraídos de los datos antiguos. Las técnicas basadas en el modelo reciben su nombre ya que producen un modelo simbólico, el cual puede ser interpretado por un usuario. Algunas técnicas simbólicas supervisadas bien conocidas son los árboles de decisión y las reglas de asociación, mientras que una técnica de aprendizaje simbólico no supervisado es el agrupamiento conceptual de Michalski (Michalski, 1980), (Michalski and Stepp, 1983) y los trabajos EB (Fisher, 1987) y GCF (Talavera and Béjar, 2001).

Si bien los métodos basados en similitudes son flexibles ya que pueden ser aplicados a cualquier tipo de dato siempre que contemos con una distancia o medida de similitud adecuada para el tipo de dato, la función o modelo aprendido puede ser inexistente, como por ejemplo en el aprendizaje basados en instancias, o puede resultar ininteligible para los usuarios al no ofrecer ninguna explicación sobre las decisiones tomadas. En particular, las técnicas de agrupamiento basadas en similitudes descubren los agrupamientos usando una medida numérica de similitud entre los ejemplos al mismo tiempo que buscan maximizar la similitud entre los elementos de cada grupo descubierto y minimizar la similitud entre los elementos de grupos diferentes. En esas técnicas, los agrupamientos resultantes carecen de descripciones conceptuales, dificultando su interpretación. Por ejemplo, consideremos un agrupamiento de documentos en donde se han descubierto cinco grupos o clases de documentos con características similares de acuerdo con una función de similitud dada. Sabemos que los elementos en cada uno de los grupos han sido agrupados porque ellos se encuentran cercanos en su espacio métrico (o son similares) con respecto a la distancia o función de similitud subyacente usada en el algoritmo; sin embargo, resultaría de mayor utilidad poder conocer también cuáles son las características comunes a los documentos en cada uno de los grupos, es decir, contar con grupos enriquecidos con patrones (o generalizaciones) que describan las propiedades comunes a los elementos de cada grupo.

Con el propósito de salvar este problema, en la literatura se han propuesto diversos métodos híbridos. Algunos ejemplos de ellos, entre otros, son el sistema Anaprom (Golding and Rosenbloom, 1993), el sistema EACH (Exemplar-Aided Constructor of Hyperrectangles) (Salzberg, 1991), el sistema RISE (Rule Induction from a Set of Exam-

ples) (Domingos 1996), RIONA (Rule Induction with Optimal Neighbourhood Algorithm) (Góra and Wojna, 2002). Sin embargo, al combinar ambos paradigmas, suelen aparecer problemas de inconsistencia entre la distancia subyacente y las generalizaciones descubiertas. En esta categoría se encuentra HDCC, el cual integra ambos paradigmas de aprendizaje bajo la forma de una aproximación general al agrupamiento jerárquico que integra distancias y generalizaciones. HDCC brinda un algoritmo de agrupamiento jerárquico el cual produce un nuevo tipo de jerarquía de grupos llamada dendrograma conceptual. Los dendrogramas conceptuales son dendrogramas cuyos grupos cuentan con descripciones conceptuales obtenidas a partir de la aplicación combinada de operadores de generalización y distancias.

Una de las mayores contribuciones de HDCC es su marco teórico, que brinda la posibilidad de analizar la presencia de inconsistencias entre distancias y operadores de generalización cuando son usados de forma combinada bajo el algoritmo de agrupamiento de HDCC. Para esto, el marco cuenta con un conjunto de herramientas matemáticas que nos permite llevar a cabo análisis del grado de consistencia entre pares de distancias y operadores de generalización. Esto permite conocer de antemano cuán consistentes será la jerarquía inducida por una distancia dada y los patrones descubiertos, producto de un determinado operador de generalización. En otras palabras, podemos conocer de antemano si los elementos cubiertos por los patrones, resultantes del uso de un operador de generalización dado, son cercanos con respecto a la distancia subyacente en su espacio métrico.

La principal contribución de este trabajo es una extensión al marco teórico de HDCC. Para analizar el nivel de afinidad entre generalizaciones y distancias, el marco, originalmente, define tres niveles de consistencia. A su vez, para cada nivel, define la correspondiente propiedad que los operadores de generalización deben satisfacer a fin de alcanzar ese nivel de consistencia. Estos niveles de consistencia garantizan que los dendrograma conceptuales resultantes reflejen en mayor o menor medida la distribución de los elementos en su espacio métrico.

La extensión al marco, que acá presentamos, se basa en conceptos propuestos en el trabajo de Estruch (2008), en la cual sumamos un nuevo nivel de consistencia para los dendrogramas conceptuales basándonos fundamentalmente en el concepto de generalizaciones basadas en distancias definido por Estruch.

El resto de este trabajo se encuentra organizado como sigue. En la sección “Conceptos Previos”, presentamos de forma resumida algunos conceptos necesarios para la comprensión del resto del trabajo. Los conceptos principales de la aproximación al agrupamiento conceptual jerárquico basado en distancias HDCC son dados en la sección “El algoritmo HDCC” y la sección “Consistencia entre operadores de generalización y distancias”, en donde se describe el marco teórico. En la sección “Dendrogramas basados en distancias”, presentamos la extensión al marco así como algunos ejemplos de operadores de generalización basados en distancias. Finalmente, la sección “Conclusiones” cierra el trabajo con algunas conclusiones y trabajo futuro.

CONCEPTOS PREVIOS

Una manera de definir una generalización de un conjunto finito de elementos E en un espacio métrico (X, d) es de forma extensional, como un conjunto que contiene a E. Sin embargo, esta definición por extensión no dice nada acerca del concepto o patrón que los elementos del conjunto comparten.

Un patrón $p \in \mathcal{L}$, donde $\mathcal{L}$ es un lenguaje de patrones, es una forma intencional de representar un conjunto de elementos en un conjunto X. Por

ejemplo, dados los números $\{10, 3, 8\}$ un patrón que los describe podría ser el intervalo cerrado $[3, 10]$.

Adicionalmente, los elementos descriptos por un patrón p son denotados por $Set(p)$, referido como la cobertura del patrón, y decimos que un elemento $x \in X$ es cubierto por un patrón p si $x \in Set(p)$. Por ejemplo, $Set([3, 10]) = \{3, 4, 5, 6, 7, 8, 9, 10\}$ y 6 es cubierto por el patrón $[3, 10]$.

En particular, en el contexto de HDCC, interesan las generalizaciones de pares de elementos y de pares de patrones, así que se dice que un patrón es el resultado de una función que transforma dos elementos de un espacio métrico y cuya cobertura incluye al menos los dos elementos, o es el resultado de una función que transforma dos patrones y cuya cobertura incluye al menos los elementos en la unión de las coberturas de ambos patrones. Estas funciones reciben el nombre de operadores binarios de generalización y operadores binarios de generalización de patrones, respectivamente.

Las siguientes definiciones introducen formalmente estos conceptos. La Definición 1 formaliza el concepto de operador binario de generalización sobre un espacio métrico, mientras que el concepto de operador binario de generalización de patrones es dado en la Definición 2.

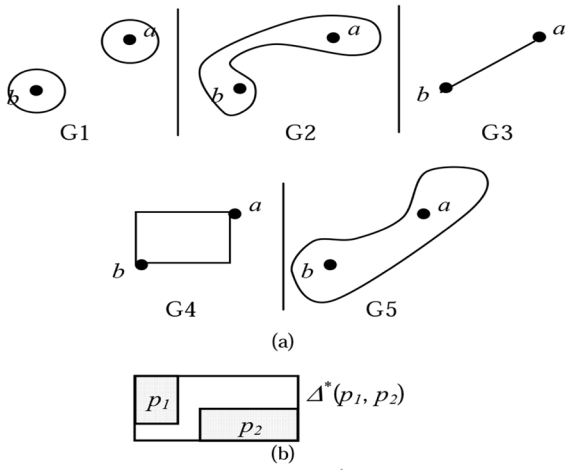

Figura 1. (a) Cinco posibles generalizaciones de dos puntos en $\Re^2$. (b) Una generalización de dos patrones p_1 y p_2 en $\Re^2$.

DEFINICIÓN 1

Sea (X, d) un espacio métrico y L un lenguaje de patrones. Un operador binario de generalización es una función $\Delta: X \times X \rightarrow \mathcal{L}$ tal que dados $x_1 \in X$, $x_2 \in X$, $\Delta(x_1, x_2) = p$, con $p \in \mathcal{L}$ y $x_1 \in Set(p)$ y $x_2 \in Set(p)$.

La figura 1(a) muestra cinco generalizaciones (patrones) posibles de dos puntos a y b en el espacio métrico $(\mathcal{R}_2, d)$, donde d es la distancia Euclidea.

DEFINICIÓN 2

Sea (X, d) un espacio métrico y $\mathcal{L}$ un lenguaje de patrones. Un operador binario de generalización de patrones es una función $\Delta^*: \mathcal{L} \times \mathcal{L} \rightarrow \mathcal{L}$ tal que dados $p_1 \in \mathcal{L}$ y $p_2 \in \mathcal{L}$, $\Delta^*(p_1, p_2) = p$, con $p \in \mathcal{L}$ y $Set(p_i) \subseteq Set(p)$ $(i \in \{1,2\})$.

En la figura 1(*b*), mostramos una generalización posible para los dos patrones p_1 y p_2 en $\mathcal{L}$, siendo L el conjunto de todos los rectángulos. Notemos que cuando $\mathcal{L} = X$, los operadores Δ^* y Δ pueden ser iguales. Esto sucede, por ejemplo, cuando trabajamos con átomos de un lenguaje de primer orden, ya que Δ^* y Δ podrían ambos instanciarse en el operador de generalización menos general de Plotkin (lgg) (Plotkin, 1970).

EL ALGORITMO HDCC

El algoritmo HDCC está basado en el algoritmo de agrupamiento jerárquico aglomerativo tradicional, el cual construye una jerarquía de grupos a partir de los elementos individuales, fusionando los grupos de manera progresiva (ver (Berkhin, 2006) para mayor información). La unión de los grupos está dirigida por una distancia entre los mismos, conocida como distancia de enlace. Generalmente,

la distancia de enlace es determinada por (a) la máxima distancia entre los elementos de cada grupo, referida como distancia de enlace completa $d^c_{\mathcal{L}}$; (b) la mínima distancia entre los elementos de cada grupo, referida como distancia de enlace simple $d^s_{\mathcal{L}}$; (c) la distancia media entre los elementos de cada grupo, referida como distancia de enlace a la media $d^a_{\mathcal{L}}$; (d) la mínima distancia entre los prototipos de los grupos, referida como distancia de enlace a los prototipos $d^p_{\mathcal{L}}$, entre otras. En el resto de este trabajo, solo consideramos $d^c_{\mathcal{L}}$, $d^s_{\mathcal{L}}$, $d^a_{\mathcal{L}}$, y $d^p_{\mathcal{L}}$; usamos $d_{\mathcal{L}}$ para referirnos a cualquiera de ellas.

HDCC toma como base el algoritmo jerárquico tradicional y lo adapta de manera tal que hace trabajar juntos dos operadores de generalización (un operador binario de generalización y un operador binario de generalización de patrones) y dos distancias (una distancia de enlace $d_{\mathcal{L}}$ y una distancia d entre los elementos de un espacio métrico (X, d)). En cada iteración, luego de enlazar los dos grupos más cercanos de acuerdo a la distancia de enlace empleada, el algoritmo fusiona este nuevo grupo con aquellos grupos cubiertos por su generalización o patrón. El patrón de cada grupo provee una descripción para los elementos de su grupo, los cuales se encuentran cercanos en el espacio métrico de acuerdo a las distancias subyacentes $d_{\mathcal{L}}$ y d, pero también describe todos aquellos elementos que, aunque no se encuentran lo suficientemente cerca para ser enlazados por la distancia, son cubiertos por el patrón. Para representar la jerarquía de los grupos y conceptos, HDCC produce un dendrograma extendido llamado dendrograma conceptual. Un dendrograma conceptual muestra no sólo una jerarquía de grupos, cuyos elementos han sido agrupados dirigidos por las distancias $d_{\mathcal{L}}$ y d, sino que también provee una jerarquía de generalizaciones (patrones), que nos brindan descripciones de los elementos que componen cada uno de los grupos en la jerarquía. En un dendrograma conceptual las líneas sólidas representan a los grupos enlazados por similitud o proximidad en el espacio métrico, mientras que las líneas punteadas se corresponden con aquellos grupos enlazados por generalización.

Las figura 2 (a) y (b) muestran un ejemplo sencillo de un dendrograma tradicional y el correspondiente dendrograma conceptual para la evidencia $\{aa, aab, abb, aabbbbbb\}$, en ambos casos usando la distancia de enlace simple. Los elementos pertenecen al espacio métrico (X, d), donde X representa el conjunto de las listas finitas de símbolos sobre el alfabeto $\sum=\{a, b\}$ y d es la distancia de edición o distancia de Levenshtein (1966). El patrón $p=aa^*$ cubre el grupo $\{aa, aab, aabbbbbb\}$, el cual se ha formado considerando, en primer lugar, la distancia entre grupos y, en segundo lugar, la cobertura del patrón aa^* resultante de la aplicación del operador de generalización elegido.

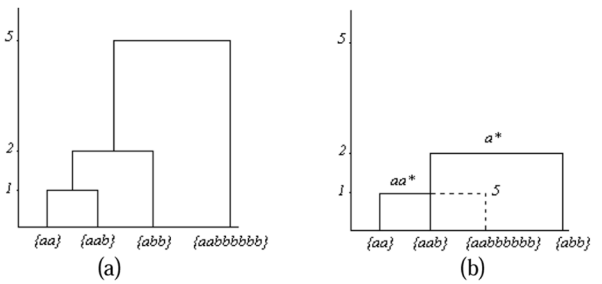

Figura 2. (a) Dendrograma tradicional y (b) conceptual.

CONSISTENCIA ENTRE OPERADORES DE GENERALIZACIÓN Y DISTANCIAS

Al integrar técnicas de aprendizaje basadas en distancias con técnicas de aprendizaje conceptual, por lo general, surgen problemas de consistencia. Las distancias entre los elementos de un espacio métrico y las generalizaciones de dichos elementos no siempre “van de la mano”.

Para ilustrar este problema, consideremos nuevamente el ejemplo de la figura 2. Al aplicar

el algoritmo tradicional jerárquico aglomerativo con enlace simple $d^s_{\mathcal{L}}$ y distancia de edición d, los primeros grupos en enlazarse son *{aa}* y *{aab}*. Por otro lado, HDCC producirá un patrón para el grupo *{aa, aab}*, en este caso *aa** (ver figura 2 (b)). Claramente, existe una inconsistencia entre los elementos descriptos por el patrón *aa* (aa, aab, aaa, aaba, aabb,…)* producido por el operador de generalización empleado y los grupos inducidos por $d^s_{\mathcal{L}}$ y d. Notemos que *aa** cubre el ejemplo *aabbbbbb* pero no cubre el ejemplo *abb*, que es más cercano a los elementos de *{aa, aab}* de acuerdo con la distancia de edición, como se ve en la figura 3.

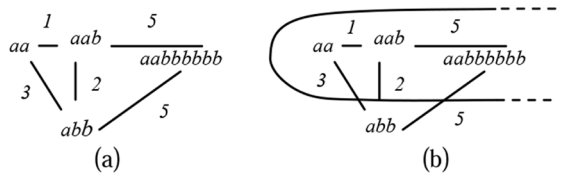

Figura 3 (a) Distancias de edición entre listas de símbolos sobre el alfabeto $\sum$*= {a, b}. (b)* Cobertura del patrón *aa** .

Esta inconsistencia entre distancias y generalizaciones hace que las formas de los dendrogramas conceptuales y tradicionales difieran de manera significativa. La forma exacta de un dendrograma conceptual dependerá no sólo de la distancia d entre los elementos de la evidencia y de la distancia de enlace $d_{\mathcal{L}}$ empleada sino también de los operadores de generalización elegidos.

Sobre la base de la similitud entre los dendrogramas conceptuales y tradicionales, se han definido tres niveles de consistencia ente las distancias y los operadores de generalización cuando vienen usados de forma conjunta en HDCC (Funes et al. 2008) (Funes, 2008). Cuanto más similares son los dendrogramas más consistente es la distancia con respecto al operador de generalización empleado. En las próximas sub secciones, resumimos los tres niveles y, en la sección "Dendrogramas basados en distancias", presentamos una extensión al marco teórico de HDCC, en la forma de una nueva propiedad para los operadores de generalización, la cual garantiza un nivel mínimo de consistencia entre distancias y generalizaciones.

DENDROGRAMAS EQUIVALENTES

En algunos casos, el dendrograma conceptual y el tradicional resultan isomórficos. Esta situación aparece cuando los patrones descubiertos cubren exclusivamente aquellos grupos que también son enlazados por la distancia. Por lo tanto, se dice que un dendrograma conceptual es equivalente a un dendrograma tradicional si y sólo si para cada grupo C del dendrograma todos sus hijos se encuentran enlazados a una misma distancia l. Este concepto se encuentra formalizado en la Definición 3.

DEFINICIÓN 3

Sea T el árbol que resulta de HDCC. T es equivalente al dendrograma tradicional sii $\forall$ $(C, p, l) \in T$, $\forall(C_1, p_1, l_1),\ldots, (C_n, p_n, l_n)$ hijos de (C, p, l) en T: $d_{\mathcal{L}}(C_i, C_j, d) = l$, con $i \neq j$; $i, j = 1, \ldots, n$.

Para lograr el máximo grado de consistencia, es decir, el de los dendrogramas equivalentes, los operadores de generalización deben producir patrones tales que cada vez que HDCC enlace dos grupos C_1 y C_2 a una distancia de enlace mínima l, el patrón p asociado al nuevo grupo no cubra ningún otro grupo C cuyas distancias l_1 y l_2 a C_1 y a C_2, respectivamente, sea mayor que l. Los operadores de generalización que satisfacen esta propiedad son referidos como operadores fuertemente acotados por $d_{\mathcal{L}}$.

Intuitivamente, un operador binario de generalización de patrones es fuertemente acotado por $d_{\mathcal{L}}$ si y sólo si para cualquier par de patrones p_1, p_2 y para cualquier par de conjuntos C_1 y C_2 cubiertos por p1 y p_2, respectivamente, las distancias de enlace a C_1

y a C_2 desde los nuevos elementos cubiertos por la generalización de p_1 y p_2 son iguales o menores que la distancia de enlace entre C_1 y C_2, es decir, los nuevos elementos cubiertos por la generalización de p_1 y p_2 caen en las bolas de radio $d_{\mathcal{L}}(C_1, C_2, d)$ y centro en los puntos de enlace de C_1 y de C_2. Los puntos de enlace son, en el caso de $d^s_{\mathcal{L}}$, los dos elementos de C_1 y C_2 más cercanos entre sí; los dos más distantes en dcL; los prototipos en el caso de $d^p_{\mathcal{L}}$ y los centroides en el caso de $d^a_{\mathcal{L}}$ (asumiendo que el espacio métrico es continuo). La propiedad de acotabilidad fuerte es expresada formalmente en la Definición 4.

DEFINICIÓN 4

Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones y $d_{\mathcal{L}}$ una distancia de enlace. Un operador binario de generalización de patrones Δ* es fuertemente acotado por $d_{\mathcal{L}}$ sii $\forall\, p_1, p_2 \in \mathcal{L}, C_1 \in Set(p_1), C_2 \subseteq Set(p_2), C \subseteq Set(\Delta^*(p_1, p_2)) - (Set(p_1) \cup Set(p_2))$: $d_{\mathcal{L}}(C, C_1, d) \leq d_{\mathcal{L}}(C_1, C_2, d) \vee d_{\mathcal{L}}(C, C_2, d) \leq d_{\mathcal{L}}(C_1, C_2, d)$.

La Definición 5 formaliza la misma propiedad para los operadores binarios de generalización, los que son usados por HDCC cuando generaliza pares de grupos unitarios. Puesto que la distancia de enlace $d_{\mathcal{L}}$ entre los conjuntos unitarios reduce, en todos los casos, a la distancia d entre elementos del espacio métrico, un operador binario de generalización D se dice fuertemente acotado por d (y no por $d_{\mathcal{L}}$).

DEFINICIÓN 5

Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones. Un operador binario de generalización Δ es fuertemente acotado por d sii $\forall\, e, e_1, e_2 \in X$: si $e \in Set(\Delta(e_1, e_2))$ entonces $d(e, e_1) \leq d(e_1, e_2) \vee d(e, e2) \leq d(e_1, e_2)$.

La distancia de enlace $d_{\mathcal{L}}$ empleada afecta la propiedad de acotabilidad de los operadores de generalización. Por lo tanto, un operador de generalización podría ser fuertemente acotado bajo una cierta distancia de enlace $d_{\mathcal{L}}$ pero no bajo otra, aun conservando la misma distancia d.

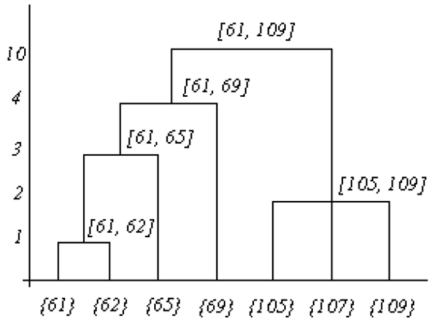

Figura 4. Dendrograma conceptual equivalente al tradicional.

En (Funes et al., 2008), los autores han demostrado que la propiedad de acotabilidad fuerte es una condición suficiente para los dendrogramas equivalentes, lo cual viene establecido en la Proposición 1.

PROPOSICIÓN 1

Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones para X, Δ un operador binario de generalización, Δ* un operador binario de generalización de patrones y $d_{\mathcal{L}}$ una distancia de enlace. Para toda evidencia $E \subseteq X$, el dendrograma conceptual T que resulta de HDCC $(E, X, d, \Delta^*, \Delta, d_{\mathcal{L}})$ es equivalente al dendrograma tradicional si los operadores de generalización Δ y Δ* son fuertemente acotados por $d_{\mathcal{L}}$.

La Figura 4 muestra un dendrograma conceptual para un conjunto de números reales, el cual bajo $d^s_{\mathcal{L}}$ es equivalente a su correspondiente dendrograma tradicional. $\mathcal{L}$ está dado por el conjunto de los intervalos finitos cerrados en $\mathcal{R}$, d por la diferencia absoluta, mientras que $\Delta^*(p_1, p_2)$ viene dado por el intervalo [a, b], donde a es el mínimo valor entre los dos límites inferiores de p_1 y p_2, y b es el máximo valor de los dos límites superiores. $\Delta(e_1, e_2)$ se define como $[min(e_1, e_2), max(e_1, e_2)]$.

DENDROGRAMAS QUE PRESERVAN EL ORDEN

En algunos casos, los dendrogramas conceptuales, si bien no resultan ser equivalentes a los tradicionales, pueden solo preservar el orden en el cual los grupos son enlazados por $d_{\mathcal{L}}$; es decir, los patrones descubiertos no deben cubrir ningún grupo que se encuentre más lejos que uno más cercano que no es cubierto. Estos dendrogramas conceptuales son referidos como dendrogramas que preservan el orden.

Intuitivamente, un dendrograma conceptual que preserva el orden es uno en el que, para cualquier nodo (C, p, l) en la jerarquía, sus hijos se encuentran enlazados a una misma distancia l o son atraídos por p con una distancia de enlace que es menor que todas las distancias de enlace a los grupos que no son cubiertos por p. Este concepto es formalizado por la definición 6.

DEFINICIÓN 6

Sea (X, d) un espacio métrico y T el árbol que resulta de HDCC. T preserva el orden sii $\forall (C, p, l) \in T, \forall (C_i, p_i, l_i)$ hijo de $(C, p, l), \exists (C_i, p_i, l_i)$ hijo de (C, p, l) tal que $d_{\mathcal{L}}(C_i, p_i, d) = l \vee (\mathrm{dL}(C_i, p_i, d) < d_{\mathcal{L}}(C', C_i, d) \wedge d_{\mathcal{L}}(C_i, C_j, d) < d_{\mathcal{L}}(C', \mathrm{Cj}, d))$, para todo $(C', p', l') \in T$, $C' \not\subset Set(p)$.

Para lograr este nivel de consistencia, los operadores de generalización deben producir patrones tales que cada vez que HDCC enlace dos grupos C_1 y C_2, con patrones p_1 y p_2, cualquier otro grupo Ccubierto por la generalización de p_1 y de p_2 pero no enlazado por la distancia $d_{\mathcal{L}}$ deberá tener distancias de enlace a C_1 y a C_2 menores que las distancias a C_1 y a C_2 desde cualquier otro grupo C' que el patrón no cubra. Este concepto es formalizado por la propiedad de acotabilidad débil para los operadores binarios de generalización de patrones dada en la definición 7. Análogamente, la definición 8 establece la misma propiedad para los operadores binarios de generalización, mientras que la proposición 2 establece que la acotabilidad débil es una condición suficiente para la los dendrogramas que preservan el orden (ver0 (Funes et al., 2008) para la demostración de la Proposición 2).

DEFINICIÓN 7

Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones y $d_{\mathcal{L}}$ una distancia de enlace. Un operador binario de generalización de patrones Δ^* es débilmente acotado por $d_{\mathcal{L}}$ sii $\forall p_1, p_2 \in \mathcal{L}, C_1 \subseteq Set(p_1), C_2 \subseteq Set(p_2), C \subseteq Set(\Delta^*(p_1, p_2)) - (Set(p_1) \cup Set(p_2)), C' \not\subset Set(\Delta^*(p_1, p_2)) : (d_{\mathcal{L}}(C, C_1, d) \leq d_{\mathcal{L}}(C_1, C_2, d) \wedge d_{\mathcal{L}}(C, C_2, d) \leq d_{\mathcal{L}}(C_1, C_2, d)) \wedge (d_{\mathcal{L}}(C, C_1, d) < d_{\mathcal{L}}(C', C_1, d) \wedge d_{\mathcal{L}}(C, C_2, d) < d_{\mathcal{L}}(C', C_2, d))$.

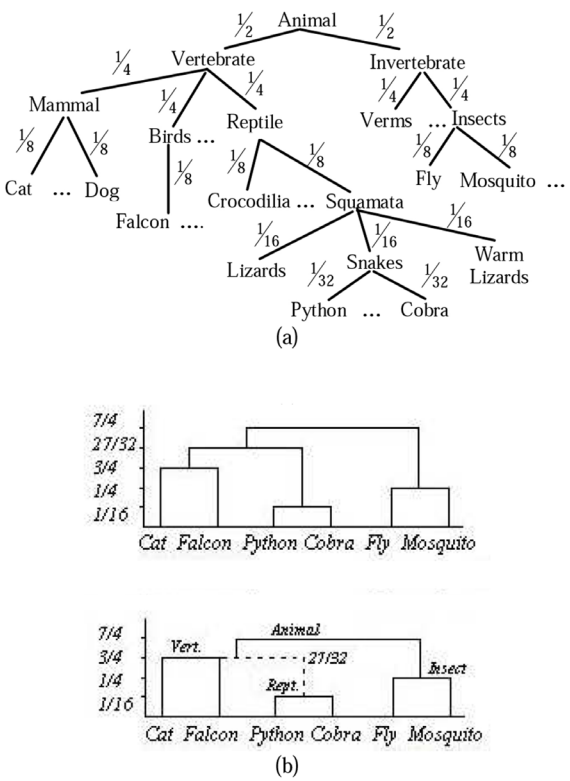

Figura 5. (a) La relación R representada como árbol. (b) Dendrograma conceptual que preserva el orden y el correspondiente dendrograma tradicional.

DEFINICIÓN 8

Sea (X, d) un espacio métrico y $\mathcal{L}$ un lenguaje de patrones. Un operador binario de generalización Δ es débilmente acotado por d sii $\forall e, e', e_1, e_2 \in X$: si $e \in Set(\Delta(e_1, e_2)) \wedge e' \notin Set(\Delta(e_1, e_2))$ entonces $d(e, e_1) \leq d(e_1, e_2) \vee d(e, e_2) \leq d(e_1, e_2) \vee ((d(e, e_1) < d(e', e_1) \vee d(e, e_2) < d(e', e_2))$.

PROPOSICIÓN 2

Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones, Δ un operador binario de generalización, Δ * un operador binario de generalización de patrones y $d_{\mathcal{L}}$ una distancia de enlace. Para cualquier evidencia $E \subseteq X$, el dendrograma conceptual T que resulta de HDCC($E, X, d,$ Δ*, Δ, $d_{\mathcal{L}}$) preserva el orden si los operadores Δ y Δ* son débilmente acotados por $d_{\mathcal{L}}$ y d, respectivamente.

La figura 5 (b) muestra un ejemplo de un dendrograma de datos nominales que preserva el orden bajo la distancia de enlace simple $d^s_{\mathcal{L}}$. El ejemplo aplica una distancia inducida por una relación R, donde R es un orden parcial. R está definida como xRy si x es un y. La figura 5 (a) muestra una parte de la relación R como una jerarquía de árbol. La distancia entre dos elementos es la suma de los costos asociados a cada arco del camino más corto que conecte los dos elementos. El costo de un arco en el nivel i del árbol (contando desde la raíz) está dado por $w_i = 1/2^i$. $\Delta(e_1, e_2)$ se define como el ancestro directo de e_1 y e_2 si $e_1 \neq e_2$, en otro caso es igual a e_1, y $\Delta^*(p_1, p_2)$ es definido de forma análoga. En lafigura 5 (b) podemos ver el dendrograma conceptual y el correspondiente dendrograma tradicional. La evidencia está formada solamente por elementos en las hojas de R. Los nodos internos de R definen el lenguaje de patrones.

DENDROGRAMAS ACEPTABLES

Algunos operadores de generalización, aunque no (débilmente) acotados por una distancia, conducen a dendrogramas que son consistentes con la distancia en un sentido más amplio. El razonamiento detrás de esto es que un patrón no debería cubrir nueva evidencia cuya distancia a la vieja evidencia sea mayor que la mayor de las distancias entre los elementos de la antigua evidencia. Los operadores que producen este tipo de patrones, así como los dendrogramas que resultan de su aplicación, son referidos como aceptables. La definición 9 formaliza este concepto para Δ*.

DEFINICIÓN 9

Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones, y $d^c_{\mathcal{L}}$ la distancia de enlace completo. Un operador binario de generalización de patrones Δ* es aceptable sii $\forall p_1, p_2 \in \mathcal{L}, e \in Set(\Delta^*(p_1, p_2)), \exists e' \in Set(p_1) \cup Set(p_2) : d(e, e') \leq d^c_{\mathcal{L}}(Set(p_1), Set(p_2), d)$.

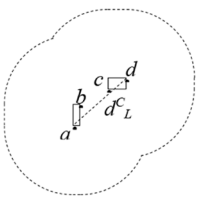

Fig. 6. Cobertura máxima para un patrón aceptable.

La propiedad de aceptabilidad es independiente de $d_{\mathcal{L}}$. Sólo depende de la distancia d, específicamente de $max\{ d(x, y) : x \in Set(p_1), y \in Set(p_2) \}$. Para simplificar la notación se ha usado en la definición 9 la distancia de enlace completo $d^c_{\mathcal{L}} Set(p_1), Set(p_2), d)$ en lugar de $max\{ d(x, y) : x \in Set(p_1), y \in Set(p_2) \}$.

La figura 6 muestra la unión de los círculos con centros en los puntos de enlace completo tomados de los conjuntos de cobertura $Set(p_1)$ y $Set(p_2)$,

respectivamente, y con radios $d^c_{\mathcal{L}}(Set(p_1), Set(p_2), d)$, que determina la máxima cobertura para la generalización de los patrones p_1 y p_2 producida por una un operador de generalización aceptable. En este ejemplo, los puntos de enlace completo están dados por a y d, tomados de la cobertura de los dos patrones p_1 y p_2 (rectángulos mínimos) mostrados en la figura para la evidencia $\{a, b, c, d\}$ en $\mathcal{R}2$.

La propiedad de aceptabilidad para operadores binarios de generalización lleva a la misma condición que para los operadores de generalización Δ fuertemente acotados (ver definición 5). Por lo tanto, los dendrogramas aceptables son aquellos dendrogramas que resultan de la aplicación de un par de operadores de generalización (Δ, Δ*) que son fuertemente acotado y aceptable, respectivamente.

En este nivel de consistencia, los dendrogramas conceptual y tradicional pueden diferir significativamente.

DENDROGRAMAS BASADOS EN DISTANCIAS

Un concepto interesante relativo a las generalizaciones ha sido introducido por Estruch en (Estruch, 2008). En dicho trabajo, el autor trata la problemática de las inconsistencias entre operadores de generalización y distancias, definiendo un marco en el que los paradigmas de Aprendizaje Automático basado en distancias y simbólico pueden integrarse de manera consistente. Según Estruch, la integración puede llevarse a cabo de manera consistente cuando las generalizaciones son, de acuerdo a lo que él define, basadas en distancias.

El concepto de generalización basada en distancia puede verse, de forma intuitiva, como un patrón que conecta todos los ejemplos generalizados de manera tal que cualquiera de estos ejemplos puede ser alcanzado desde cualquiera de los otros ejemplos a través de la distancia moviéndonos en el espacio métrico siempre dentro de la generalización.

Para definir este concepto, el autor introduce la noción de conexión, significando que cuando un patrón generaliza un par de elementos e_1 y e_2, podemos movernos en el espacio métrico de un elemento al otro a través de elementos similares (cercanos) sin salirnos de la generalización. Por ejemplo, si generalizamos el conjunto de números enteros {2, 10}, esperamos poder ir desde el 2 al 10 a través del camino $2 \rightarrow 3 \rightarrow ... \rightarrow 9 \rightarrow 10$ sin salir de la generalización. Esto significa que, dentro del espacio métrico, podemos movernos de forma gradual de un elemento a otro pasando por elementos que también son cubiertos por la generalización de los dos elementos en cuestión.

Intuitivamente, la idea detrás del concepto de generalización basada en distancia de Estruch es que los elementos intermedios (en términos de la distancia) deberían estar incluidos en la generalización. Sin embargo, no cualquier camino debería considerarse, sino que una generalización debería incluir los caminos más cortos que conecten los elementos que conducen a la generalización. Esto significa que todos los elementos en el camino más corto entre dos elementos generalizados (en muchos espacios métricos, el segmento recto que los conecta) deben estar incluidos en la generalización.

El concepto de elemento intermedio en un espacio métrico es obtenido cambiando la propiedad de desigualdad triangular de los espacios métricos por una igualdad: dado un espacio métrico (X, d) y dos elementos $x, z \in X$, un elemento $y \in X$ está entre medio de x y z, o es un elemento intermedio con respecto a d, si $d(x, z) = d(x, y) + d(y, z)$.

El autor también extiende esta idea a conjuntos de ejemplos, lo que nos lleva a la noción de nervio y de esqueleto, usados para definir el concepto de operador de generalización basado en distancia.

Intuitivamente, dado un conjunto de evidencia $E \not\subset X$, un nervio de E es un grafo conectado μ donde cada elemento de E es un vértice en μ y vice versa.

El esqueleto con respecto al nervio μ de E es el conjunto de elementos en X tales que se encuentran en medio de cualquier par de vértices conectados en forma directa en μ.

Finalmente, se dice que un operador de generalización es basado en distancia si para todo conjunto de evidencia E, existe un nervio μ conectando los puntos en E tal que el esqueleto de μ esté incluido en la generalización de E. En otras palabras, la generalización de un conjunto de evidencia debe cubrir no sólo los elementos de la evidencia sino también deben existir caminos dentro de la generalización para movernos de un elemento a otro sin salirnos de ella.

¿Cuál es la razón por la que interesa que los elementos intermedios a los elementos generalizados sean también cubiertos por la generalización? Para responder a esta pregunta, consideremos un par de secuencias *abb* y *acc* y la distancia de edición. Dos generalizaciones posibles de estas secuencias son los patrones **bb+*cc* and **a**. Sin embargo, preferimos *a*a** ya que movernos en el espacio métrico desde *abb* hasta *acc* a través del camino $abb \rightarrow_1 ab \rightarrow_1 a \rightarrow_1 ac \rightarrow_1 acc$ es más suave o más directo que el camino $abb \rightarrow_1 acbb \rightarrow_1 accbb \rightarrow_2 acc$. Podemos ver que en el segundo camino, el elemento intermedio *accb* que se encuentra entre *accbb* y *acc* no está incluido en la generalización **bb+*cc*. Esto nos conduce a un paso con una distancia mayor a 1 o nos fuerza a salirnos de la generalización.

Si analizamos el nivel más bajo de consistencia en el marco de HDCC, podremos ver que los operadores de generalización aceptables dan lugar a generalizaciones donde los elementos cubiertos deben encontrase a una distancia de los elementos generalizados que sea menor o igual a la máxima distancia entre los elementos generalizados. Sin embargo, esta definición considera aceptables también aquellas generalizaciones que "dejan huecos" o, dicho de otra manera, que no cubren elementos intermedios los cuales, sin embargo, de acuerdo al concepto de generalización basada en distancia también deberían ser incluidos. Para ilustrar este problema, consideremos por ejemplo el patrón $[1, 3] \cup [9, 11]$ que resulta de la generalización de la evidencia $\{2, 10\}$ en el espacio métrico de los enteros con la diferencia absoluta como distancia. Notemos que esta generalización satisface la condición de aceptabilidad: cualquiera de los elementos en la cobertura de $\Delta(2, 10) = Set([1, 3] \cup [9, 11]) = \{1, 2, 3, 9, 10, 11\}$ se encuentra a una distancia (al menos a uno de los elementos generalizados) que es menor que la (máxima) distancia entre los elementos generalizados ($d(2, 10) = 8$). Sin embargo, podemos ver que algunos elementos intermedios, aquellos entre 3 y 9, se encuentran por fuera de la cobertura del patrón. No parece razonable considerar consistentes con la distancia aquellos patrones que excluyen elementos intermedios a los elementos de la evidencia generalizada ya que precisamente esos elementos son los que constituyen la distancia entre la evidencia generalizada.

La aplicación de las ideas de Estruch a los operadores de generalización Δ* y Δ usados por HDCC resulta bastante directa. Así, es que hemos llamado a este nuevo nivel de consistencia como el nivel de los dendrogramas basados en distancia. A continuación, damos las correspondientes definiciones.

La definición 10 establece la condición necesaria y suficiente para los operadores binarios de generalización basados en distancia. Informalmente, un operador binario de generalización Δ es basado en distancia si y sólo si para cualquier par de puntos en un espacio métrico, cualquier punto intermedio se encuentra incluido en la generalización de los dos puntos.

[1] Dadas dos duplas de números reales (a1, ..., an) y (b1, ...,bn) la distancia de Manhattan d = Σ |ai – bi| i= 1,...,n

DEFINICIÓN 10

Sea (X, d) un espacio métrico. Un operador binario de generalización Δ es basado en distancia sii $\forall\ e_1, e_2, e \in X$: si $d(e_1, e) + d(e, e_2) = d(e_1, e_2)$ entonces $e \in Set(\Delta(e_1, e_2))$. Consideremos las generalizaciones mostradas en la figura 1 (a) para los puntos a y b en el espacio métrico $(\mathcal{R}^2, d)$, siendo d la distancia Euclidea. Sólo G3, G4 y G5 son generalizaciones basadas en distancias con respecto a la distancia d. En G1 y G2 algunos elementos entre a y b se encuentran excluidos de las generalizaciones. Dado que estamos usando la distancia Euclidea, los elementos entre a y b esponden al segmento más corto que conecte a con b en el espacio métrico $(\mathcal{R}^2, d)$. Sin embargo, si considerásemos un espacio métrico diferente, por ejemplo el espacio métrico definido en $\mathcal{R}^2$ por la distancia de Manhattan[1], los elementos entre a y b se encuentran ubicados en el rectángulo delimitado por a y b. Por lo tanto, la única generalización en la figura 1(a) que es basada en distancia con respecto a la distancia de Manhattan es G4, la cual corresponde precisamente al rectángulo delimitado por a y b.

Extendiendo este concepto a los operadores binarios de generalización de patrones, informalmente, podemos decir que un operador binario de generalización de patrones Δ* es basado en distancia si dado cualquier par de patrones p_1, p_2 resultantes de la generalización de dos grupos C_1 y C_2 existen dos nervios μ_1 y μ_2, tales que el esqueleto del nervio que resulta de la unión de μ_1 y μ_2 mas el agregado de un nuevo arco que conecta los puntos de enlace de C_1 y C_2 es cubierto por la generalización de p_1 y p_2 . Sin embargo, si analizamos este concepto en el marco de HDCC, vemos que es suficiente pedir que la generalización cubra sólo los elementos que se encuentran entre los dos puntos de enlace, ya que la cobertura del esqueleto asociado a la unión de los nervios μ_1 y μ_2 está dada por la definición de operador de generalización binario de patrones (una generalización de dos patrones debe cubrir al menos los elementos en la cobertura de los patrones que generaliza). La definición 11 formaliza este concepto.

DEFINICIÓN 11

Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones y $d_{\mathcal{L}}$ una distancia de enlace. Un operador binario de generalización de patrones Δ* es basado en distancia sii $\forall\ p_1, p_2 \in \mathcal{L}, C_1 \subseteq Set(p_1), C_2 \subseteq Set(p_2), e \in X$: $if\ d(e_1, e) + d(e, e_2) = d(e_1, e_2)$ entonces $e \in Set(\Delta^*(p_1, p_2))$, con e_1 y e_2 los puntos de enlace en $d_{\mathcal{L}}(C_1, C_2)$.

La figura 7 muestra diversas generalizaciones de patrones, todas para la misma evidencia $\{a, b, c, d, e, f\}$ en $\mathcal{R}^2$ y usando la distancia Euclidea. En la figura7 (a), el lenguaje de patrones que hemos adoptado es el conjunto de rectángulos mínimos en $\mathcal{R}^2$. La figura muestra dos patrones p_1 y p_2 (el rectángulo que generaliza los puntos f, d y e y el rectángulo que generaliza los puntos a, b y c, respectivamente) y su generalización, dada por el operador de generalización de patrones $\Delta^*(p_1, p_2)$, que corresponde al mayor de los rectángulos mostrados en la figura. Gráficamente, podemos ver que Δ* es basado en distancia ya que todos los puntos intermedios entre los dos puntos de enlace (cualquiera sea la distancia de enlace usada), representados por el segmento que conecta ambos puntos, estará incluido en la región confinada entre las dos líneas punteadas y, por lo tanto, en la generalización de p_1 y p_2, es decir, en el rectángulo más grande.

En la figura 7 (b), hemos usado un lenguaje de patrones diferente y la distancia de enlace simple para la misma evidencia. En este caso, los patrones son uniones de rectángulos mínimos. La figura muestra dos patrones p_1 y p_2, ambos consistiendo, en este caso, de la unión de dos rectángulos. La generalización de p_1 y p_2, $\Delta^*(p_1$ y $p_2)$, en este caso,

está dada por la unión de todos los rectángulos en la figura (incluido el rectángulo de líneas punteadas). Como se puede ver en la figura, los puntos intermedios entre los dos puntos de enlace simple a y e, descriptos por el segmento punteado que los conecta, están incluidos en el rectángulo de líneas punteadas, el que es parte de la generalización $\Delta^*(p_1, p_2)$.

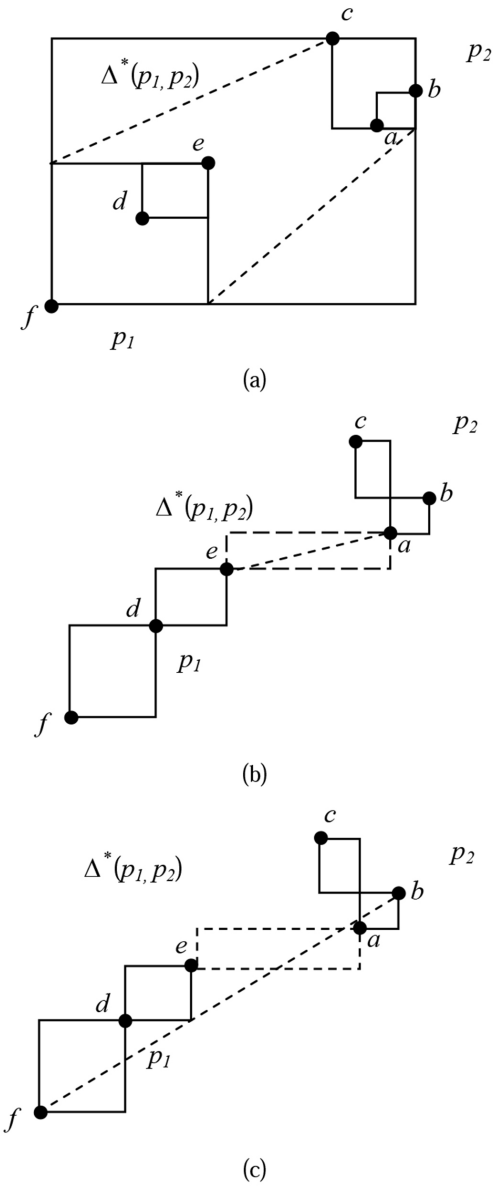

Fig. 7. Ejemplos de generalizaciones basadas en distancias de patrones en Â2.

En la figura 7 (c) hemos cambiado la distancia de enlace al enlace completo. Los puntos de enlace completo, en este caso, están dados por f y b. Como se puede ver en la figura, algunos de los puntos intermedios, descriptos por el segmento punteado que conecta f y b, no caen dentro de la generalización $\Delta^*(p_1, p_2)$, por lo que este operador no es basado en distancia con respecto a la distancia de enlace completo $d^c_{\mathcal{L}}$ y la distancia Euclidea d.

Si consideramos la distancia de Manhattan para los tres ejemplos mostrados en la figura 7, los operadores binarios de generalización de patrones Δ* empleados en los ejemplos también son basados en distancias en los casos (a) y (b) ya que todos los elementos que se encuentran entre los puntos de enlace con respecto a la distancia de Manhattan están incluidos en las respectivas generalizaciones, mientras que en el caso (c), existen puntos intermedios que caerán fuera de la generalización.

Finalmente, diremos que un dendrograma conceptual es basado en distancia si este es el resultado de la aplicación de dos operadores de generalización D* y D basados en distancias con respecto a las distancias d y $d_{\mathcal{L}}$.

Una consecuencia importante de emplear operadores de generalización en HDCC que sean basados en distancias es la existencia de pares de operadores y distancias que, al emplearlos en forma combinada, producirán patrones basados en distancias, llevando al aprendizaje de dendrogramas conceptuales también basados en distancia. En efecto, varios tipos de datos, tanto estructurados como no estructurados, han sido propuestos en (Estruch, 2008), dejándonos abierta la posibilidad de emplear tales pares de operadores y distancias desde la óptica de HDCC. En (Estruch, 2008) podemos encontrar combinaciones consistentes para datos numéricos, nominales, conjuntos, listas, tuplas, grafos, átomos and clausulas.

CONCLUSIONES

La aproximación HDCC al agrupamiento jerárquico basado en distancias brinda un algoritmo genérico que puede ser instanciado para diferentes tipos de datos de manera consistente; esto significa que podemos enriquecer los dendrogramas tradicionales producto del algoritmo de agrupamiento jerárquico aglomerativo tradicional produciendo dendrogramas conceptuales en HDCC y preservando la consistencia entre patrones y distancias, en la medida que contemos con pares compatibles de operadores de generalización y distancias para el tipo de dato involucrado.

Para esto, HDCC nos provee de un marco teórico con un conjunto de herramientas matemáticas, necesarias para diseñar instancias consistentes del algoritmo HDCC. Esto significa que podemos llevar a cabo a priori análisis teóricos del nivel de consistencia alcanzado para cualquier par de operadores de generalización y distancias dados para un cierto tipo de dato.

En este trabajo hemos presentado una extensión a dicho marco teórico en la forma de un nuevo nivel de consistencia entre distancias y generalizaciones. Este nuevo nivel de consistencia, que hemos llamado nivel de los dendrogramas basados en distancia, tiene sus bases en la propuesta realizada por Estruch (2008). En dicho trabajo, el autor analiza la relación entre distancias y generalizaciones y propone un marco donde los métodos simbólicos y los métodos basados en distancias pueden ser integrados de manera consistente y provee un conjunto de operadores de generalización que producen patrones basados en distancias, que pueden ser usados en HDCC, ampliando nuestro conjunto de operadores y distancias compatibles.

Aunque los dendrogramas resultantes de la aplicación de operadores de generalización basados en distancia pueden diferir con respecto a los tradicionales, como es el caso también de los dendrogramas aceptables, los dendrogramas basados en distancia satisfacen una importante propiedad (la intrinsicabilidad) que no es necesariamente alcanzada por las generalizaciones aceptables. La propiedad de intrinsicabilidad garantiza la existencia en el espacio métrico de un segmento de elementos intermedios que conecta la evidencia y que también está incluida en la generalización. De esta manera, podemos movernos, guiados por la distancia subyacente, desde un elemento de la evidencia hacia otro a través de un camino de elementos vecinos sin salirnos de la generalización.

Otra importante contribución de este trabajo es la adición de nuevos pares de operadores y distancias para datos numéricos, nominales, conjuntos, listas, tuplas, grafos, átomos y clausulas, a partir de los resultados teóricos dados en (Estruch, 2008), a los ya disponibles en el marco HDCC (operadores fuertemente acotado bajo cualquier distancia de enlace para datos nominales y numéricos con las distancia discreta y la diferencia absoluta, respectivamente; la composabilidad de tuplas de datos nominales y numéricos (Funes, 2008) (Funes et al., 2009), y un par de operadores de generalización y distancias altamente consistentes para secuencias que producen dendrogramas equivalentes (Funes et al., 2011).

Como parte del trabajo futuro, planeamos estudiar las relaciones existentes entre el nuevo nivel presentado en este trabajo y los tres niveles de consistencia previamente propuestos en HDCC. También se planea seguir enriqueciendo el marco teórico, analizando nuevas combinaciones de distancias y operadores de generalización.

REFERENCIAS

Estruch, “Bridging the gap between distance and generalisation:Symbolic learning in metric spaces”, Tesis (PhD thesis), DSIC, Universidad Politécnica de

Valencia, España, 229, (2008).

Berkhin, "A Survey of Clustering Data Mining Techniques", Grouping Multidimensional Data, pp. 25-71, Springer, (2006).

Jain, Murty, and Flynn, "Data clustering: a review", ACM Comput. Survey, Vol. 31, Nº 3, pp. 264-323, (1999).

Cover and Hart, "Nearest neighbour pattern classification", IEEE Transactions on Information Theory, pp. 13-27, (1967).

MacQueen, "Some methods for classification and analysis of multivariate observations", in Proc. of the 5th Berkeley Symposium on Math. Statistics and Probability, pp. 281–297, University of California Press, (1967).

Michalski, "Knowledge Acquisition Through Conceptual Clustering: A Theoretical Framework and an Algorithm for Partitioning Data into Conjunctive Concepts", Policy Analysis and Information Systems, Vol. 4, Nº 3, pp. 219-244, (1980).

Michalski, and Stepp, "Learning from Observation: Conceptual Clustering", in Michalski et al (eds.) Machine Learning: An Artificial Intelligence Approach, pp. 331-363. TIOGA Publishing Co, (1983).

Fisher, "Knowledge Acquisition Via Incremental Conceptual Clustering", Machine Learning 2: 139-172, Kluwer Acadenfic Publishers, (1987).

Talavera and Béjar, "Generality-Based Conceptual Clustering with Probabilistic Concepts", IEEE Transactions on Pattern Analysis & Machine Inteligence 23(2), (2001).

Golding and Rosenbloom, "Improving rule-based systems through case-based reasoning", Readings in knowledge acquisition and learning: automating the construction and improvement of expert systems, pages 759-764, Morgan Kaufmann, (1993).

Salzberg, "A nearest hyperrectangle learning method", Machine Learning, 6:251-276, (1991).

Domingos, "Unifying instance-based and rule-based induction". Machine Learning, 24(2):141-168, (1996).

Góra and Wojna, "Riona: A classifier combining rule induction and k-NN method with automated selection of optimal neighbourhood", Proc. of the 13th European Connference on Machine Learning (ECML'02), LNCS, pages 111{123. Springer, (2002).

Plotkin, "A note on inductive generalization", Machine Intelligence, 5:153–163, (1970).

Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals", Soviet Physics Doklady, 10:707–710, (1966).

Funes, Ferri, Hernández-Orallo and Ramirez-Quintana, "Hierarchical Distancebased Conceptual Clustering", In Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 349–364. Springer, Heidelberg, (2008).

Funes, "Agrupamiento Conceptual Jerárquico Basado en Distancias, Definición e Instanciación para el Caso Proposicional". Tesis (Master en Ingeniería de Software, Métodos Formales y Sistemas de Información), DSIC, Universidad Politécnica de Valencia, España, (2008).

Funes, Ferri, Hernández-Orallo, Ramírez-Quintana, "An Instantiation of Hierarchical Distance-based Conceptual Clustering for Propositional Learning", LNAI 5476, pp. 637–646. Springer-Verlag Berlin Heidelberg, (2009).

Funes, Ferri, Hernández-Orallo, Ramírez-Quintana, "An Instantiation for Sequences of Hierarchical Distance-based Conceptual Clustering", Proceedings of JAIIO 2011, Córdoba, Argentina, (2011).

Fecha de Presentación: 15/05/15
Fecha de Aprobación: 08/06/15