Universidad Politécnica de Valencia

Departamento de Sistemas
Informáticos y Computación

Tesis de Máster

Agrupamiento Conceptual Jerárquico Basado en Distancias

Definición e Instanciación para el Caso Proposicional

Candidata:

Ana Funes

Directores:

María José Ramírez Quintana
José Hernández Orallo

– Diciembre de 2008 –

Trabajo parcialmente financiado por beca del proyecto
ALFA LERNet AML/19.0902/97/0666/II-0472-FA y la
Universidad Nacional de San Luis

Correo Electrónico de la autora: afunes@dsic.upv.es

Dirección de la autora:

Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera, s/n
46022 Valencia
España

Agradecimientos

Quiero agradecer a todos aquellos que de una forma u otra han contribuido en la realización de este trabajo.

En primer lugar, a mis supervisores, María José Ramírez Quintana y José Hernández Orallo, quienes me han guiado a lo largo de mi estancia en Valencia, aportando valiosas ideas, solventando muchas de mis dudas y contribuyendo a mejorar mi formación. También, quiero agradecer al resto de los integrantes del grupo de Minería de Datos, especialmente a Cèsar Ferri quién siempre ha estado siguiendo muy de cerca mi trabajo y a Vicent Estruch a quién en reiteradas ocasiones recurrí en ayuda.

Asimismo, mi agradecimiento va a todo el resto del grupo ELP por su camaradería y, en especial, a su directora, María Alpuente, por haberme acogido, con la calidez que la caracteriza, en su grupo y haberme brindado un agradable lugar de trabajo.

No puedo dejar de agradecer a mis amigos Daniel Romero, Pedro Ojeda, Alexei Lescaylle, Rafa Navarro y Christophe Joubert por los innumerables momentos gratos compartidos.

También quiero dar las gracias a la Universidad Nacional de San Luis por haber hecho posible mi estancia en la Universidad Politécnica de Valencia y en consecuencia haber contribuido a mi formación académica.

Finalmente, quiero agradecer especialmente a mis seres queridos, a mi hijo Juan y a Aristides, por su paciencia y el apoyo incondicional que siempre me han brindado.

Resumen

Un problema que aparece asociado a algunas técnicas de Minería de Datos es su falta de comprensibilidad. Este es un problema que afecta a las técnicas basadas en distancia, tanto para tareas de agrupamiento así como de clasificación. Aunque varias de estas técnicas han demostrado ser útiles en la práctica al ofrecer buenas predicciones, no brindan una descripción, patrón o generalización que justifique el porqué de la decisión tomada para cada individuo. Así, por ejemplo, si bien es de mucha utilidad conocer que una cierta molécula pertenece a un grupo porque se encuentra cercana a los otros elementos del grupo de acuerdo a una cierta distancia, es de mayor utilidad poder conocer, además, cuáles son las propiedades comunes a todos los elementos del grupo (en este caso, por ejemplo, podrían ser las propiedades químicas o físicas de las moléculas).

La fuente del problema es la dicotomía existente entre las distancias y las generalizaciones. Es bien conocido que las distancias y las generalizaciones dan lugar a dos aproximaciones diferentes en la Minería de Datos y el Aprendizaje Automático. Por un lado, nos encontramos con las técnicas basadas en distancias en donde lo único que necesitamos es contar con una función de distancia o medida de similitud para poder trabajar con ellas. Sin embargo, aunque estas técnicas nos ofrecen esta flexibilidad, no nos proveen patrones o explicaciones que justifiquen las decisiones tomadas. Por el otro lado, tenemos las técnicas basadas en modelos, las cuales, a diferencia de las anteriores, se basan en la idea de que una generalización o patrón descubierto a partir de un conjunto de datos puede ser usado para describir aquellos nuevos datos cubiertos por él.

Al combinar ambas técnicas, un problema importante que surge es conocer si los patrones descubiertos son consistentes con la distancia subyacente. En particular, en el caso de la tarea de agrupamiento, el problema es conocer si, al usar una técnica de agrupamiento basada en distancia, los patrones descubiertos para cada grupo son consistentes con la distancia empleada para construir los grupos, ya que es posible que surjan inconsistencias cuando la noción de generalización y

distancia son tratadas de forma independiente. Con esto queremos significar que para un conjunto de ejemplos y una generalización de los mismos, se espera que aquellos ejemplos que se encuentren cercanos en un espacio métrico de acuerdo a su distancia sean cubiertos por la generalización, mientras que aquellos que estén lejos se espera que se encuentren fuera de la cobertura de la generalización.

En este trabajo analizamos la relación existente entre aquellos grupos obtenidos a partir de un agrupamiento jerárquico tradicional basado en distancias y los conceptos que pueden ser obtenidos por generalización a partir de la jerarquía resultante. Mostramos, a través de ejemplos, que pueden surgir muchas inconsistencias ya que no siempre la distancia subyacente es compatible con el operador de generalización conceptual empleado. Con el fin de sobrellevar este problema, proponemos un nuevo algoritmo que integra el agrupamiento jerárquico basado en distancia con el agrupamiento conceptual. De esta forma, los nuevos dendrogramas obtenidos permiten mostrar claramente cuándo un elemento ha sido integrado a un grupo porque se encuentra "cercano" en el espacio métrico a los otros elementos del grupo o sólo porque se encuentra cubierto por el correspondiente concepto. Consecuentemente, aunque la nueva jerarquía puede diferir con respecto a la original, la trazabilidad métrica es clara.

Teniendo en cuenta esto último, introducimos tres niveles de consistencia entre los operadores de generalización y las distancias empleadas sobre la base de la divergencia existente entre la jerarquía de grupos obtenida por la distancia de enlace y la nueva jerarquía resultante de nuestro algoritmo. Para ello, definimos tres propiedades diferentes para los operadores de generalización cuya satisfacción determina el grado de consistencia existente entre un operador y una distancia.

Finalmente, llevamos a cabo una instanciación para el caso proposicional, donde proponemos un conjunto de pares de distancia y operador de generalización, los cuales usados en forma conjunta trabajan consistentemente para datos numéricos, nominales y tuplas.

Índice general

Índice de figuras

Índice de tablas

1

Introducción

Distancia y generalización son los conceptos subyacentes a dos aproximaciones diferentes en Aprendizaje Automático (o de Máquina). La noción de similitud, que es un concepto más amplio que el de distancia, es la base para muchas técnicas de inferencia inductiva en las cuales se espera que elementos similares se comporten de forma similar. Una distancia, en cambio, no sólo formaliza la noción de similitud entre casos o individuos sino que nos brinda propiedades adicionales del espacio métrico, las cuales son explotadas de manera beneficiosa por muchas técnicas de aprendizaje. Estas técnicas son conocidas como técnicas basadas en distancias. Algunos métodos populares basados en similitud son por ejemplo k-NN (k Nearest Neighbours) [Cover and Hart, 1967], el algoritmo de agrupamiento k-medias [MacQueen, 1967] y los discriminantes de Fisher [Fisher, 1936].

La noción de generalización es también otro concepto importante en Aprendizaje Automático. Todo aprendizaje inductivo involucra algún tipo de generalización. A diferencia de los métodos basados en distancias, algunas aproximaciones se basan en la idea de que una generalización o patrón descubierto a partir de los datos antiguos puede ser usado para describir aquellos nuevos datos cubiertos por el patrón. Estas técnicas son conocidas como basadas en modelos o simbólicas ya que producen un modelo que puede ser interpretado por el usuario. Entre las técnicas más conocidas en aprendizaje supervisado[1] podemos mencionar los árboles de decisión y las reglas de asociación; en aprendizaje no supervisado

[1]Según [Hernández-Orallo et al., 2004], el aprendizaje supervisado se refiere a los métodos predictivos (la clasificación y, algunas raras veces, a la regresión), mientras que el aprendizaje no supervisado a los métodos descriptivos. El método no supervisado por excelencia es el agrupamiento, mientras que el resto de las técnicas descriptivas podría considerarse que no lo son realmente.

podemos citar como ejemplo el agrupamiento conceptual de Michalski [Michalski, 1980; Michalski and Stepp, 1983].

Se puede decir que las técnicas basadas en distancia son bastante intuitivas, y también flexibles en el sentido de que nos basta con contar con una función de distancia adecuada al tipo de datos en cuestión para poder aplicarlas. Sin embargo, estas técnicas no nos ofrecen ningún tipo de patrón o explicación que justifique el porqué de una decisión tomada para un individuo dado. En particular, las técnicas de agrupamiento basadas en distancias acomodan los elementos en grupos basándose en medidas numéricas de similitud entre los elementos; por lo tanto, los grupos resultantes carecen de descripciones conceptuales haciendo difícil su interpretación. Por ejemplo, si bien es de utilidad poder conocer que un documento pertenece a un grupo de documentos porque, de acuerdo a una cierta medida de similitud, se encuentra próximo a los elementos del grupo, es mucho mejor poder conocer además cuáles son las características comunes que describen dichos documentos.

Una aproximación bien conocida para el agrupamiento basado en distancias es el agrupamiento jerárquico [Jain et al., 1999; Berkhin, 2006]. En el agrupamiento jerárquico, los datos son divididos en grupos a lo largo de varios pasos de particionado formando una jerarquía de grupos. Se parte de un único grupo que contiene todos los elementos hasta obtener n grupos que contienen un único elemento. Dependiendo de la manera en que la jerarquía es construida, los agrupamientos jerárquicos pueden clasificarse como aglomerativos (bottom-up) o divisivos (top-down).

Una aproximación diferente a la tarea de agrupamiento es aquella conocida como agrupamiento conceptual definido por Michalski [Michalski, 1980; Michalski and Stepp, 1983]. El agrupamiento conceptual resuelve el problema de interpretación de los grupos formando grupos que pueden ser descritos por propiedades que involucran relaciones sobre un conjunto selecto de atributos. Un sistema de agrupamiento conceptual acepta un conjunto de descripciones de objetos y produce una partición sobre las observaciones. Esas descripciones pueden ser vistas como generalizaciones de los grupos, las cuales son expresadas en la forma de patrones comunes a todos los elementos del grupo.

En esta tesis, presentamos una aproximación general a la tarea de agrupamien-

to jerárquico, de manera tal que usando una distancia para construir la jerarquía de grupos también se producen al mismo tiempo patrones que describen cada uno de los grupos descubiertos. La parte central de esta aproximación es un algoritmo para agrupamiento conceptual jerárquico basado en distancias. El aspecto clave aquí , que no ha sido tenido en cuenta por otros métodos de agrupamiento conceptual que usan distancias, es considerar si la jerarquía inducida por una distancia y los patrones descubiertos son consistentes o no; es decir, la pregunta a hacernos es si todos los elementos cubiertos por un patrón se encuentran próximos en el espacio métrico de acuerdo a la distancia subyacente. Para responder a esta pregunta, en primer lugar, es necesario poder mostrar gráficamente y de forma clara cuándo esta situación ocurre. Esto nos ha llevado al desarrollo de una nueva representación gráfica de los dendrogramas[2], los cuales hemos llamado dendrogramas conceptuales. Por otro lado, también necesitamos poder analizar a priori si dichas inconsistencias aparecerán o no. Esto último ha dado lugar al desarrollo de tres niveles de consistencia entre distancias y generalizaciones, así como a la definición de propiedades que aseguran, en un mayor o menor grado, que el agrupamiento conceptual también refleja la distribución de los ejemplos en el espacio métrico. Esto significa que si, para un problema dado, somos capaces de demostrar alguna de estas propiedades como ciertas, sabremos de antemano que la jerarquía de grupos resultante es al mismo tiempo consistente con la distancia y los conceptos expresados por cada patrón en la jerarquía.

1.1. Estado del arte

Esta lista no exhaustiva de trabajos constituye una panorámica de los principales trabajos relacionados con el tema de esta tesis. Dado que el objetivo de la misma es proveer un marco conceptual para la integración del aprendizaje simbólico a una técnica de agrupamiento basada en distancias, revisaremos aquellas contribuciones más relevantes a ambas aproximaciones.

Existen, en la literatura, varios algoritmos de agrupamiento que generan descripciones de conceptos. Por un lado, tenemos aquellas técnicas tradicionales que

[2]Un dendrograma es una representación gráfica de un árbol, usada para representar la jerarquía de grupos resultante del agrupamiento jerárquico.

provienen del área del agrupamiento conceptual, como por ejemplo CLUSTER/2 [Michalski and Stepp, 1983] y COBWEB [Fisher, 1987]. Por otro lado, nos encontramos con aquellas que, usando un subconjunto de la lógica de primer orden, aplican técnicas tradicionales de agrupamiento basadas en distancias. Entre las técnicas de este segundo grupo, podemos citar KBG [Bisson, 1992], RDBC [Kirsten and Wrobel, 1998], C 0.5 [De Raedt and Blockeel, 1997], TIC [Blockeel et al., 1998] y COLA-2 [Emde, 1994a], entre otras.

Nuestra propuesta es cercana a KBG y RDBC en el sentido que todos usan una estrategia bottom-up basada en distancias; sin embargo, RDBC es puramente basado en distancias, es decir, no emplea operadores de generalización. La principal diferencia con estos dos algoritmos está en el lenguaje de representación. Mientras KBG y RDBC están basado en un subconjunto de una lógica de primer orden, nuestra aproximación es general y, en consecuencia, aplicable a cualquier lenguaje de representación. KBG, si bien usa un algortimo jerárquico de agrupamiento al mismo tiempo que generaliza, reemplaza cada uno de los grupos por su correspondiente generalización ya que, en este caso, los patrones y la evidencia se encuentran expresados en el mismo lenguaje, por lo que la estructura resultante contiene en las hojas los elementos de la evidencia y en los nodos intermedios las generalizaciones. Dado que las coberturas de las generaliziones pueden solaparse, la estructura del clustering es un grafo acíclico dirigido y no un árbol. RDBC, en cambio, al igual que nuestra aproximación, se encuentra basado en el algoritmo jerárquico aglomerativo. Para calcular la similitud entre los ejemplos, RDBC usa la función de similitud de RIBL [Emde and Wettschereck, 1996], muy similar a la empleada por KBG.

C 0.5 se basa en el sistema de ILP TILDE [Blockeel and De Raedt, 1998] (el cual induce, a partir de ejemplos clasificados, árboles de decisión lógicos) para llevar a cabo la tarea de agrupamiento. En lugar de seleccionar los literales de partición sobre la base de la entropía o el error esperado valiéndose de la clase, C 0.5 selecciona, el literal de partición que maximiza la distancia entre los dos grupos y para lo cual debe valerse además de un operador de refinamiento bajo θ-subsunción [Plotkin, 1970; Muggleton and De Raedt, 1994]. De esta manera, el aprendizaje de árboles de decisión es visto como un caso especial de aprendizaje de jerarquía de conceptos. En un árbol de decisión para agrupamiento, cada hoja

corresponde a un grupo de ejemplos; el resto de los nodos corresponde a un predicado pero también a una descripción de los subgrupos formados en cada nodo, formando de esta forma una taxonomía o jerarquía de conceptos. Para calcular la distancia entre los grupos C 0.5 emplea una distancia que computa el prototipo de cada grupo y se basa en el uso de una distancia entre elementos proporcionada por el usuario, la cual puede tener en cuenta o no la clase, determinando de esta forma si el proceso de aprendizaje es supervisado o no. TIC (Top down Induction of Clustering trees) es una actualización de C 0.5 donde se le agraga un método de poda del árbol.

COLA-2 es un sistema que implementa la aproximación CCG (conceptual-clustering-based generalisation) [Emde, 1994b]. Consiste en un primer paso donde se capturan las similitudes entre las instancias no clasificadas para encontrar un conjunto de clases haciendo uso de de KBG. El segundo paso, que consiste en seleccionar una clase, puede ser entendido como una búsqueda en el espacio de posibles generalizaciones dado por el grafo de generalizaciones que resulta del primer paso.

CLUSTER/2 es un sistema de agrupamiento conceptual conjuntivo donde cada grupo es descrito por un único concepto conjuntivo. Se encuentra restringido a un número fijo de atributos nominales sobre los cuales se pueden establecer relaciones de orden. Durante el proceso de apendizaje existe un espacio de hipótesis compuesto de conjunciones de expresiones relacionales sobre los atributos (por ejemplo, $x1 = rojo$, $x1 \leq azul$). CLUSTER/2 usa una función de calidad para encontrar los mejores k grupos. Dicha función se basa en un balance entre la complejidad y la dispersión, donde la complejidad es una medida del número de conjunciones en cada concepto (existiendo un concepto por grupo) y la dispersión es el número de ejemplos no cubiertos por cada concepto. El proceso consiste en elegir al azar inicialmente k ejemplos, llamados semillas. Usando el operador de generalización $m - boundstar$ se recorre el espacio de hipótesis buscando las $m_i \leq m$ mejores descripciones más generales que cubren cada una de las semillas e_i pero no a las otras, para luego especializar cada una de estas descripciones convenientemente. A partir de los $m_1 \times m_2 \times ... \times m_k$ posibles conceptos el algoritmo se queda con aquél cuya calidad es mejor de acuerdo a la función de calidad. El proceso es repetido eligiendo k nuevas semillas tales que sean cubiertas por el

concepto elegido, hasta que se alcanza un número de iteraciones especificadas por el usuario o la calidad de los conceptos no puede mejorarse.

COBWEB es un sistema de agrupamiento conceptual incremental que organiza los ejemplos en un árbol de clasificación. Cada nodo representa una clase y es rotulada por un concepto probabilistico que sumariza las distribuciones de atributo-valor de los objetos clasificados bajo el nodo. Utiliza una función de calidad, de naturaleza probabilistica también, que se basa en la medida CU (por Category Utility) de [Gluck and Corter, 1985] y que mide el número esperado de valores de atributos que pueden ser adivinados correctamente para un elemento arbitrario perteneciente a uno de los k grupos. El árbol de conceptos es inferido incremetalmente, ubicando cada nuevo ejemplo en una categoría (grupo) existente o creando una nueva para lo cual hace uso de operadores que permiten (a) ubicar un objeto en una clase existente; (b) crear una nueva clase; (c) unir dos nodos y (d) separar nodos y de la función de calidad, todo esto a medida que se lleva a cabo una búsqueda hill-climbing en todo el espacio jerárquico de conceptos.

Concluyendo, podemos decir que nuestra propuesta es diferente a todas estas técnicas de agrupamiento conceptual que usan medidas de similitud no sólo en que la nuestra es general, y en consecuencia, aplicable a cualquier tipo de dato, siempre que dispongamos de un operador de generalización y distancia adecuada, sino que viene acompañada de una nueva representación gráfica que permite visualizar a posteriori las divergencias entre el operador de generalización y la distancia empleada, además de contar con la posibilidad de analizar a priori las condiciones que deben ser satisfechas a fin de asegurar que los dendrogramas conceptuales son consistentes con la distancia subyacente.

Nuestro trabajo se encuentra fuertemente relacionado al trabajo presentado en [Estruch, 2008], donde el autor analiza la relación entre distancias y generalizaciones y propone un marco donde ambos paradigmas pueden ser integrados de forma consistente. En ese trabajo el análisis es llevado a cabo para operadores de generalización n-arios definidos sobre la evidencia, es decir, sobre conjuntos de ejemplos y no sobre patrones que describen la evidencia como es realizado en esta tesis.

1.2. Contribuciones de esta tesis

La principal contribución de este trabajo es el desarrollo de un marco conceptual para integrar de forma general y práctica el agrupamiento jerárquico basado en distancias y el agrupamiento conceptual, dando lugar a un algoritmo de agrupamiento jerárquico que es en sí un operador de generalización n-ario construido a partir de operadores de generalización binarios definidos sobre patrones y elementos de un espacio métrico.

Adicionalmente, vale aclarar que nuestra aproximación es general, en el sentido que puede ser aplicada a cualquier distancia, lenguaje de patrones y operadores de generalización. Consecuentemente, la idea es directamente aplicable a datos estructurados. Una posible instanciación nos podría brindar, por ejemplo, descripciones o generalizaciones para grupos de fórmulas atómicas resultantes de la aplicación del operador de generalización menos general (lgg) de Plotkin [Plotkin, 1970] al mismo tiempo que el proceso de agrupamiento usara una distancia para átomos adecuada, tal como por ejemplo la distancia definida en [Ramon et al., 1970]. Otra instanciación directa podría ser, por ejemplo, para listas usando expresiones regulares como lenguaje de patrones junto a la distancia de edición. En esta tesis presentamos un caso concreto de instanciación de nuestro marco para datos estructurados: tuplas de datos numéricos y nominales. Proponemos un juego de operadores y distancias para cada uno de estos tipos de datos, tras analizar formalmente cuáles de las propiedades que definimos en nuestro marco son satisfechas por ellos. Mostramos, además, por medio de experimentos para el caso proposicional, que nuestra aproximación nos permite no sólo obtener descripciones útiles de cada uno de los grupos en la jerarquía sino que se preserva al mismo tiempo la calidad del agrupamiento.

A continuación se presentan las contribuciones originales que configuran el cuerpo de esta tesis de master así como las publicaciones a las que han dado lugar.

Algoritmo de agrupamiento conceptual jerárquico basado en distancias (ver capítulo 3)

Tomando como base el algoritmo tradicional de agrupamiento jerárquico basado en distancias propusimos una extensión al mismo, que permitiera obtener con-

ceptos asociados a cada uno de los grupos al mismo tiempo que se construye la jerarquía.

Dicha extensión se basa en una simple modificación al algoritmo base, consistente en obtener un patrón o generalización para cada nuevo grupo inducido por la distancia utilizando, inicialmente, un operador de generalización binario y, posteriormente, otro operador binario de generalización de patrones, los cuales son dados como parte de las entradas del algoritmo. A medida que el algoritmo obtiene cada uno de estos patrones, va reorganizando los grupos de manera tal que elementos de aquellos grupos en la jerarquía que son completamente cubiertos por el patrón son incorporados al nuevo grupo inducido por la distancia.

Esta simple modificación permite obtener para cada grupo su correspondiente descripción conceptual lo que aporta una mayor comprensión de los mismos, a la vez que la reorganización de los grupos aporta una mayor consistencia entre la distancia y los operadores de generalización en aquellos casos donde no la hubiere.

Consistencia entre distancias y operadores de generalización (ver capítulo 4)

Con el fin de conocer a priori cuán consistentes resultarán, en el marco de nuestro algoritmo de agrupamiento, los operadores de generalización con una distancia dada, hemos definido tres niveles diferentes de consistencia basándonos en las divergencias existentes entre el dendrograma resultante del agrupamiento conceptual jerárquico (dendrograma conceptual) y el correspondiente obtenido por el algoritmo tradicional de agrupamiento jerárquico.

El máximo grado de consistencia entre la distancia subyacente y el operador de generalización viene dado cuando el dendrograma conceptual resultante bajo una distancia y operador dados es equivalente al tradicional. El siguiente nivel que hemos definido es aquel en el que los dendrogramas conceptuales preservan el orden de los grupos con respecto al dendrograma tradicional. Finalmente, el nivel de menor consistencia es aquel que hemos llamado de aceptabilidad, en donde los dendrogramas conceptuales son considerados aceptables siempre que sean el resultado del uso de operadores de generalización aceptables, es decir, operadores que generen patrones donde los nuevos elementos cubiertos por un patrón no se encuentren nunca a una distancia mayor que la máxima distancia entre los

antiguos elementos.

Simultáneamente para cada uno de los niveles de consistencia hemos dado las condiciones suficientes que un operador de generalización debe satisfacer con respecto a la distancia a fin de obtener dendrogramas conceptuales que sean equivalentes, que preserven el orden o que sean aceptables.

De esta manera, dados distancias y operadores de generalización, podemos analizar a priori qué tipo agrupamiento obtendremos; es decir, podemos calificar el grado de consistencia que tendrán los patrones resultantes de un operador de generalización con respecto a la distancia subyacente y optar ante varias alternativas por aquellas combinaciones operador de generalización-distancia que resulten con un mayor grado de consistencia.

Trabajo. A. Funes, C. Ferri, J. Hernández-Orallo, M. J. Ramírez-Quintana. *Hierarchical Distance-based Conceptual Clustering*. Proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2008, Part II, LNAI 5212, pp. 349 - 364. Springer (2008).

Instanciación para agrupamiento proposicional (ver capítulo 5)

Considerando que la aproximación propuesta al agrupamiento conceptual jerárquico es general, el marco puede ser instanciado de diversas formas y para diferentes tipos de datos. En este caso, estudiamos la relación existente entre distancia y operadores de generalización para el agrupamiento proposicional, analizando las propiedades que son satisfechas sobre la base de los tres niveles de consistencia definidos. En primer lugar, analizamos la relación existente entre las distancias y operadores más usuales para datos numéricos y nominales para finalmente hacer lo mismo para tuplas.

Experimentos (ver capítulo 6)

Con el objetivo de comprobar que la calidad de los grupos de las jerarquías obtenidas a partir del algoritmo para agrupamiento conceptual jerárquico no se encuentra degradada, llevamos a cabo una serie de experimentos, utilizando las distancias y operadores de generalización, propuestos en el capítulo 5, para datos

proposicionales. Asimismo, pudimos ilustrar con estos experimentos algunos de los resultados teóricos obtenidos anteriormente.

Trabajo. A. Funes, C. Ferri, J. Hernández-Orallo, M. J. Ramírez-Quintana. *An Instantiation of Hierarchical Distance-based Conceptual Clustering for Propositional Learning*. Sometido a la 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-09), 27 al 30 de Abril de 2009, Bangkok, Thailand.

1.3. Estructura de la tesis

El resto del trabajo se encuentra organizado como sigue.

En el capítulo 2 introducimos algunos conceptos previos y terminología, ambos necesarios para el desarrollo de nuestra aproximación al agrupamiento conceptual jerárquico basado en distancias, la cual es presentada en el capítulo 3. En el capítulo 4 introducimos nuestro marco teórico para el análisis de consistencia entre distancias y operadores de generalización en el ámbito de nuestra aproximación. En el capítulo 5 presentamos una instanciación para agrupamiento proposicional. En el capítulo 6 se describen los experimentos que fueron llevados a cabo sobre la instanciación para tuplas de datos nominales y numéricos y se reportan los resultados obtenidos. Finalmente, el capítulo 7 cierra nuestro trabajo con las conclusiones y el trabajo futuro.

2

Preliminares

En este capítulo, recordamos algunos conceptos preliminares e introducimos alguna terminología, ambos necesarios para el entendimiento del trabajo.

En primer lugar, en la sección 2.1 recordamos la noción de espacio métrico y distancia, así como las definiciones de algunas funciones de distancia que son usadas en el trabajo. A continuación en la sección 2.2 introducimos algunas definiciones necesarias para aclarar la terminología usada a lo largo de esta tesis, en relación a conceptos tales como generalización, operador de generalización embebido en un espacio métrico y su respectiva extensión para patrones. Finalmente, en la sección 2.3 recordamos las nociones básicas del algoritmo de agrupamiento jerárquico aglomerativo sobre el cual se basa nuestra aproximación.

2.1. Espacios métricos

Las distancias son funciones que nos permiten cuantificar la similitud entre dos objetos. Estas funciones transforman pares de objetos en números reales. Cuanto menor es este número, más similares son los objetos.

La idea de asignar distancias a pares de puntos es precisamente lo que da origen a los espacios métricos.

Definición 1 *Un espacio métrico es un par (X, d) donde X es un conjunto no vacío y d es una función real llamada distancia o métrica definida sobre* $X \times X$ *que satisface las siguientes propiedades:*

i) $d(x,y) \geq 0 \; \forall \; x, \; y \in X$ *and* $d(x,y) = 0 \iff x = y$

ii) $d(x,y) = d(y,x) \; \forall \; x, \; y \in X$

iii) $d(x,z) \leq d(x,y) + d(y,z) \; \forall \; x, \; y, \; z \in X$

Es decir, i) las distancias son siempre positivas y el único punto a distancia cero de un punto dado es él mismo; ii) una distancia es una función simétrica; iii) una distancia satisface la desigualdad triangular: la longitud de uno de los lados de un triángulo es menor o igual a la suma de los otros dos lados.

En la terminología del aprendizaje basado en distancias diremos que un elemento está "cerca" en lugar de decir que un elemento es "similar" a otro.

Existen muchas distancias en la literatura para diversos tipos de datos. En [Hutchinson, 2002] se ofrece un catálogo de métricas. Las ecuaciones 2.1, 2.2, 2.3 muestran algunas de las distancias mas conocidas en $\mathbb{R}^n$.

$$Distancia\ Euclidea \quad d(x,y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{2.1}$$

$$Distancia\ de\ Manhattan \quad d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2.2}$$

$$Distancia\ de\ Chebyshev \quad d(x,y) = max_{1<i<n}|x_i - y_i| \tag{2.3}$$

A continuación presentamos algunas otras distancias definidas para otros tipos de datos que son empleadas en el trabajo.

Distancias entre átomos

- La función $d : \mathcal{A}_L \times \mathcal{A}_L \rightarrow \mathbb{Z}^2$, donde $\mathcal{A}_L$ es un conjunto de átomos en un lenguaje de primer orden L, que se encuentra definida en [Ramon et al., 1970], calcula la distancia existente entre un par de átomos de primer orden.

  Esta distancia se basa en la diferencia existente de ambos átomos con respecto a la generalización menos general (lgg) de los mismos. Su definición está dada por la ecuación 2.4, donde a_1, a_2 son dos átomos en $\mathcal{A}_L$.

  $$d(a_1, a_2) = size(a_1) - size(lgg(a_1, a_2)) + size(a_2) - size(lgg(a_1, a_2)) \tag{2.4}$$

  La función $size(a)$ se define como un par (F, V) donde F es el números de símbolos de predicados y de función que curren en el átomo a, y V es la suma de las ocurrencias al cuadrado de cada variable en a.

  Cuando el lgg de dos átomos a_1 y a_2 se encuentra indefinido lo representamos mediante el símbolo $\top$, que es considerado el elemento más general, de manera que $size(lgg(a_1, a_2)) = size(\top) = (0, 1)$.

Ejemplo 1 *Los siguientes ejemplos ilustran el uso de esta distancia.*

$$\begin{aligned}
d(p(x),q(a)) &= size(p(x)) - size(\top) + size(q(a)) - size(\top) \\
&= (1,1) - (0,1) + (2,0) - (0,1) = (3,-1) \\
d(p(x),q(x)) &= size(p(x)) - size(\top) + size(q(x)) - size(\top) \\
&= (1,1) - (0,1) + (1,1) - (0,1) = (2,0) \\
d(p_1(x,y,y),p_1(a,x,x)) &= size(p_1(x,y,y)) - size(p_1(x,y,y)) + \\
&\quad + size(p_1(a,x,x)) - size(p_1(x,y,y)) \\
&= (1,5) - (1,5) + (2,4) - (1,5) = (1,-1) \\
d(p_1(x,y,y),p_1(a,x,y)) &= size(p_1(x,y,y)) - size(p_1(x,y,z)) + \\
&\quad + size(p_1(a,x,y)) - size(p_1(x,y,z)) \\
&= (1,5) - (1,3) + (2,2) - (1,3) = (1,1)
\end{aligned}$$

Aunque estas distancias son expresadas como pares de enteros, un orden lexicográfico es definido sobre el conjunto de pares el cual hace posible su comparación:

$$(F_1,V_1) < (F_2,V_2) \Longleftrightarrow F_1 < F_2 \vee (F_1 = F_2 \wedge V_1 < V_2)$$

De acuerdo a este orden, los valores resultantes en el ejemplo 1 muestran que $p(x)$ está mas cerca de $q(x)$ que de $q(a)$ y que $p_1(x,y,y)$ está más cerca de $p_1(a,x,x)$ que de $p_1(a,x,y)$.

- Otra distancia para átomos es la función $d : \mathcal{E}_L \times \mathcal{E}_L \to \mathbb{R}^2$, propuesta por Shan-Hwei Nienhuys-Cheng en [Cheng, 1997] que se encuentra definida para el conjunto $\mathcal{E}_L$ de átomos básicos y términos básicos de un lenguaje de primer orden L y que está dada por la ecuación 2.5.

$$\begin{aligned}
d(e,e) &= 0, \quad \forall e \in \mathcal{E}_L \\
d(p(s_1,...,s_n),q(t_1,...,t_m)) &= 1, \quad p \neq q \\
d(p(s_1,...,s_n),p(t_1,...,t_n)) &= \tfrac{1}{2n}\textstyle\sum_{i=1}^{n} d(s_i,t_i)
\end{aligned} \tag{2.5}$$

Ejemplo 2

$$\begin{aligned}
d(q(a,a,a),q(a,b,b)) &= \tfrac{1}{2\times 3}(d(a,a)+d(a,b)+d(a,b)) = \tfrac{1}{6}(0+1+1) = \tfrac{1}{3} \\
d(q(a,a,a),p(b,a,b)) &= 1
\end{aligned}$$

2.2. Generalización

Intuitivamente, la generalización de un conjunto finito de elementos E en un espacio métrico (X, d) podría ser definida extensionalmente como un conjunto que contiene a E. Sin embargo, este tipo de definición por extensión no nos dice nada acerca del concepto o *patrón* que describe a todos los elementos de E. Por ejemplo, una generalización expresada como el conjunto de strings $ab, abab, ababab, abababab...$ puede no resultar tan clara como la descripción o patrón $(ab)^+$.

En el ejemplo anterior podemos observar que hemos usado un lenguaje $\mathcal{L}$ para dar las descripciones conceptuales distinto del lenguaje 2^X usado para las descripciones por extensión. Asimismo, podemos notar que, dependiendo del lenguaje de patrones $\mathcal{L}$ que elijamos, podremos expresar ciertas generalizaciones pero no otras. Por ejemplo, el patrón que describe las cadenas de caracteres que contienen el mismo número de a's que de b's no puede ser expresado por el mismo lenguaje que usamos para el ejemplo anterior, es decir, el lenguaje de las expresiones regulares.

Diremos entonces que un patrón $p \in \mathcal{L}$ puede ser considerado una forma de representar los elementos de E y, al mismo tiempo, denotaremos con $Set(p)$ a los elementos descritos o cubiertos por p. Esto nos lleva a la noción de cobertura y de operador de generalización.

Nos referiremos a $Set(p)$ como la *cobertura* del patrón p. Consecuentemente, diremos que un elemento $x \in X$ *es cubierto* por un patrón p si $x \in Set(p)$ y que p es una *generalización* de un conjunto E si y sólo si $E \subseteq Set(p)$.

Para obtener una generalización o patrón expresado en un lenguaje de patrones $\mathcal{L}$, dada una evidencia E, necesitamos contar con una función que nos transforme los elementos de la evidencia en patrones que describan al menos esos elementos. Esta idea es formalizada por la definición 2.

Definición 2 *Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones y $E \subseteq X$.*

Un operador de generalización es una función $\Delta_X : 2^X \to \mathcal{L}$ tal que $\Delta_X(E) = p$, donde $p \in \mathcal{L}$, $E \subseteq Set(p)$.

Por ejemplo, una posible generalización para un conjunto de puntos en $\mathbb{R}^2$ puede ser definida como el mínimo rectángulo que incluye todos los puntos en el

conjunto (ver figura 2.1). En este caso, el lenguaje de patrones $\mathcal{L}$ viene dado por el conjunto de los rectángulos.

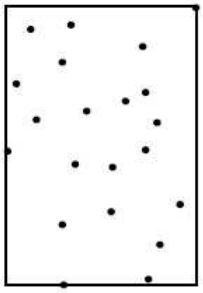


Figura 2.1: Una posible generalización de un conjunto de puntos en $\mathbb{R}^2$.

Si restringimos E a conjuntos de dos elementos, denotamos con Δ al operador binario de generalización . En la definición 3 damos su formalización.

Definición 3 *Sea (X, d) un espacio métrico y $\mathcal{L}$ un lenguaje de patrones.*

Un operador binario de generalización es una función $\Delta : X \times X \to \mathcal{L}$ tal que dados $x_1, x_2 \in X, \Delta(x_1, x_2) = p$, donde $p \in \mathcal{L}$, $x_1 \in Set(p)$ y $x_2 \in Set(p)$.

Por lo tanto, podemos decir que un operador binario de generalización simplemente transforma dos elementos de X en un patrón que los representa.

La figura 2.2 muestra cinco posibles generalizaciones de dos puntos en el espacio métrico $(\mathbb{R}^2, d)$, donde d es la distancia Euclídea.

Figura 2.2: Cinco posibles generalizaciones de dos puntos en $\mathbb{R}^2$.

De igual manera, así como generalizamos conjuntos de elementos o pares de elementos, podemos generalizar pares de patrones ya que los patrones describen elementos.

Definición 4 *Sea (X, d) un espacio métrico y $\mathcal{L}$ un lenguaje de patrones.*

Un operador binario de generalización de patrones es una función $\Delta^* : \mathcal{L} \times \mathcal{L} \to \mathcal{L}$ tal que dados $p_1, p_2 \in \mathcal{L}$, $\Delta^*(p_1, p_2) = p$, donde $Set(p_i) \subseteq Set(p)$ $(i \in 1, 2)$.

La definición 4 establece que una generalización de dos patrones debe describir al menos los mismos elementos descritos por ambos patrones.

Podemos notar que para aquellos casos en los que el lenguaje de patrones $\mathcal{L}$ es igual al conjunto X, como ocurre por ejemplo con el operador *lgg* para átomos, los operadores Δ^* y Δ podrían ser definidos de igual manera.

En la figura 2.3 mostramos una posible generalización para dos patrones p_1 y p_2 en un lenguaje de patrones $\mathcal{L}$, donde $\mathcal{L}$ es el conjunto de rectángulos de ejes paralelos. La generalización $\Delta^*(p_1, p_2)$ se encuentra definida como el mínimo rectángulo que cubre los patrones (rectángulos) p_1 y p_2.

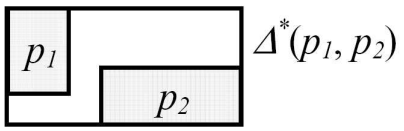


Figura 2.3: Una generalización de dos patrones en $\mathbb{R}^2$.

2.3. Agrupamiento jerárquico

Los algoritmos de agrupamiento jerárquico [Johnson, 1987] construyen una jerarquía de grupos, llamada *dendrograma*, a partir de los elementos individuales que conforman la evidencia. Las hojas de esta jerarquía corresponden a los elementos individuales, mientras que los nodos intermedios corresponden a subconjuntos de ejemplos que particionan la evidencia.

En la figura 2.4 se muestra un ejemplo de un árbol de agrupamiento o dendrograma. El nodo raíz contiene todos los elementos de la evidencia, en este caso $E = \{A, B, C, D, E, F, G\}$. Cada nodo en el árbol se encuentra a un nivel diferente, el cual viene dado por la distancia existente entre los grupos. En general, para una evidencia formada por n ejemplos, el árbol tendrá n niveles, donde cada nivel i determina un agrupamiento K_i diferente.

- Nivel 1: $K_1 = \{A, B, C, D, E, F, G\}$
- Nivel 2: $K_2 = \{\{A, B, C, D\}, \{E, F, G\}\}$
- Nivel 3: $K_3 = \{\{A, B\}, \{C, D\}, \{E, F, G\}\}$

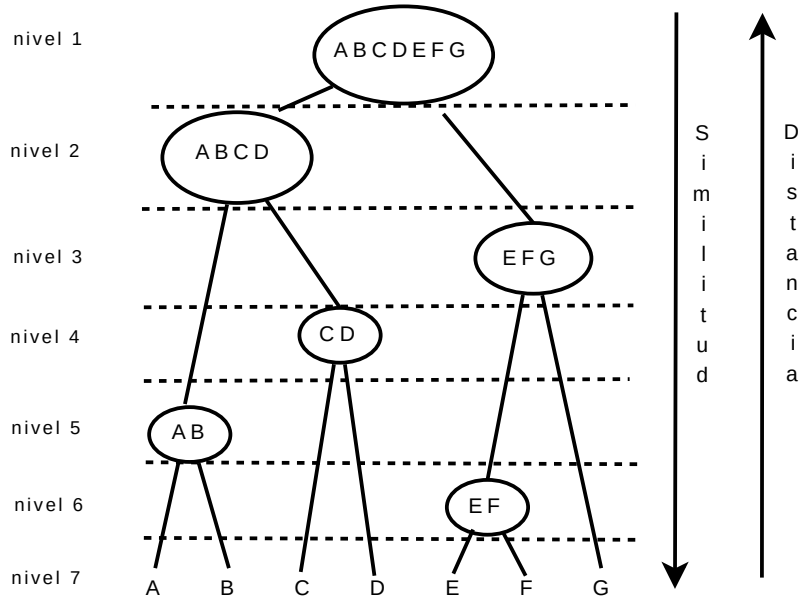


Figura 2.4: Un árbol de agrupamiento o dendrograma y sus n niveles.

- Nivel 4: $K_4 = \{\{A, B\}, \{C, D\}, \{E, F\}, \{G\}\}$
- Nivel 5: $K_5 = \{\{A, B\}, \{C\}, \{D\}, \{E, F\}, \{G\}\}$
- Nivel 6: $K_6 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E, F\}, \{G\}\}$
- Nivel 7: $K_7 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}, \{G\}\}$

Los algoritmos de agrupamiento jerárquico pueden ser clasificados en dos tipos, dependiendo de la forma en que construyen la jerarquía de grupos.

Por un lado tenemos el agrupamiento *divisivo* o top-down en donde, partiendo de la raíz, la cual corresponde a un único grupo con todos los elementos de la evidencia, se van dividiendo los grupos hasta que cada grupo contienen un único elemento que corresponde a una hoja del árbol.

Por otro lado, tenemos el agrupamiento *aglomerativo* o bottom-up, cuando la jerarquía se construye empezando por las hojas hasta llegar a la raíz. Inicialmente, cada hoja corresponde a un grupo formado por un único elemento, los cuales se van uniendo hasta obtener un único grupo que contiene toda la evidencia.

Los grupos son unidos sobre la base de la distancia existente entre ellos, para lo cual se emplea una distancia entre grupos referida como la *distancia de enlace o*

enlazado. Existen diversas distancias de enlace. Usualmente, la distancia de enlace entre dos grupos C_1 y C_2 viene determinada por la máxima distancia entre los elementos de cada grupo, y es conocida como *distancia de enlace completa* y que denotaremos a lo largo de este trabajo como d_L^c; también puede estar dada por la mínima distancia entre los elementos de los grupos (*distancia de enlace simple*, d_L^s); por la distancia media entre los elementos de los grupos (*distancia de enlace a la media*, d_L^a) o también por la mínima distancia entre los prototipos de los grupos (*distancia de enlace a los prototipos*, d_L^p), entre otras.

Formalmente, dichas distancias se definen como sigue[1]:

Sea (X, d) un espacio métrico, $C_1 \subseteq X$, $C_2 \subseteq X$ dos grupos de elementos de X,

$d_L^c(C_1, C_2, d) = max\{\, d(x, y) : x \in C_1,\, y \in C_2 \,\}$

$d_L^s(C_1, C_2, d) = min\{\, d(x, y) : x \in C_1,\, y \in C_2 \,\}$

$d_L^a(C_1, C_2, d) = \frac{1}{|C_1||C_2|} \sum_{x \in C_1} \sum_{y \in C_2} d(x, y)$

$d_L^p(C_1, C_2, d) = d(x_1, x_2)$, con x_1 y x_2 los prototipos[2] de C_1 y C_2

A lo largo del trabajo usaremos d_L para referirnos a cualquiera de ellas.

Como hemos mencionado, en el agrupamiento jerárquico aglomerativo, el proceso de agrupamiento comienza por las hojas de la jerarquía. Veamos un poco más en detalle cómo funciona este algoritmo. Cada una de las hojas de la jerarquía corresponde a un grupo con un único elemento de la evidencia. En el paso siguiente, se unen aquellos dos grupos C_1 y C_2 que se encuentran más próximos de acuerdo a la distancia de enlace d_L en uso, para formar un nuevo grupo $C_{1,2}$ que pasa a ser el padre de C_1 y de C_2 en el árbol. De esta manera, el nuevo conjunto de grupos pasa a estar formado por $C_{1,2}$ mas todas las hojas exceptuando a C_1 y a C_2. Este proceso es repetido hasta que el conjunto de grupos ha reducido a un solo grupo que contiene a todos los elementos de la evidencia.

[1] $|E|$ denota la cardinalidad del conjunto E.

[2] El *prototipo* de un conjunto es aquel elemento que se encuentra en el centro del conjunto con respecto al espacio métrico; es decir, que minimiza la suma de las distancias al resto de los elementos del conjunto. Si dicho elemento no pertenece al conjunto, se le denomina *centroide*.

La forma tradicional de representar la jerarquía de grupos no es la de la figura 2.4 sino que en la práctica se usa otra representación gráfica equivalente. En la figura 2.5 podemos ver dicha representación alternativa para el dendrograma correspondiente al árbol de la figura 2.4.

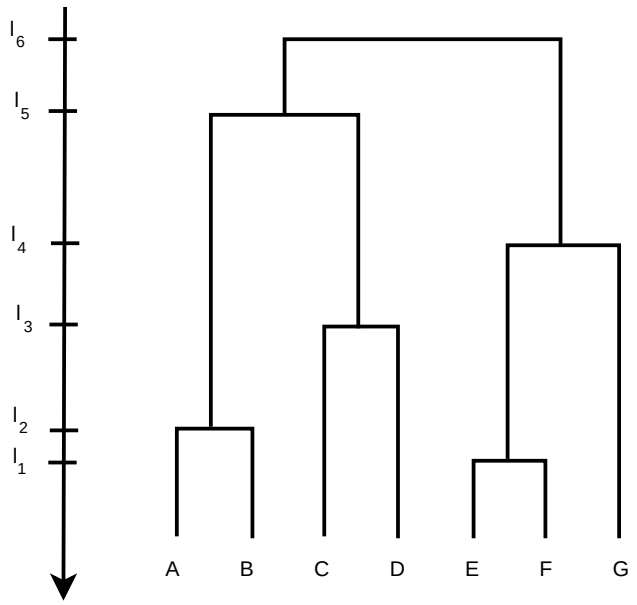


Figura 2.5: Dendrograma equivalente al de la figura 2.4.

En la figura se puede ver que el ejemplo etiquetado E fue el primero unido por la distancia de enlazado al elemento F, formando así el grupo $\{E, F\}$ a distancia de enlace l_1. En el próximo paso se formó el grupo $\{A, B\}$ a distancia de enlace l_2, y así sucesivamente hasta que se obtuvo un único grupo conteniendo toda la evidencia, al unirse los grupos $\{A, B, C, D\}$ y $\{E, F, G\}$ a distancia l_6.

3

Aproximación al agrupamiento conceptual jerárquico basado en distancias

La aproximación al agrupamiento conceptual jerárquico basado en distancias que presentamos en esta sección se basa en el algoritmo de agrupamiento jerárquico aglomerativo.

En primer lugar, en la sección 3.1 presentamos el problema que motivó el desarrollo de nuestra aproximación la cual desarrollamos en la sección 3.2.

3.1. Motivación

Es sabido que existen dos aproximaciones o enfoques diferentes en Aprendizaje Automático. Por un lado tenemos aquellas técnicas, tanto de aprendizaje supervisado como no supervisado, conocidas como *basadas en el modelo*, *impacientes* o *no retardadas* (eager) que producen un modelo explícito a partir de los datos de entrenamiento. Así, por ejemplo, una red neuronal [Haykin, 1998] [Isasi and Galván, 2003], una vez entrenada, es una función o modelo que puede ser utilizada para predecir nuevos ejemplos. Sin embargo, la red neuronal entrenada es una caja negra ya que no podemos saber por qué ante cada nueva evidencia produce un cierto resultado. Es decir, el modelo producido por la red neuronal no es inteligible. No obstante, existen otras técnicas impacientes que producen modelos comprensibles para el usuario, haciendo que los modelos o hipótesis sean expresadas en un lenguaje de patrones más próximo al usuario. Estas técnicas

basadas en el modelo son conocidas como *simbólicas*. Un ejemplo de ellas son los árboles de decisión los cuales pueden describirse como un conjunto de reglas fácilmente comprensibles por el usuario.

Por otro lado, nos encontramos con aquellas técnicas conocidas como *retardadas* o *perezosas* (lazy) en donde no se construye un modelo sino que la técnica actúa ante cada nueva predicción o consulta. Ejemplos de estas técnicas son los métodos CBR (Case-based reasoning) [Sycara et al., 1992] y k-NN [Cover and Hart, 1967]. Un subgrupo importante de estas técnicas son aquéllas conocidas como *basadas en distancias* o *basadas en similitud* en las que la idea subyacente es que elementos similares o cercanos deben comportarse de forma similar. Vale aclarar que esta clasificación no es excluyente y que existen técnicas no retardadas basadas en distancias como por ejemplo los discriminantes de Fisher [Fisher, 1936] en donde una vez encontrados los discriminantes los ejemplos pueden ser despreciados, o k-medias [MacQueen, 1967], en donde una vez encontrada la partición de los ejemplos, podemos usar sus centroides para clasificar nuevos ejemplos.

Las técnicas basadas en distancias, si bien funcionan muy bien para muchos problemas y tienen la ventaja adicional de que pueden ser instanciadas para diversos tipos de datos siempre que se disponga de la distancia adecuada, tienen asociado el inconveniente de la comprensibilidad. La comprensibilidad es un aspecto muchas veces importante a ser considerado en la selección de una hipótesis o modelo ya que en muchos problemas el usuario necesita encontrar modelos inteligibles a partir de los datos. Por ejemplo, en un un compañía financiera es necesario contar con algún tipo de explicación para dar a aquellos clientes a los cuales se les ha rechazado un préstamo; o por ejemplo, en un modelo de diagnóstico, dado que es el médico quién tiene la última palabra, es necesario que éste pueda comprender las razones que ha utilizado el sistema para determinar el tipo de patología que ha predicho.

Teniendo en cuenta las características de ambas aproximaciones, resulta interesante contar con algoritmos que combinen lo mejor de ambas técnicas, es decir, algoritmos que puedan ser adaptados a cualquier representación de los datos y que a la vez nos proporcionen un modelo que pueda ser expresado simbólicamente. Consecuentemente, en este caso, no sólo los datos serán una entrada al

algoritmo sino que también la distancia y operador de generalización deberán ser dados como entrada.

En este trabajo, combinamos ambas técnicas sobre la base del algoritmo de agrupamiento jerárquico aglomerativo. Sin embargo, esta integración, como veremos, debe ser hecha cuidadosamente ya que pueden surgir ciertas inconsistencias entre los datos considerados similares con respecto a la distancia empleada y los datos descritos por los patrones descubiertos por generalización. En otras palabras, puede ocurrir que los datos cubiertos por un patrón no se encuentren cercanos en el espacio métrico de acuerdo a la distancia subyacente.

Para ilustrar el problema consideremos una primera aproximación al agrupamiento conceptual jerárquico basado en distancias que integra ambas técnicas. Esta primera aproximación trabaja sobre la base del algoritmo tradicional de agrupamiento jerárquico aglomerativo. Aquí, los patrones son descubiertos o bien a medida que se forma cada nuevo grupo o bien en un post-proceso, en ambos casos usando un operador de generalización Δ_X sobre cada grupo. El problema viene ilustrado con el ejemplo 3.

Ejemplo 3 *Dado el espacio métrico* (X, d) *donde* X *es el conjunto de las listas finitas de símbolos sobre el alfabeto* $\Sigma = \{a, b\}$ *y* d *es la distancia de edición o distancia de Levenshtein [Levenshtein, 1966].*

Supongamos una evidencia E *formada por cuatro listas de símbolos aa, aab, abb, aabbbbbb. En la figura 3.1 se muestran las cuatro listas y las distancias de edición entre algunas de ellas, considerando el coste de un reemplazo como el coste de una supresión mas una inserción.*

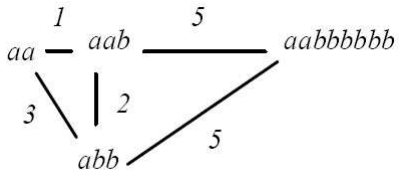


Figura 3.1: Cuatro listas en el espacio métrico (Σ, d).

Supongamos además que estamos usando la distancia de enlace simple y que el lenguaje de patrones $\mathcal{L}$ *usado es el de las expresiones regulares. De acuerdo a la distancia de enlace simple, los primeros dos grupos más cercanos son* $\{aa\}$

y $\{aab\}$*, que son enlazados a distancia 1, luego* $\{aa, aab\}$ *es enlazado con* $\{abb\}$ *a distancia 2 y finalmente* $\{aabbbbbb\}$ *es enlazado a* $\{aa, aab, abb\}$ *a distancia 5, como se muestra en la figura 3.2. En la misma figura podemos ver que el operador de generalización* Δ_X *usado como entrada al algoritmo ha producido como generalización del grupo* $\{aa, aab\}$ *el patrón* $aa*$ *y el patrón* $a*$ *como generalización de* $\{aa, aab, abb\}$ *y de* $\{aa, aab, abb, aabbbbbb\}$*, respectivamente.*

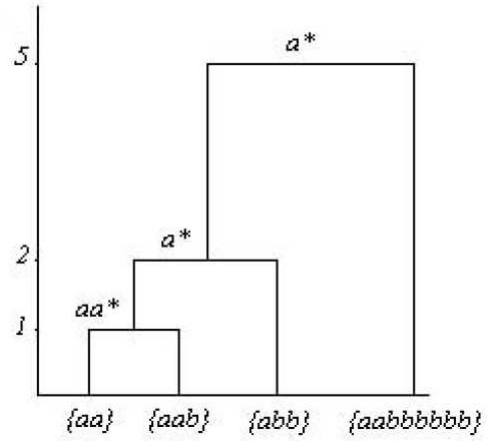


Figura 3.2: Dendrograma resultante bajo enlace simple y patrones, para los ejemplos de la figura 3.1.

Con esta primera aproximación, podemos ver claramente las inconsistencias que pueden ocurrir entre la distancia de edición d, la distancia de enlace simple d_L^s y el operador de generalización Δ_X usados. Si observamos la figura 3.3, el patrón $aa*$ no cubre solamente el grupo $\{aa, aab\}$ sino que también cubre al grupo $\{aabbbbbb\}$, lo que de por sí no sería un problema . Sin embargo, podemos ver que existe una inconsistencia métrica entre los elementos descritos por $aa*$ y los grupos inducidos por la distancia, ya que $aa*$ cubre $\{aabbbbbb\}$ pero no cubre a $\{abb\}$, el cual de acuerdo a la distancia de enlace simple d_L^s en combinación con la distancia de edición d, se encuentra más cerca de $\{aa, aab\}$ de lo que se encuentra $\{aabbbbbb\}$.

Notemos que el proceso de agrupamiento, en esta primera aproximación, es dirigido solamente por las distancias d y d_L, mientras que los patrones obtenidos por la generalización Δ_X de cada grupo no intervienen para nada en la estructura de

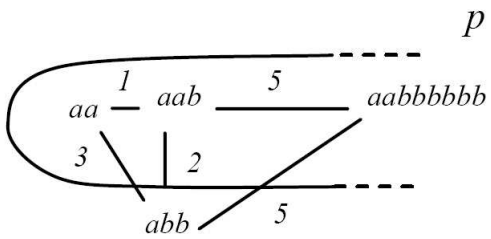


Figura 3.3: Cobertura del patrón $aa*$.

la jerarquía de grupos pudiendo ocurrir, como en este caso, que los agrupamiento descritos por los patrones no coincidan con los formados por las distancias.

En consecuencia, debemos diseñar una manera de mostrar claramente en cada grupo qué elementos han sido atraídos por el patrón por cobertura y cuáles por la distancia, o debemos reconsiderar si los operadores de generalización y las distancias son consistentes entre sí, o ambas cosas.

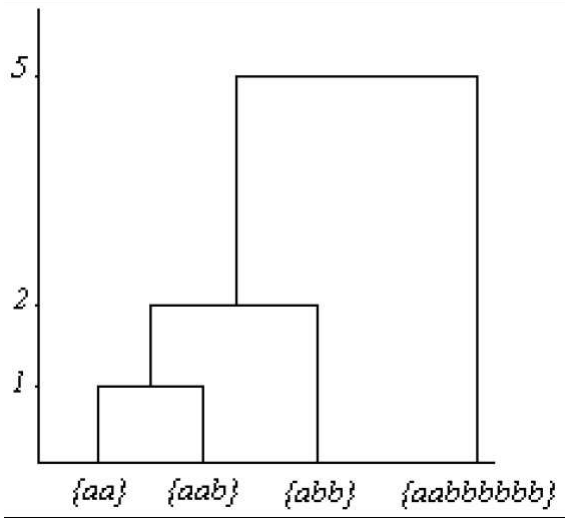
3.2. Algoritmo de agrupamiento HDCC

Como una solución al problema de inconsistencia planteado en la sección 3.1, proponemos nuestra aproximación al agrupamiento conceptual jerárquico basado en distancias.

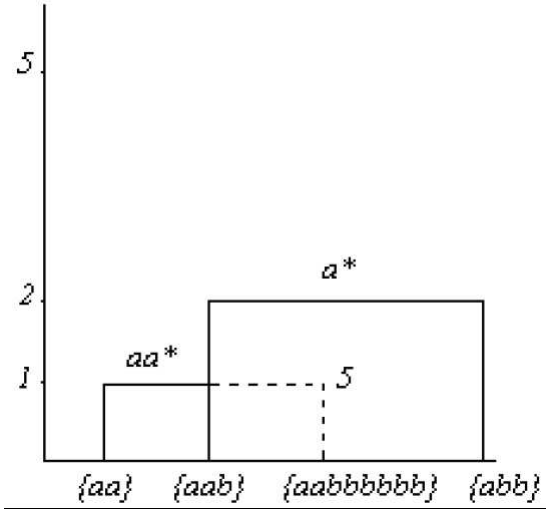
Al igual que en la aproximación anterior, nuestro algoritmo, que hemos llamado HDCC por **H**ierarchical **D**istance-based **C**onceptual **C**lustering, trabaja sobre la base del algoritmo de agrupamiento jerárquico aglomerativo pero, a diferencia de la aproximación mostrada en la sección anterior, HDCC usa operadores binarios de generalización de patrones (ver definición 4) y operadores binarios de generalización definidos para elementos de un espacio métrico (ver definición 3) para construir las generalizaciones de los grupos.

HDCC permite que generalización y distancia trabajen juntos llevando a cabo una simple modificación al algoritmo base. Esta modificación consiste en unir a cada nuevo grupo todos aquellos grupos que se encuentran cubiertos por la generalización del nuevo grupo. De esta manera, los patrones finales proveen una descripción común a todos los elementos que se encuentran próximos de acuerdo a las distancias subyacentes d y d_L y, al mismo tiempo, de aquellos que, aunque

no se encuentran tan próximos como para haber sidos captados por la distancia, se encuentran cubiertos por el patrón.



(a) Dendrograma tradicional



(b) Dendrograma conceptual

Figura 3.4: Dendrogramas tradicional vs. dendrograma conceptual.

Para representar la jerarquía de agrupamientos resultante, usamos un dendrograma extendido que hemos llamado *dendrograma conceptual*. Un dendrograma conceptual no sólo brinda la información tradicional sobre qué elementos se encuentran en cada grupo sino que, además, provee una descripción en la forma de patrón de las propiedades comunes de los elementos de cada grupo. Como

puede observarse, en el dendrograma conceptual de la figura 3.4(b), las líneas sólidas enlazan los grupos unidos por la distancia de enlace d_L, mientras que las líneas punteadas muestran aquellos grupos que han sido enlazados por la generalización. Por ejemplo, el patrón $aa*$ cubre el grupo $\{aa, aab, aabbbbbb\}$, el cual ha sido formado considerando, en primer lugar, la distancia de enlazado entre los grupos $\{aa\}$ y $\{aab\}$ y, en segundo lugar, la cobertura del grupo $\{aabbbbbb\}$ por el patrón $aa*$.

Asimismo, en la figura 3.4 se puede ver la diferencia existente entre el dendrograma resultante del algoritmo tradicional de agrupamiento jerárquico (figura 3.4(a)) y el dendrograma conceptual (figura 3.4(b)) producido por HDCC para el ejemplo en curso.

El algoritmo enfrenta el problema antes expuesto llevando a cabo un test de cobertura a continuación de cada proceso de generalización y reorganizando los grupos para hacerlos consistentes con el último patrón descubierto. Más específicamente, este proceso de cobertura-reorganización consiste, simplemente, en unir a cada nuevo grupo C con patrón p todos aquellos grupos que se encuentran completamente cubiertos por p, es decir, incluidos en $Set(p)$. De esta manera, en esta aproximación, a diferencia de la aproximación inicial mostrada en la sección anterior, estos grupos agregados conceptualmente pueden jugar un papel muy diferente en la construcción de la jerarquía.

El algoritmo 1 muestra un pseudo código para HDCC.

Algoritmo 1 ***HDCC***

Entrada:

$E = \{e_1, e_2, ..., e_n\} \subseteq X$, *una distancia* d, *con* (X, d) *un espacio métrico;*

$\Delta^* : \mathcal{L} \times \mathcal{L} \to \mathcal{L}$ *un operador binario de generalización de patrones;*

$\Delta : X \times X \to \mathcal{L}$ *un operador binario de generalización;*

$d_L : 2^X \times 2^X \times (X \times X \to \mathbb{R}) \to \mathbb{R}$ *una distancia de enlace.*

Salida:

Un árbol T *de grupos y generalizaciones.*

Pasos:

1. $S \leftarrow \{\{e_1\}, \{e_2\}, ..., \{e_n\}\}$.
2. *Insertar* $(\{e_i\}, \Delta(e_i, e_i), 0)$ *como una hoja de* T, *para todo* $\{e_i\}$ *en* S.
3. *Mientras* $S \neq \{E\}$
   1. *Calcular* $d_L(C_i, C_j, d)$ *entre cada par de grupos* $C_i, C_j \in S$ *con* $i < j$, *usando la distancia* d.
   2. *Calcular el patrón* $p_{C_{xy}}$ *del grupo* C_{xy} *como* $\Delta^*(p_{C_x}, p_{C_y})$, *donde* $C_{xy} = C_x \cup C_y$, p_{C_x}, p_{C_y} *son los patrones de* C_x *y* C_y *respectivamente, y* C_x *y* C_y *son los grupos más cercanos en* S *de acuerdo a* d_L.
   3. $S \leftarrow S \cup \{C_{xyz}\}$ *con* $C_{xyz} = C_{xy} \cup C_z$ *y* $C_z = \{e \mid e \in C_i \wedge C_i \in S \wedge C_i \subseteq Set(p_{C_{xy}})\}$.
   4. *Insertar* $(C_{xyz}, p_{C_{xy}}, d_{L_{C_{xy}}})$ *en* T *como el nodo padre de* $(C_x, p_{C_x}, d_{L_{C_x}})$, $(C_y, p_{C_y}, d_{L_{C_y}})$ *y de los nodos* $(C_i, p_{C_i}, d_{L_{C_i}})$ *donde* $C_i \in S$ *y* $C_i \subseteq Set(p_{C_{xy}})$.
   5. $S \leftarrow S - \{C_i\}$ *para todo* C_i *tal que* $C_i \subseteq Set(p_{C_{xy}})$.
4. *Retornar* T.

HDCC acepta las siguientes entradas:

(a) Un conjunto E de n ejemplos y una distancia d de un espacio métrico (X, d).

(b) Δ^*, un operador binario de generalización de patrones definidos sobre un lenguaje de patrones $\mathcal{L}$ compatible[1] con X.

(c) Un operador binario de generalización $\Delta : X \times X \to \mathcal{L}$ para computar las generalizaciones de cada uno de los grupos unitarios iniciales $\{e\}$ como $\Delta(e, e)$.

(d) Una distancia de enlace $d_L : 2^X \times 2^X \times (X \times X \to \mathbb{R}) \to \mathbb{R}$ para calcular las distancias entre dos grupos usando la distancia d entre elementos de X.

La salida es un árbol T donde cada nodo corresponde a un grupo con sus correspondientes patrones y distancias de enlazado (mostradas sobre el eje y del dendrograma). HDCC de hecho actúa como un operador de generalización n-ario sobre cada grupo de la jerarquía.

En el primero y segundo paso, HDCC inicializa el conjunto S y el árbol T. Inicialmente S consiste de n grupos, donde cada uno de ellos contiene un elemento de E, y T es un árbol con n hojas, donde cada hoja es una 3-upla formada por un conjunto unitario $\{e\} \in S$, su correspondiente patrón computado como $\Delta(e, e)$ y la distancia de enlace (en este caso $d_L(\{e\}, \{e\}, d) = 0$).

Los pasos 3.1 a 3.5 son repetidos hasta que S contenga un único grupo. En el paso 3.1, HDCC calcula las distancias de enlace d_L entre los grupos, la cual depende de la distancia d que es pasada como entrada al algoritmo. Por ejemplo, si como distancia de enlace d_L se usa la distancia de enlace completo, la distancia de enlace $d_L(C_i, C_j, d)$ será igual a la máxima distancia d existente entre un par de elementos en C_i y C_j.

Una vez que las distancias de enlace han sido calculadas, en el paso 3.2. HDCC calcula el patrón $p_{C_{xy}}$ como la generalización de los patrones p_{C_x} y p_{C_y} de C_x y C_y, donde C_x y C_y son los grupos que se encuentran a la menor distancia de enlace. $p_{C_{xy}}$ es usado en el paso 3.3 para crear el nuevo grupo C_{xyz}. C_{xyz} estará formado por los elementos de C_x y C_y más todos aquellos elementos que se encuentran en aquellos grupos de S que son completamente cubiertos por el patrón $p_{C_{xy}}$. El nuevo grupo C_{xyz} es agregado a S a su vez que C_x y C_y son eliminados de él.

En el paso 3.4., HDCC inserta un nuevo nodo en el dendrograma T consistente

[1]Por compatible queremos significar que los patrones deben describir elementos en X

en una 3-tupla formada por el grupo C_{xyz}, su generalización $p_{C_{xy}}$ y la distancia de enlace $d_L(C_x, C_y, d)$. Este nodo es insertado como padre de los nodos $(C_x, \_, \_)$ y $(C_y, \_, \_)$ así como de aquellos nodos $(C_i, \_, \_)$, donde cada uno de los grupos C_i se haya completamente cubiertos por $p_{C_{xy}}$.

Los grupos C_i son eliminados de S en el paso 3.5. Notar que son estos grupos C_i los que en la versión gráfica del dendrograma conceptual son enlazados con líneas punteadas.

Podemos observar que cada patrón p en el dendrograma conceptual es calculado como las aplicaciones sucesivas de los operadores Δ y Δ^*. En general tendremos que, para obtener los patrones de toda la jerarquía, Δ debe ser aplicado n veces mientras que Δ^* es aplicado a lo sumo $n-1$ veces. Notar, además, que en un paso del algoritmo, varios grupos pueden ser unidos bajo un mismo patrón, en consecuencia el dendrograma conceptual resultante puede contener menos niveles que el dendrograma tradicional, alcanzándose la raíz más rápidamente.

El ejemplo 4 ilustra cómo trabaja HDCC bajo la distancia de enlace simple d_L^s.

Ejemplo 4 *Sea E la evidencia consistente en el conjunto de puntos en $\mathbb{R}^2$ mostrados por la figura 3.5; $\mathcal{L}$ el lenguaje de patrones empleado, definido como el conjunto de rectángulos de ejes paralelos. Asumimos, además, que los operadores de generalización Δ y Δ^* son mínimos, es decir $\Delta(e_1, e_2)$ y $\Delta^*(p_1, p_2)$ devuelven el mínimo rectángulo que incluye a los puntos e_1 y e_2 y a los rectángulos p_1 y p_2, respectivamente.*

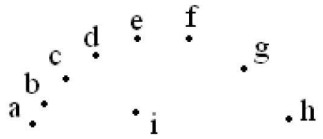
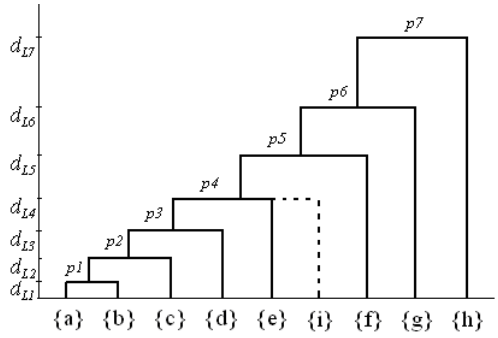
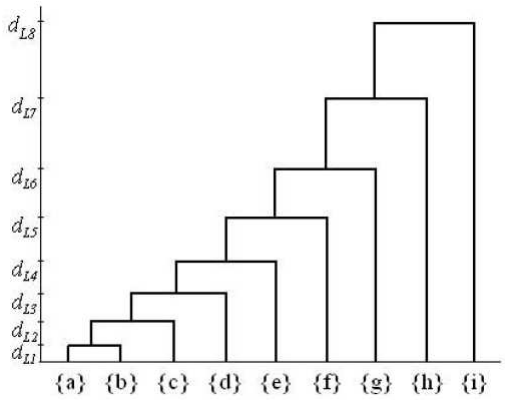


Figura 3.5: Un conjunto de puntos en $\mathbb{R}^2$.

La figura 3.6(a) muestra el dendrograma conceptual resultante. Los grupos $\{a, b\}$, $\{a, b, c\}$ y $\{a, b, c, d\}$ han sido formados dirigidos por la distancia. Sin embargo, el grupo $\{a, b, c, d, e, i\}$ se formó porque la distancia de enlazado unió $\{e\}$ a $\{a, b, c, d\}$ en primer lugar, y luego $\{i\}$ fue enlazado a $\{a, b, c, d, e\}$ por el patrón p_4 que también lo cubre, como se muestra en la figura 3.7.



(a) Dendrograma conceptual.



(b) Dendrograma tradicional.

Figura 3.6: Dendrogramas tradicional y conceptual usando la distancia de enlace simple d_L^s.

Notar que en el algoritmo de agrupamiento jerárquico tradicional cuya salida es mostrada en la figura 3.6(b), $\{i\}$ es el último grupo enlazado por la distancia.

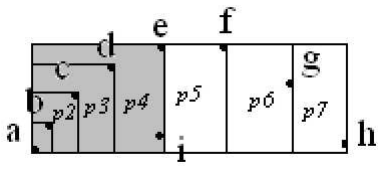


Figura 3.7: Los patrones $p_1, ..., p_7$ descubiertos por HDCC. El área sombreada muestra la evidencia cubierta por p_4.

4

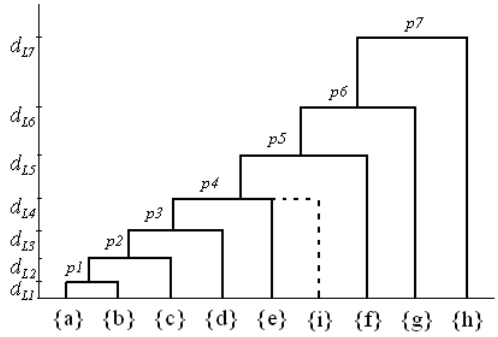
Consistencia entre distancias y operadores de generalización

La forma exacta del dendrograma conceptual, así como si contendrá líneas punteadas o no, dependerá no sólo de la distancia d y de los operadores de generalización Δ^* y Δ, sino que también dependerá de la distancia de enlace d_L usada. Esto es ilustrado por el ejemplo 5.

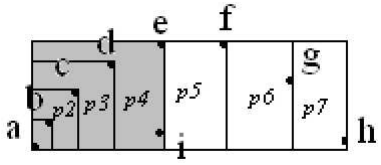
Ejemplo 5 *Consideremos nuevamente la evidencia del ejemplo 4 mostrada en la figura 3.5 y el lenguaje de patrones $\mathcal{L}$ definido como el conjunto de rectángulos de ejes paralelos. Supongamos, además, que los operadores de generalización $\Delta(e_1, e_2)$ y $\Delta^*(p_1, p_2)$ devuelven el mínimo rectángulo que incluye a los puntos e_1 y e_2 o a los rectángulos p_1 y p_2, respectivamente, y que la distancia d entre los ejemplos es la distancia Euclídea (dada en la ecuación 2.1).*

En la figura 4.1 podemos ver que los dendrogramas conceptuales que resultan de aplicar la distancia de enlace simple d_L^s (figura 4.1(a)) y completo d_L^c (figura 4.1(c)) para la misma evidencia, así como los respectivos patrones resultantes (figuras 4.1(b) y 4.1(d)) son muy diferentes.

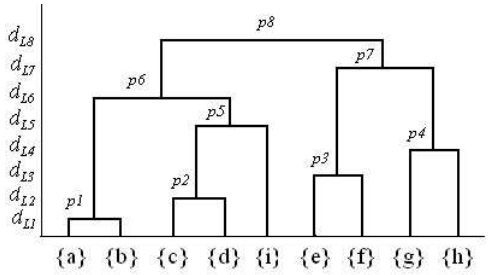
Podemos observar también que cuanto más similar es el dendrograma conceptual con respecto al dendrograma tradicional, más consistente es la distancia subyacente con respecto a los operadores de generalización empleados, ya que los grupos enlazados por la distancia coincidirán con aquellos enlazados por el patrón. Por ejemplo, al aplicar la distancia de enlace completo d_L^c en combinación con los operadores del ejemplo 5, el dendrograma conceptual de la figura 4.1(c) es equivalente a su correspondiente dendrograma tradicional.



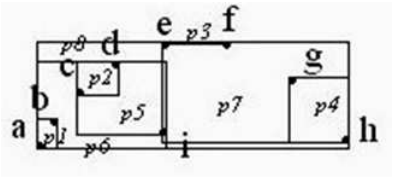
(a) Distancia de enlace simple d_L^s.



(b) Patrones bajo enlace simple.



(c) Distancia de enlace completo d_L^c .



(d) Patrones bajo enlace completo.

Figura 4.1: Diferencias entre dendrogramas usando diferentes distancias de enlace.

Teniendo en cuenta las consideraciones anteriores, hemos definido tres grados o niveles de consistencia entre las distancias y los operadores de generalización basados en la similitud entre el dendrograma conceptual y el tradicional. Cuanto más similares son los dendrogramas, mas consistente es la generalización con la distancia.

Estos tres niveles de consistencia van desde el más alto, que hemos llamado de *equivalencia*, pasando por el nivel intermedio de *preservación del orden* hasta el nivel mas bajo, que hemos llamado de *aceptabilidad*, y que son presentados en las secciones 4.1, 4.3 y 4.5, respectivamente.

El aporte mas importante de este capítulo son las propiedades de *acotabilidad fuerte*, *débil* y *aceptabilidad* que hemos definido en cada uno de los niveles de consistencia para los operadores de generalización Δ^* y Δ con respecto a las distancias d_L y d y que hemos probado como condiciones suficientes para obtener dendrogramas conceptuales equivalentes, que preservan el orden y aceptables, respectivamente. Ellas son presentadas en las secciones 4.2, 4.4 y 4.6.

4.1. Dendrogramas equivalentes

En algunos casos, el dendrograma conceptual es isomorfo al dendrograma tradicional. Esto ocurre cuando cada uno de los patrones descubiertos no cubre ningún otro grupo aparte de aquellos enlazados por la distancia; es decir, cada nuevo grupo, o bien, se forma por la unión de los grupos que se encuentra a la menor distancia de enlace l, o es un grupo unitario (una hoja del árbol). En consecuencia, decimos que un dendrograma conceptual es equivalente al correspondiente dendrograma tradicional si para cada grupo C que no es una hoja, todos sus hijos se encuentran enlazados a la misma distancia l. Este concepto es formalizado en la definición 5.

Definición 5 *Sea T el árbol resultante de HDCC. T es equivalente al dendrograma tradicional sii*

$\forall$ *nodo* $(C, p, l) \in T : |C| = 1 \vee (\forall\ (C_i, p_i, l_i)$ *hijo de* (C, p, l), $\exists (C_j, p_j, l_j) \in T$ *hijo de* (C, p, l) *tal que* $d_L(C_i, C_j, d) = l)$.

Nótese, sin embargo, que un dendrograma conceptual equivalente podría contener líneas punteadas. Esto es por la manera en que HDCC forma los grupos. En

cada iteración, HDCC une en primer lugar aquellos dos grupos que se encuentran más próximos de acuerdo a la distancia de enlace d_L. Sin embargo, podrían existir más de un par de grupos que se encuentren a la misma distancia de enlace. Dado que sólo dos de estos pares de grupos son enlazados por la distancia en cada pasada, cada uno de los otros grupos que se encuentre a la misma distancia de enlazado puede ser atraído por el patrón (y mostrado con líneas punteadas) antes de ser enlazado por la distancia en la siguiente iteración del algoritmo.

4.2. Operadores de generalización fuertemente acotados

La pregunta que nos planteamos acá es bajo qué condiciones podemos garantizar que, dados un par de operadores de generalización Δ^* y Δ, una distancia d entre los elementos del espacio métrico y una distancia de enlace d_L, HDCC producirá dendrogramas conceptuales equivalentes a la versión tradicional para cualquier evidencia E.

Si queremos obtener dendrogramas equivalentes, cada vez que HDCC determina los dos grupos más cercanos C_1 y C_2 a una distancia de enlace l, el correspondiente patrón p no debería cubrir ningún otro grupo C cuyas respectivas distancias l_1 y l_2 a C_1 y C_2 sean mayores que l. Notemos que tanto l_1 como l_2 deben ser menores o iguales que l. Si son menores, HDCC enlazaría antes C con C_1 o con C_2. Si son iguales, como dijimos antes, serán enlazadas por el patrón a la misma distancia l junto con C_1 y C_2.

Decimos que los operadores de generalización que generan patrones cuyas coberturas satisfacen esta condición son *fuertemente acotados por la distancia de enlace* d_L.

Intuitivamente, un operador binario de generalización de patrones es fuertemente acotado por la distancia de enlace d_L cuando para cualquier par de patrones p_1 y p_2, y para cualquier par de conjuntos C_1 y C_2 incluidos en la cobertura de p_1 y p_2 respectivamente, las respectivas distancias de enlace desde C_1 y C_2 a los nuevos elementos cubiertos por la generalización de p_1 y p_2 son menores o iguales a la distancia de enlace entre C_1 y C_2; es decir, los nuevos elementos cubiertos

por la generalización de los patrones p_1 y p_2 caen dentro de las bolas cerradas[1] de radio $d_L(C_1, C_2, d)$ y cuyos centros son los puntos de enlace de C_1 y de C_2. Notemos que los puntos de enlace, en el caso de la distancia de enlace simple d_L^s, corresponden a los dos elementos más cercanos entre C_1 y C_2; a los dos elementos más lejanos entre C_1 y C_2, en el caso de la distancia de enlace completo d_L^c y a los prototipos de C_1 y C_2, en el caso de la distancia de enlace a los prototipos d_L^p.

El concepto de operador binario de generalización de patrones fuertemente acotado es formalizado en la definición 6.

Definición 6 *Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones y d_L una distancia de enlazado.*

Un operador binario de generalización de patrones Δ^* es fuertemente acotado por la distancia de enlazado d_L sii
$\forall p_1,\ p_2 \in \mathcal{L},\ C_1 \subseteq Set(p_1),\ C_2 \subseteq Set(p_2),\ C \subseteq Set(\Delta^*(p_1, p_2)) - (Set(p_1) \cup Set(p_2)) : d_L(C, C_1, d) \leq d_L(C_1, C_2, d) \vee d_L(C, C_2, d) \leq d_L(C_1, C_2, d)$.

Ejemplo 6 *La figura 4.2 muestra los grupos $\{a, b, c\}$ y $\{d, e, f\}$ que se han formado dirigidos por la distancia de enlace simple d_L^s. Los patrones usados son uniones de rectángulos de ejes paralelos. El rectángulo A corresponde a la generalización de $\{a, b\}$, $A \cup B$ de $\{a, b, c\}$, C de $\{d, e\}$, $C \cup D$ de $\{d, e, f\}$ y $A \cup B \cup C \cup D \cup E$ de $\{a, b, c, d, e, f\}$.*

La unión de las dos bolas cerradas (círculos en el caso de $\mathbb{R}^2$) determina el área donde los nuevos elementos en la generalización de $A \cup B$ y $C \cup D$ deberían estar si el operador de generalización Δ^* fuese fuertemente acotado por la distancia de enlazado simple d_L^s.

Como podemos ver, este operador sí es fuertemente acotado por d_L^s ya que los nuevos elementos en la generalización de $A \cup B$ y $C \cup D$ son aquellos cubiertos por el mínimo rectángulo (en este ejemplo E) que une a los puntos más cercanos en los patrones $A \cup B$ y $C \cup D$. En el caso de la distancia d_L^s, los puntos de enlazado se encontrarán entre ellos cayendo la generalización dentro de la unión de las dos bolas con centro en los puntos de enlace y radio d_L^s.

Sin embargo, si cambiamos el operador de generalización Δ^* al de los rectángulos mínimos podremos ver en la figura 4.3 que este operador Δ^* no es fuertemente

[1]Sea (X, d) un espacio métrico. Una *bola cerrada* de radio $r > 0$ centrada en un punto x de X, usualmente denotada por $B(x; r)$, es definida como $\{y \in X \mid d(y, x) \leq r\}$.

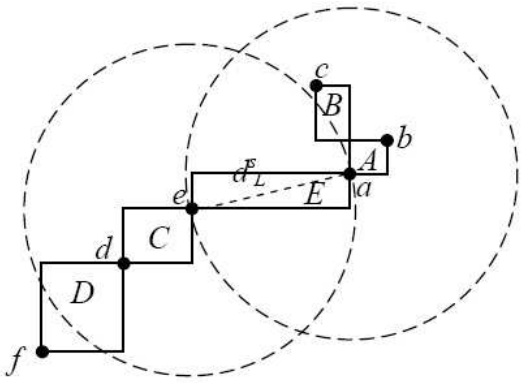


Figura 4.2: Región válida para el patrón producto de un operador Δ^* fuertemente acotado por d_L^s.

acotado por d_L^s ya que parte de los puntos cubiertos por el rectángulo C (que es la generalización de los rectángulos A y B) caen fuera del área determinada por la unión de los dos círculos con centro en los puntos de enlace a y e y radio $d(a, e)$.

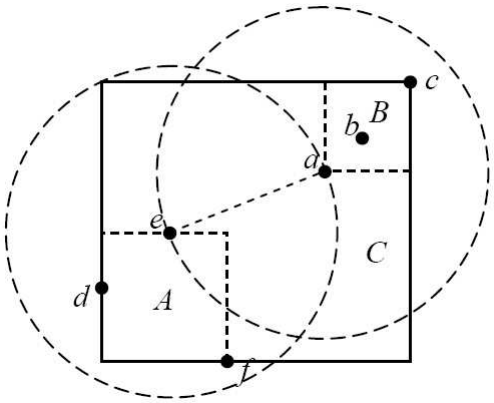


Figura 4.3: Patrón producto de un operador Δ^* que no es fuertemente acotado por la distancia de enlazado d_L^s.

La misma propiedad debe ser definida para los operadores binarios de generalización Δ los cuales son usados por HDCC para obtener las generalizaciones de los grupos unitarios.

Dado que la distancia de enlace d_L entre dos conjuntos unitarios se reduce a la distancia d entre los elementos de los dos conjuntos y puesto que los ope-

radores binarios de generalización Δ están definidos para pares de elementos en X, hablaremos de que un operador binario de generalización Δ es fuertemente acotado por la distancia d.

Por analogía con el razonamiento hecho para Δ^*, para que un operador Δ sea fuertemente acotado por d se debe cumplir que para cualquier par de elementos e_1 y e_2, las distancias desde e_1 y e_2 a cualquier otro elemento e cubierto por la generalización de e_1 y e_2 debe ser menor o igual a la distancia entre e_1 y e_2; es decir, e debe estar incluido dentro de las bolas cerradas $B(e_1; d(e_1, e_2))$ y $B(e_2; d(e_1, e_2))$. Esta idea es formalizada en la definición 7.

Definición 7 *Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones.*

Un operador de generalización binario 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 $d(e, e_1) \leq d(e_1, e_2) \vee d(e, e_2) \leq d(e_1, e_2)$.

Ejemplo 7 *La figura 4.4 muestra gráficamente el área en $\mathbb{R}^2$ donde estarán incluidos todos los elementos cubiertos por el patrón p resultante de la generalización de dos punto e_1 y e_1 computado por un operador binario de generalización Δ fuertemente acotado.*

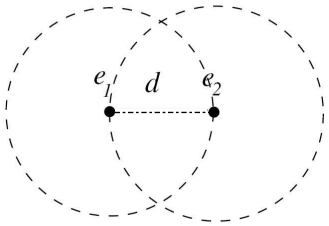


Figura 4.4: Cobertura máxima para un patrón computado por un operador binario de generalización Δ fuertemente acotado por la distancia d.

Ejemplo 8 *Las cuatro generalizaciones de dos puntos e_1 y e_2 en $(\mathbb{R}^2, d)$, con d la distancia Euclídea, mostradas en la figura 4.5 son producto de operadores de generalización fuertemente acotados por d, excepto la cuarta donde algunos de los puntos cubiertos por la generalización caen fuera de las bolas con centro en los puntos e_1 y e_2 y radio $d(e_1, e_2)$.*

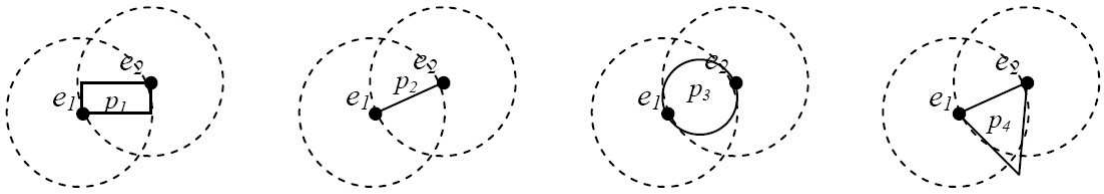


Figura 4.5: Ejemplos de patrones obtenidos por operadores Δ que son y no son fuertemente acotados por la distancia Euclídea en $\mathbb{R}^2$.

Hemos visto, a través del ejemplo 5, que la distancia de enlace d_L usada afecta la forma de los dendrogramas resultantes. Esto ocurre, en realidad, porque la distancia de enlazado d_L afecta la propiedad de acotabilidad de los operadores de generalización Δ^*. Es decir, un operador de generalización Δ^* puede ser fuertemente acotado por una distancia de enlace pero no por otra. Esto mismo ocurre con los operadores Δ y la distancia d.

Ejemplo 9 *Como vimos en la figura 4.3, el operador de generalización Δ^* de los rectángulos mínimos no es fuertemente acotado por la distancia de enlace simple d_L^s. Sin embargo, Δ^* es fuertemente acotado por d_L^c puesto que cada patrón describe un rectángulo que está determinado por los dos puntos más lejanos en p_1 y p_2, como puede verse en la figura 4.6, y los nuevos elementos en el rectángulo están contenidos en la unión de $B(e_1; d(e_1, e_2))$ y $B(e_2; d(e_1, e_2))$ donde e_1 y e_2 son los puntos de enlace (en este caso, los dos mas distantes).*

La proposición 1 establece que la acotabilidad fuerte de los operadores de generalización Δ^* y Δ es una condición suficiente para obtener dendrogramas equivalentes.

Proposición 1 *Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones para X, Δ un operador de generalización binario, Δ^* un operador de generalización binario de patrones y d_L una distancia de enlazado.*

Para toda evidencia $E \subseteq X$, el dendrograma conceptual T que resulta de HDCC(E, X, d, Δ^*, Δ, d_L) es equivalente al dendrograma tradicional si los operadores Δ^* y Δ son fuertemente acotados por d_L y d, respectivamente.

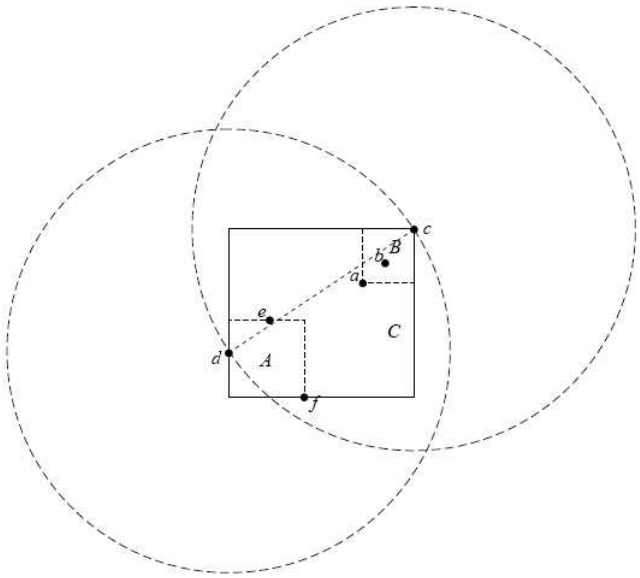


Figura 4.6: Patrón producto de un operador Δ^* que es fuertemente acotado por la distancia de enlazado d_L^c.

Demostración. Hay dos casos diferentes a considerar en T: (a) las hojas y (b) los nodos internos.

(a) En el primer paso, HDCC construye n grupos unitarios $\{e\}$ y sus correspondientes generalizaciones $\Delta(e, e)$. Siendo que Δ está fuertemente acotado por d, por Definición 7 tenemos que

$\forall e', e \in E$: si $e' \in Set(\Delta(e, e))$ entonces $d_L(e', e, d) \leq d_L(e, e, d)$.

Dado que $d_L(e, e, d) = 0$ y d_L es siempre positiva, entonces $d_L(e', e, d) = 0$. El único elemento e' que satisface esto es $e' = e$. En consecuencia, después que un patrón es calculado ningún otro elemento será agregado al grupo por HDCC.

(b) En los pasos siguientes, cada nodo nuevo (C, p, l) en T está formado por la unión de los dos grupos $(C_1, p_1, l_1), (C_2, p_2, l_2)$ cuya distancia de enlace l es la mínima, y donde p es calculado como $\Delta^*(p_1, p_2)$.

Siendo que Δ^* y Δ son operadores de generalización tenemos que $C \subseteq Set(p)$ y $C_1 \subseteq Set(p_1)$ y $C_2 \subseteq Set(p_2)$.

(b.1) Si $\Delta^*(p_1, p_2)$ no cubre ningún otro grupo además de C_1 y C_2, tenemos que C $= C_1 \cup C_2$ y $d_L(C_1, C_2, d) = l$.

(b.2) Supongamos que existe otro hijo (C_3, p_3, l_3) de (C, p, l) tal que $C_3 \subseteq Set(\Delta^*(p_1, p_2)) - (Set(p_1) \cup Set(p_2))$. Dado que Δ^* es fuertemente acotado por d_L, por Definición 6 tenemos $d_L(C_3, C_1, d) \leq d_L(C_1, C_2, d) \vee d_L(C_3, C_2, d) \leq d_L(C_1, C_2, d)$. Sin embargo, $d_L(C_3, C_1, d)$ debe ser igual a $d_L(C_1, C_2, d)$ y $d_L(C_3, C_2, d)$ también debe ser igual a $d_L(C_1, C_2, d)$, caso contrario HDCC habría fusionado antes C_1 con C_3 o C_2 con C_3 que C_1 y C_2.

En consecuencia, de (b.1) y (b.2), podemos concluir que $d_L(C_i, C_j, d) = l$ para cualquier hijo $(C_i, p_i, l_i), (C_j, p_j, l_j)$ de (C, p, l).

Finalmente de (a) y (b) concluimos que T es equivalente al dendrograma tradicional por definición 5. □

4.3. Dendrogramas que preservan el orden

La condición para tener dendrogramas equivalentes, sin embargo, es demasiado fuerte para muchos tipos de datos y operadores de generalización, puesto que fuerza a generalizaciones mínimas.

En efecto, un operador de generalización de patrones Δ^* cuya cobertura $Set(\Delta^*(p_1, p_2))$ es igual a $Set(p_1) \cup Set(p_2)$ para todo par de patrones p_1, p_2 en $\mathcal{L}$, es fuertemente acotado por d_L dado que no existe ningún nuevo elemento en $Set(\Delta^*(p_1, p_2)) - Set(p_1) \cup Set(p_2)$ que deba satisfacer la condición exigida en la definición 6 de operador fuertemente acotado, por lo que todos la cumplen. Análogamente, el operador Δ es fuertemente acotado por la distancia d cuando la generalización es mínima, es decir, cuando la cobertura de $\Delta(e_1, e_2)$ es igual a $\{e_1, e_2\}$ para todo par de elementos $e_1, e_2 \in X$.

Algunas veces, para un par de operadores de generalización Δ^* y Δ, una distancia entre los elementos del espacio métrico d y una distancia de enlazado d_L dada, el dendrograma conceptual, aunque no equivalente al dendrograma tradicional, puede preservar el orden en el cual los grupos son enlazados por la distancia.

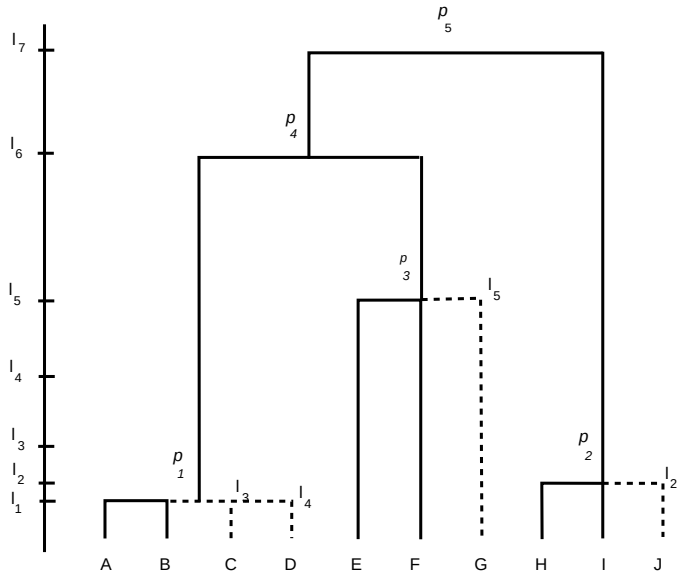
Que un dendrograma conceptual preserve el orden significa para nosotros que los elementos en los grupos del dendrograma conceptual no pueden estar mezclados con respecto al dendrograma tradicional. Es decir, si un grupo fue descubierto por HDCC entonces ese grupo debe existir también dentro de los grupos descubiertos por el algoritmo tradicional. Sin embargo, es importante notar que en el dendrograma tradicional pueden existir grupos que no sean descubiertos por HDCC. Esto es porque HDCC puede enlazar varios grupos en un mismo nivel de la jerarquía atraídos por el patrón, por lo que el número de grupos en la jerarquía conceptual va a ser menor o igual que el número de grupos en el dendrograma tradicional.

Si T_C y T_T son, respectivamente, los árboles resultantes de HDCC y de la versión tradicional del algoritmo, podemos decir, sin tener en cuenta las distancias de enlazado de cada uno de sus nodos, que T_C es una abstracción de T_T.

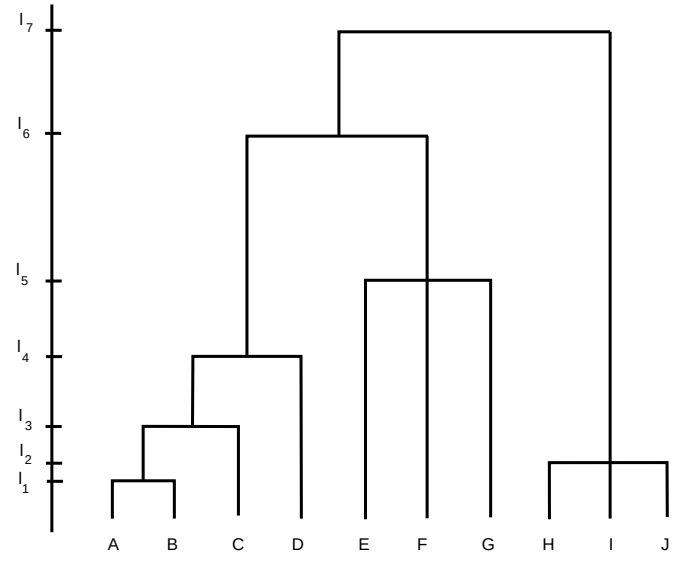
Ejemplo 10 *La figura 4.7 muestra un dendrograma conceptual que preserva el orden más la correspondiente versión tradicional del mismo.*

Podemos comprobar que los grupos formados por HDCC se encuentran incluidos entre los formados por la versión tradicional del algoritmo de agrupamiento jerárquico. Los grupos en la jerarquía conceptual son los mostrados en la figura 4.8(a), mientras que los formados por la versión tradicional del algoritmo son mostrados en la figura 4.8(b).

Notemos también en la figura 4.7(a) que si bien todos los grupos en el dendrograma conceptual se encuentran dentro de los grupos que el algoritmo tradicional formó en algún nivel de la jerarquía, no se forman necesariamente a la misma distancia. Por ejemplo, el grupo $\{A, B, C, D\}$ que antes se formaba a una distancia l_4, en el dendrograma conceptual se forma a una distancia $l_1 < l_4$. Esto se debe al hecho de que para que un dendrograma preserve el orden, un patrón generado por HDCC para un grupo C nunca deberá cubrir un grupo C' dejando sin cubrir a otro grupo C'' que se encuentre a una distancia de enlazado de C menor que la distancia de enlazado entre C y C'. Por ejemplo, el grupo $\{D\}$ que es cubierto por el patrón p_1 se enlaza con $\{A, B\}$ a distancia l_4 y el grupo $\{C\}$ que se enlaza a distancia $l_3 < l_4$ también es cubierto por el patrón, y a su vez los clusters $\{E, F, G\}$ y $\{H, I, J\}$ que no son cubiertos por p_1 se enlazan a distancias l_5 y l_6 que son mayores que l_4.

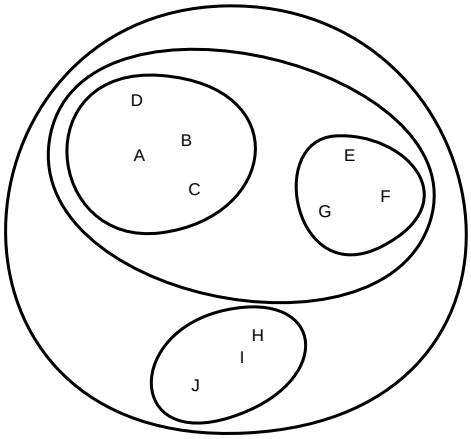


(a) Dendrograma conceptual.

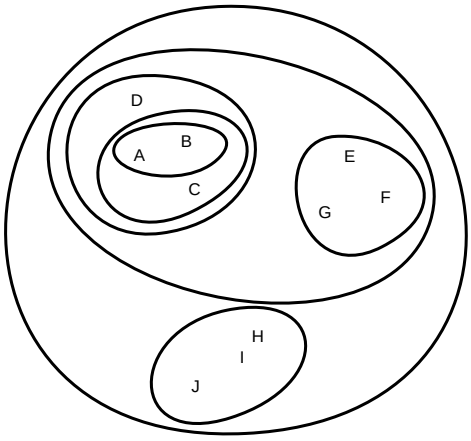


(b) Dendrograma tradicional.

Figura 4.7: Dendrograma conceptual que preserva el orden y su correspondiente dendrograma tradicional.



(a) Jerarquía de inclusiones en el dendrograma conceptual.



(b) Jerarquía de inclusiones en el dendrograma tradicional.

Figura 4.8: Jerarquías de inclusiones en los dos tipos de dendrogramas.

Mas específicamente, un dendrograma conceptual que preserva el orden es uno donde para todo nodo (C, p, l) en la jerarquía, sus hijos son enlazados a la misma distancia l o son enlazados por el patrón a una distancia de enlace menor de la que se enlazaría cualquier otro grupo en la evidencia no cubierto por el patrón.

Este concepto es formalizado por la definición 8.

Definición 8 *Sea T el árbol resultante 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), $\exists$ un hijo (C_j, p_j, l_j) de (C, p, l) tal que $d_L(C_i, C_j, d) = l \vee d_L(C_i, C_j, d) < d_L(C', C_i, d) \wedge d_L(C_i, C_j, d) < d_L(C', C_j, d)$, para todo $(C', p', l') \in T$ con $C' \not\subset Set(p)$.

4.4. Operadores de generalización débilmente acotados

Para obtener dendrogramas conceptuales que preserven el orden, cada vez que HDCC une los dos grupos mas cercanos C_1 y C_2 con patrones p_1 y p_2, cualquier otro grupo C cubierto por la generalización de p_1 y p_2 debe estar a una distancia de enlazado de C_1 y C_2 que sea menor a la distancia de enlazado desde C_1 y C_2 a cualquier otro grupo C' no cubierto por el patrón.

Esta idea es formalizada por la propiedad que llamamos *acotabilidad débil* y que es formalizada por la definición 9. Análogamente, la definición 10 establece la misma propiedad para operadores binarios de generalización Δ.

Definición 9 *Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones y d_L una distancia de enlace.*

Un operador binario de generalización de patrones Δ^* es débilmente acotado por d_L sii $\forall p_1, p_2 \in \mathcal{L}$, $C_1 \subseteq Set(p_1)$, $C_2 \subseteq Set(p_2)$, $C \subseteq Set(\Delta^*(p_1, p_2)) - (Set(p_1) \cup Set(p_2))$, $C' \not\subset Set(\Delta^*(p_1, p_2))$:

$d_L(C, C_1, d) \leq d_L(C_1, C_2, d) \vee d_L(C, C_2, d) \leq d_L(C_1, C_2, d) \vee (d_L(C, C_1, d) < d_L(C', C_1, d) \wedge d_L(C, C_2, d) < d_L(C', C_2, d))$.

Definición 10 *Sea (X, d) un espacio métrico y $\mathcal{L}$ un lenguaje de patrones.*

Un operador de generalización binario Δ es débilmente acotado por d sii $\forall e, e', e_1, e_2 \in X$, si $e \in Set(\Delta(e_1, e_2))$ y $e' \notin Set(\Delta(e_1, e_2))$ entonces $d(e, e_1) \leq d(e_1, e_2) \vee d(e, e_2) \leq d(e_1, e_2) \vee (d(e, e_1) < d(e', e_1) \wedge d(e, e_2) < d(e', e_2)$.

La proposición 2 surge de forma inmediata a partir de las definiciones de acotabilidad débil y fuerte. Todo operador de generalización que es fuertemente acotado por la distancia de enlace es también débilmente acotado.

Proposición 2 *Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones, d_L una distancia de enlazado, Δ un operador binario de generalización y Δ^* un operador binario de generalización de patrones.*

(i) *Si Δ^* es fuertemente acotado por d_L entonces Δ^* es débilmente acotado por d_L.*

(ii) *Si Δ es fuertemente acotado por d_L entonces Δ es débilmente acotado por d_L.*

Demostración. La parte (i) de la Proposición 2 se deduce inmediatamente de las definiciones 6 y 9 de operadores Δ^* fuerte y débilmente acotados. Cualquier operador de generalización de patrones Δ^* que es fuertemente acotado por la distancia de enlace d_L es, también, débilmente acotado dado que la Definición 9 relaja la condición de la Definición 6.
El mismo razonamiento puede aplicarse para la parte (ii) puesto que la Definición 10 relaja la condición de la Definición 7. □

Al igual que en el caso de la acotabilidad fuerte, queremos demostrar que la acotabilidad débil es una propiedad suficiente para preservar el orden en los dendrogramas conceptuales, lo que es probado en la demostración de la proposición 3.

Proposición 3 *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_L una distancia de enlazado.*

Para cualquier evidencia $E \subseteq X$, el dendrograma conceptual T que resulta de $HDCC(E, X, d, \Delta^*, \Delta, d_L)$, preserva el orden si los operadores de generalización Δ^* y Δ son débilmente acotados por d_L y d, respectivamente.

Demostración. Hay dos casos diferentes a considerar en T: (a) las hojas y (b) los nodos internos.

(a) En el primer paso, HDCC construye n nodos $(\{e\}, \Delta(e,e), l)$ con $l = 0$.

Si $\Delta(e,e)$ cubre cualquier otro elemento, éste es fusionado con $\{e\}$. Siendo que Δ es débilmente acotado por d, por Definición 10 tenemos

$\forall e, e', e_1 \in E$: si $e \in Set(\Delta(e_1, e_1))$ y $e' \notin Set(\Delta(e_1, e_1))$ entonces $d(e, e_1) \leq d(e_1, e_1) \vee (d(e, e_1) < d(e', e_1))$.

Dado que $d(e_1, e_1) = 0$ y d es positiva, tenemos que $\forall e, e', e_1 \in E$: si $e \in Set(\Delta(e_1, e_1))$ y $e' \notin Set(\Delta(e_1, e_1))$ entonces $d(e, e_1) = 0 \vee d(e, e_1) < d(e', e_1)$.

(b) En los pasos siguientes, cada nodo (C, p, l) en T se forma fusionando, en primer término, los dos grupos (C_1, p_1, l_1) y (C_2, p_2, l_2) cuya distancia de enlazado l es la mínima, y donde p es calculado como la generalización de los patrones p_1 y p_2, es decir como $\Delta^*(p_1, p_2)$. Dado que Δ^* y Δ son operadores de generalización tenemos que $C \subseteq Set(p)$ y $C_1 \subseteq Set(p_1)$ y $C_2 \subseteq Set(p_2)$.

(b.1) Si $\Delta^*(p_1, p_2)$ no cubre ningún otro grupo aparte de C_1 y C_2, tenemos que $C = C_1 \cup C_2$ y $d_L(C_1, C_2, d) = l$.

(b.2) Supongamos que existe otro hijo (C_3, p_3, l_3) de (C, p, l) tal que $C_3 \subseteq Set(\Delta^*(p_1, p_2)) - (Set(p_1) \cup Set(p_2))$. Dado que Δ^* es débilmente acotado por d_L, por Definición 9 tenemos

$$d_L(C_3, C_1, d) \leq d_L(C_1, C_2, d) \vee d_L(C_3, C_2, d) \leq d_L(C_1, C_2, d) \vee$$

$$(d_L(C_3, C_1, d) < d_L(C', C_1, d) \wedge d_L(C_3, C_2, d) < d_L(C', C_2, d))$$

para todo $C' \notin Set(\Delta^*(p_1, p_2))$.

Razonando como en la Proposición 1 tenemos $d_L(C_3, C_1, d) = d_L(C_3, C_2, d) = d_L(C_1, C_2, d) = l \vee (d_L(C_3, C_1, d) < d_L(C', C_1, d) \wedge d_L(C_3, C_2, d) < d_L(C', C_2, d))$.

Finalmente de (a) y (b) concluimos que T preserva el orden por la definición 8. □

El ejemplo 11 ilustra el caso de un dendrograma que no preserva el orden, mientras que el ejemplo 12 muestra el uso de operadores de generalización que son débilmente acotados y que en consecuencia conducen a dendrogramas que sí lo preservan.

Ejemplo 11 *El dendrograma conceptual resultante de usar los operadores de generalización del ejemplo mostrado en la figura 3.4(b) no preserva el orden bajo la distancia de enlazado simple* d_L^s.

Podemos observar que Δ^* *no es débilmente acotado por* d_L^s *ya que el patrón* $aa*$ *cubre al grupo* $\{aabbbbbb\}$ *y no cubre a* $\{abb\}$ *el cual está a una distancia de enlazado* d_L^s *de* $\{aa\}$ *y de* $\{aab\}$ *menor que* $d_L^s(\{aabbbbbb\}, \{aa, aab\})$.

Ejemplo 12 *Sea* (X, d) *un espacio métrico donde* X *viene dado por un conjunto finito de datos nominales pertenecientes a una taxonomía de animales. Así, por ejemplo, Vertebrate, Mammal, Dog, son elementos en* X.

La distancia d *entre los elementos de* X *se encuentra definida de manera similar a la distancia usada en [Estruch, 2008]. Se trata de una distancia inducida por una relación* R*, donde* R *es un orden parcial que se encuentra definido como* xRy *si "x es un y". En la figura 4.9 se muestra parte de la relación* R *representada como una jerarquía.*

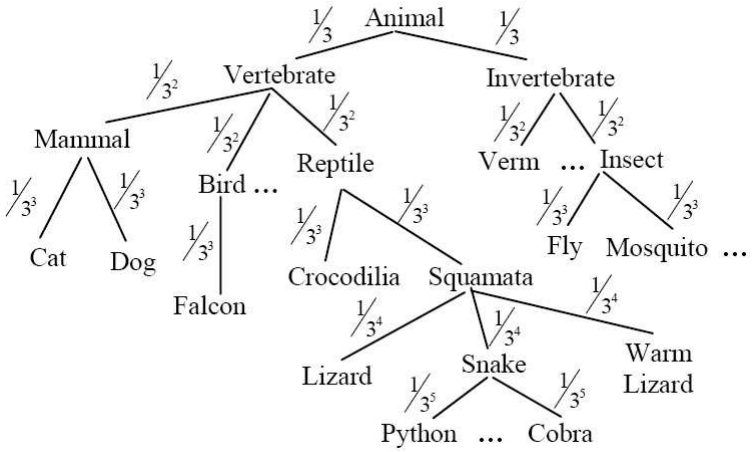


Figura 4.9: Relación R y costes asociados a cada arco, usados para el cálculo de la distancia d.

La distancia entre dos elementos está definida como la suma de los costes asociados a cada arco del camino más corto que conecta a los dos elementos. El coste de un arco de nivel i es $w_i = \frac{1}{3^i}$*, con* $i > 0$*. Así, por ejemplo, de acuerdo a la relación R de la figura 4.9, la distancia entre Cat y Cobra es* $\frac{1}{3^3} + \frac{1}{3^2} + \frac{1}{3^2} + \frac{1}{3^3} + \frac{1}{3^4} + \frac{1}{3^5}$*.*

La generalización $\Delta(e_1, e_2)$ *de dos elementos en* X*,* e_1 *y* e_2*, es definida como el mínimo ancestro de* e_1 *y* e_2*. Cuando uno de los elementos es un descendiente del otro, la generalización de ambos es el mas general de los dos elementos. Así, por ejemplo,* $\Delta(Python, Cobra) = Snake$ *y* $\Delta(Animal, Mosquito) = Animal$*.*

La generalización de patrones Δ^* *se define de forma análoga, ya que el lenguaje de patrones* $\mathcal{L} = X$*.*

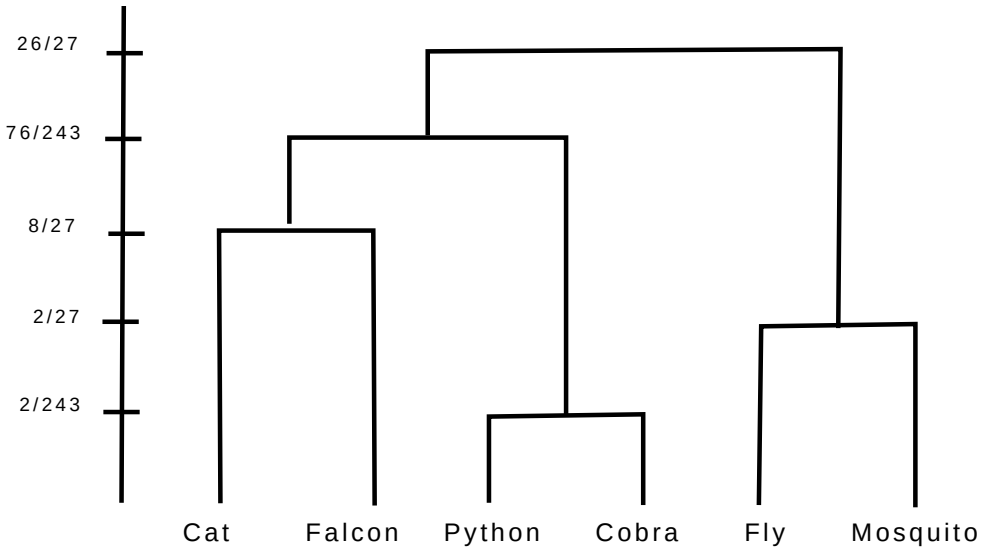
En las figuras 4.10(a) y 4.10(b) se muestran el dendrograma tradicional así como el correspondiente dendrograma conceptual para la evidencia $E = \{Cat, Falcon, Python, Cobra, Fly, Mosquito\}$*. Podemos ver que el dendrograma conceptual preserva el orden ya que el grupo* $\{Python, Cobra\}$ *que es enlazado por el patrón* $Vertebrate$ *descubierto a partir de la generalización de los patrones de los grupos* $\{Cat\}$ *y* $\{Falcon\}$ *se encuentra a una distancia de enlazado simple de* $\frac{76}{243}$ *que es menor que la distancia de enlazado simple de los grupos que no son cubiertos por el patrón (en este caso solamente* $\{Fly, Mosquito\}$*, que se enlaza a* $\frac{26}{27}$*).*

Sin embargo, esto no nos garantiza que todos los dendrogramas conceptuales resultantes del uso de esta distancia y estos operadores de generalización preserven el orden. Para ello debemos demostrar que ambos operadores de generalización son débilmente acotados por la distancia de enlace, en este caso d_L^s*.*

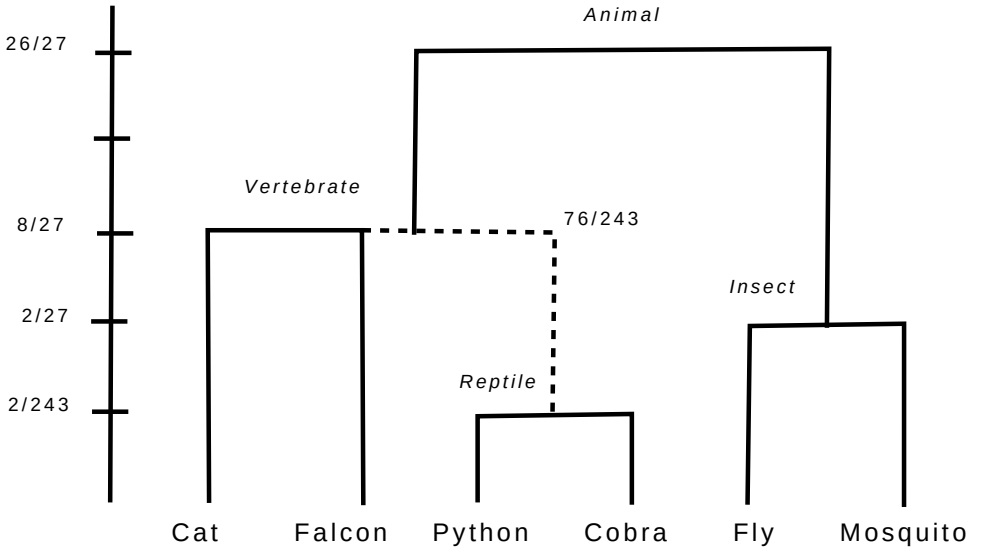
Notemos que en el peor caso todo elemento e cubierto por la generalización de e_1 *y* e_2*,* $\Delta(e_1, e_2)$*, se encuentra a una distancia máxima* $\frac{2}{3^k} + \frac{2}{3^{k+1}} + ... + \frac{2}{3^n}$ *de* e_1 *y de* e_2*, donde* $k > 0$ *es el nivel de* $\Delta(e_1, e_2)$ *y n el número de niveles en R (la raíz del árbol de encuentra en el nivel 1). Podemos notar también que, en este caso, todo elemento* e' *no cubierto por la generalización* $\Delta(e_1, e_2)$ *se encontrará a una distancia mínima de* $\frac{1}{3^{k-1}} + \frac{1}{3^k} + \frac{1}{3^{k+1}} + ... + \frac{1}{3^n}$ *de* e_1 *y* e_2*.*

Debemos probar que

$$\frac{1}{3^{k-1}} + \frac{1}{3^k} + \frac{1}{3^{k+1}} + ... + \frac{1}{3^n} > \frac{2}{3^k} + \frac{2}{3^{k+1}} + ... + \frac{2}{3^n} \iff$$



(a) Dendrograma Tradicional usando d_L^s.



(b) Dendrograma Conceptual no equivalente pero que preserva el orden usando d_L^s.

Figura 4.10: Dendrogramas conceptual que preserva el orden y tradicional para ejemplo de datos nominales jerárquicos.

$$\frac{1}{3^{k-1}} > \frac{1}{3^k} + \frac{1}{3^{k+1}} + ... + \frac{1}{3^n} \iff$$

$$\frac{1}{3^k} > \frac{1}{3^{k+1}} + ... + \frac{1}{3^{n+1}}$$

que se puede probar por inducción sobre n y k. Por lo que podemos concluir que Δ es débilmente acotado por d.

Razonando de forma análoga para Δ^*, ya que en este caso, Δ^* está definido igual que Δ, podemos concluir que el dendrograma resultante de aplicar los operadores Δ y Δ^* y las distancias d y d_L^s definidos en este ejemplo preserva el orden.

Finalmente, es interesante notar que el operador de generalización que cubre todo el espacio (el operador maximal $\Delta^*(p_1, p_2) = p$ donde $Set(p) = X$) es trivialmente débilmente acotado por cualquier distancia de enlace d_L dado que no existe ningún grupo C que satisfaga $C \nsubseteq Set(\Delta^*(p1, p2))$.

4.5. Dendrogramas aceptables

Existen algunos operadores de generalización que aunque no son (débilmente) acotados, conducen a la obtención de dendrogramas conceptuales que son consistentes con la distancia en un sentido más amplio.

La idea subyacente acá es que una generalización de un conjunto de elementos no debería cubrir nuevos elementos cuya distancia a los antiguos elementos sea mayor que la máxima distancia existente entre los antiguos elementos.

Hemos denominado a aquellos operadores que producen este tipo de patrones, así como a los dendrogramas que resultan de su aplicación, como *aceptables*.

Notemos que los dendrogramas aceptables pueden diferir considerablemente con respecto al correspondiente dendrograma tradicional.

4.6. Operadores de generalización aceptables

La definición 11 formaliza la idea de operador binario de generalización de patrones Δ^* aceptable.

Definición 11 *Sea (X, d) un espacio métrico, $\mathcal{L}$ un lenguaje de patrones y d_L^c la distancia de enlace completo. Un operador de generalización binario de patrones Δ^* es aceptable sii $\forall p_1, p_2 \in \mathcal{L}$, $e \in Set(\Delta^*(p_1, p_2))$, $\exists e' \in Set(p_1) \cup Set(p_2)$: $d(e, e') \leq d_L^c(Set(p_1), Set(p_2), d)$.*

Ejemplo 13 *La figura 4.11 muestra la región determinada por la unión de los círculos con centros en los puntos de $Set(p_1) \cup Set(p_2)$ y radios iguales a la distancia de enlace completo $d_L^c(Set(p_1), Set(p_2), d)$, o lo que es lo mismo, la máxima distancia entre los elementos en $Set(p_1)$ y los elementos en $Set(p_2)$, donde $p_1 = \Delta(a, b)$ es el mínimo rectángulo que contiene a los puntos a y b y $p_2 = \Delta(c, d)$ el mínimo rectángulo que contiene a los puntos c y d.*

La unión de dichos círculos determina la cobertura máxima para un patrón producido por un operador de generalización Δ^* para la evidencia $\{a, b, c, d\}$ en $\mathbb{R}^2$.

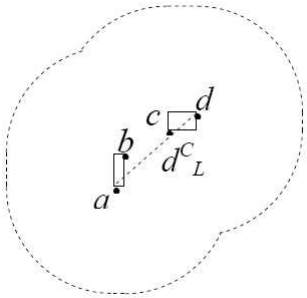


Figura 4.11: La cobertura máxima de un operador de generalización Δ^* para la evidencia $\{a, b, c, d\}$ en $\mathbb{R}^2$.

Podemos observar en la definición 11, que un operador binario de generalización de patrones es aceptable independientemente de la distancia de enlazado d_L usada para construir el dendrograma. La propiedad de aceptabilidad solamente depende de la distancia d entre los dos elementos más lejanos en los grupos. Usamos, sin embargo, $d_L^c(Set(p_1), Set(p_2), d)$ en la definición 11 para simplificar la notación ya que $d_L^c(Set(p_1), Set(p_2), d) = \text{máx}\{\, d(x, y) : x \in Set(p_1),\, y \in Set(p_2)\,\}$.

Trasladando el concepto de aceptabilidad a la generalización de dos elementos e_1 y e_2 en X tenemos que una generalización de un par de elementos para ser aceptable no debería cubrir nuevos elementos cuya distancia a los antiguos elementos sea mayor que la máxima distancia existente entre los antiguos elementos. Dado que los antiguos elementos son sólo e_1 y e_2, diremos simplemente que la generalización de e_1 y e_2 no debería cubrir elementos cuya distancia a e_1 y e_2 sea mayor que la distancia entre e_1 y e_2. Sin embargo, si observamos la definición 7, ésta es la misma condición que debe satisfacer un operador Δ para ser fuertemente acotado. Es decir, desde la perspectiva de aceptabilidad sólo los operadores Δ fuertemente acotados son aceptables.

En consecuencia, definimos como aceptables aquellos dendrogramas que resultan del uso de un operador Δ fuertemente acotado por d y un operador Δ^* aceptable.

El siguiente ejemplo ilustra la propiedad de aceptabilidad del operador de generalización Δ^* para una distancia de átomos, siendo Δ^* definido como el operador de generalización de átomos lgg.

Ejemplo 14 *Sea (X, d) un espacio métrico, donde X es el conjunto de átomos básicos de un lenguaje de primer order L y d es la distancia para átomos definida en [Cheng, 1997] y dada por la ecuación 2.5 de la sección 2.1.*

Como operador de generalización Δ^* usamos el operador lgg de Plotkin que devuelve la generalización menos general de dos átomos [Plotkin, 1970] siendo el lenguaje de patrones $\mathcal{L}$ el conjunto de átomos en el lenguaje de primer orden L.

Aunque, como veremos a continuación, Δ^* no es fuerte ni débilmente acotado por d_L^s ni por d_L^c, sí es un operador de generalización aceptable.

Sean

$$p_1 = q(a, X, c) \text{ un patrón en } \mathcal{L} \text{ y } C_1 = \{q(a, b, c), q(a, c, c)\} \subset Set(p_1);$$
$$p_2 = q(a, b, b) \text{ un patrón en } \mathcal{L} \text{ y } C_2 = \{q(a, b, b)\} \subset Set(p_2).$$

$\Delta^*(p_1, p_2) = lgg(q(a, X, c), q(a, b, b)) = q(a, X, Y)$ cubre, entre otros, al grupo $C = \{q(a, a, a)\}$. Sin embargo,

$$d_L^s(C, C_1, d) = 1/3 > d_L^s(C_1, C_2, d) = 1/6 \ y$$
$$d_L^s(C, C_2, d) = 1/3 > d_L^s(C_1, C_2, d) = 1/6 \ y$$

Por lo que, de acuerdo a la definición 6, Δ^* *no es fuertemente acotado por* d_L^s.

Supongamos ahora que:

$$p_1 = q(a, X, b, c) \text{ y } p_2 = q(a, b, X, Y)$$

Existen

$$C_1 = \{q(a, d, b, c), q(a, c, b, c)\} \subset Set(p_1) \text{ y}$$
$$C_2 = \{q(a, b, d, c), q(a, b, b, a)\} \subset Set(p_2).$$

tales que $\Delta^*(p_1, p_2) = lgg(q(a, X, b, c), q(a, b, X, Y)) = q(a, X, Y, Z)$.
Sin embargo, $q(a, X, Y, Z)$ *cubre el grupo* $C = \{q(a, a, a, b)\}$, *pero*

$$d_L^c(C, C_1, d) = 3/8 > d_L^c(C_1, C_2, d) = 1/4 \text{ y}$$
$$d_L^c(C, C_2, d) = 3/8 > d_L^c(C_1, C_2, d) = 1/4.$$

Por lo que Δ^* *tampoco es fuertemente acotado por* d_L^c.

Aunque, como vimos, Δ^* *no es fuertemente acotado por* d_L^c *y* d_L^s, *veamos que sí es aceptable.*

Sean p_1 *y* p_2 *dos patrones con aridad* n *y* $\{j_1, ..., j_k\}$ $(k \leq n)$ *las posiciones de los términos en* p_1 *y* p_2 *que no son al mismo tiempo iguales y básicos, es decir, las posiciones en donde ambos patrones no tienen términos básicos iguales.*

Notar, en consecuencia, que la distancia entre cualquier elemento $e_1 = q(t_1, ..., t_n)$ *cubierto por* p_1 *y cualquier elemento* $e_2 = q(s_1, ..., s_n)$ *cubierto por* p_2 *es menor o igual a* $\frac{k}{2n}$ *y que el valor máximo* $\frac{k}{2n}$ *es alcanzado cuando todos los términos en las posiciones* $\{j_1, ..., j_k\}$ *no unifican, haciendo que las distancias* $d(t_{j_i}, s_{j_i})$ *sean igual a 1.*

Notemos también que cualquier elemento $e \in Set(p)$ *puede estar a lo sumo a una distancia* $\frac{k}{2n}$ *de cualquier otro elemento* $e' \in Set(p)$ *ya que pueden diferir solamente en aquellos términos que no son básicos en* p. *Esto se cumple en particular para* e_1 *y* e_2 *dado que* $Set(p_1) \subseteq Set(p)$ *y* $Set(p_2) \subseteq Set(p)$.

Para concluir, vale destacar que un aspecto positivo de $\Delta(e, e) = \{e\}$, usado por el algoritmo HDCC, es que es fuertemente acotado (y aceptable). Este operador puede usualmente ser expresado en la mayoría de los lenguajes de patrones $\mathcal{L}$. Por lo tanto, solamente Δ^* debe ser analizado en la mayor parte de los casos.

Adicionalmente, la aceptabilidad de Δ^* es independiente de la distancia de enlace d_L haciendo que los resultados de aceptabilidad obtenidos sean en consecuencia extensibles a cualquier distancia de enlace.

5

Instanciación para clustering proposicional

La propuesta al agrupamiento conceptual jerárquico basado en distancias presentada en esta tesis es una aproximación general que hace posible combinar diferentes distancias y operadores de generalización para obtener conceptos que son cruciales para tareas descriptivas de minería de datos.

En este capítulo presentamos una instanciación para agrupamiento proposicional en donde los datos son expresados en término de atributos e instancias. Analizamos, en el contexto de nuestra aproximación, los distintos grados de consistencia definidos entre distancias y operadores de generalización para datos numéricos, nominales y tuplas, los cuales son los tipos de datos típicamente usados en el aprendizaje proposicional. Para ello, en primer lugar, proponemos un conjunto de operadores y distancias para los datos antes mencionados, y analizamos las propiedades que éstos satisfacen sobre la base de los tres niveles de similitud definidos entre la jerarquía obtenida por la distancia de enlazado y la jerarquía que resulta de usar los operadores de generalización, y que fueron presentados en el capítulo 4.

Demostramos que los intervalos junto con la diferencia absoluta usada como distancia para los números reales y la unión de conjuntos con la distancia discreta para los datos nominales son pares de operadores de generalización y distancias altamente consistentes, consiguiéndose el máximo grado de concordancia entre los dendrogramas tradicional y conceptual. Analizamos su uso en tuplas de nominales y reales, en donde encontramos el mismo resultado para el caso de la distancia de enlace completo, completando así la instanciación de nuestro

marco para agrupamiento proposicional.

Adicionalmente, encontramos un resultado de composabilidad para las tuplas que es obtenido independientemente de los tipos base. Esta propiedad de composabilidad permite que nuestro marco sea extendido en forma directa a tuplas de cualquier tipo de dato complejo siempre que los operadores de generalización asociados a los tipos de datos de las componentes satisfagan la propiedad pedida para las tuplas. Por ejemplo, podemos aseverar una cierta propiedad para tuplas de grafos, strings y números si conocemos a priori que las propiedades para los tipos de datos subyacentes se satisfacen.

El capítulo se encuentra organizado de la siguiente manera. En la sección 5.1 proponemos pares de operadores de generalización y distancias para datos nominales y demostramos las propiedades que son satisfechas en ese contexto. En las sección 5.2 y 5.3 hacemos lo mismo para datos numéricos y tuplas, respectivamente.

5.1. Datos nominales

Un tipo de dato nominal, también referido como enumerado o categórico, denota un conjunto finito de valores posibles que un atributo puede tomar; por ejemplo, el género de una persona, los días de la semana, los colores, etc. El tipo de dato Boolean es un caso especial de tipo de dato nominal, donde existen sólo dos valores posibles.

El espacio métrico para un tipo de datos nominal estará compuesto por un conjunto X, que es simplemente un conjunto finito de valores simbólicos, y una distancia d entre elementos de X.

Existen muchas distancias d para datos nominales. Algunas de las que más comúnmente se usan son la distancia discreta, que retorna 0 cuando ambos valores son los mismos y 1 en otro caso, y la distancia VDM (Value Difference Metric) [Stanfill and Waltz, 1986], entre otras.

En algunos casos, resulta de utilidad el uso de una distancia definida por el usuario. En el ejemplo 15 se muestra una de estas distancias.

Ejemplo 15 *Supongamos el espacio métrico* (X, d)*, donde*

$X = \{XXL, XL, L, M, S, XS, XXS\}$ y la distancia d se encuentra definida de la siguiente manera:

$$\begin{array}{rclrclrcl} d(XXL, XL) & = & 1 & d(XL, M) & = & 2 & d(L, XXS) & = & 4 \\ d(XXL, L) & = & 2 & d(XL, S) & = & 3 & d(M, S) & = & 1 \\ d(XXL, M) & = & 3 & d(XL, XS) & = & 4 & d(M, XS) & = & 2 \\ d(XXL, S) & = & 4 & d(XL, XXS) & = & 5 & d(M, XXS) & = & 3 \\ d(XXL, XS) & = & 5 & d(L, M) & = & 1 & d(S, XS) & = & 1 \\ d(XXL, XXS) & = & 6 & d(L, S) & = & 2 & d(S, XXS) & = & 2 \\ d(XL, L) & = & 1 & d(L, XS) & = & 3 & d(XS, XXS) & = & 1 \end{array}$$

Esta distancia organiza los elementos en una línea donde XXL y XXS son los puntos extremos.

Típicamente, los patrones para datos nominales son expresados como condiciones sobre los valores de los atributos; por ejemplo, $talla = XL$ o $talla \neq XL$.

Notemos que, dado que X es finito, las coberturas de los posibles patrones también son finitas. En consecuencia, ellas pueden ser expresadas extensionalmente como subconjuntos de X. Por lo tanto, en este caso, tenemos que el lenguaje de patrones $\mathcal{L}$ para los datos nominales se reduce al conjunto 2^X.

Teniendo en cuenta que los patrones son subconjuntos de X, proponemos que el operador binario de generalización para un par de valores nominales nos devuelva el conjunto que contiene ambos valores.

Proposición 4 *Sea (X, d) un espacio métrico, X un conjunto de datos nominales y 2^X el lenguaje de patrones.*

La función $\Delta_{nom} : X \times X \to 2^X$ definida por $\Delta_{nom}(e_1, e_2) = \{e_1, e_2\}$ es un operador binario de generalización para datos nominales.

Demostración. Es trivial ver que Δ_{nom} satisface la condición para ser un operador binario de generalización de acuerdo a la definición 3 ya que ambos elementos (e_1 y e_2) se encuentran en su generalización. □

Nuevamente, teniendo en cuenta que los patrones son subconjuntos de X, proponemos la unión de conjuntos como la generalización de dos patrones.

Proposición 5 *Sea (X, d) un espacio métrico, X un conjunto de datos nominales y 2^X el lenguaje de patrones.*

La función $\Delta^*_{nom} : 2^X \times 2^X \to 2^X$ definida por $\Delta^*_{nom}(s_1, s_2) = s_1 \cup s_2$ es un operador binario de generalización de patrones con respecto a 2^X.

Demostración. Al igual que antes, es trivial ver que Δ^*_{nom} satisface la condición para ser un operador binario de generalización de patrones de acuerdo a la definición 4 ya que ambos patrones (conjuntos s_1 y s_2) se encuentran incluidos en su generalización. □

La proposición 6 establece las propiedades que son satisfechas por los operadores Δ_{nom} y Δ^*_{nom} con respecto a las distancias d y d_L.

Proposición 6 *Sean Δ_{nom} y Δ^*_{nom} los operadores de generalización propuestos en las proposiciones 4 y 5 para datos nominales, d una distancia entre datos nominales y d_L una distancia de enlazado.*

(i) *Δ_{nom} y Δ^*_{nom} son fuertemente acotados por d y d_L, respectivamente.*

(ii) *Δ_{nom} y Δ^*_{nom} son débilmente acotados por d y d_L, respectivamente.*

(iii) *Δ^*_{nom} es aceptable.*

Demostración.

(i) Ambos operadores son fuertemente acotados independientemente de la distancia de enlazado empleada dado que su cobertura es la mínima posible, es decir, la cobertura de un patrón es igual a los datos que generaliza:

Dado que no existe un nuevo elemento cubierto por la generalización de los patrones p_1 y p_2 que no esté cubierto por p_1 o p_2, Δ^*_{nom} está fuertemente acotado por Definición 7.

Δ_{nom} es fuertemente acotado, también, ya que $\Delta(e_1, e_2) = \{e_1, e_2\}$ y, en consecuencia, para cualquier $e \in \{e_1, e_2\}$ podemos ver que, trivialmente, la Definición 7 se cumple.

(ii) Dado que, por el apartado (i) de la proposición 6, Δ_{nom} y Δ^*_{nom} son fuertemente acotados por d y d_L respectivamente, en consecuencia por la proposición 2 son débilmente acotados.

(iii) Δ^*_{nom} es aceptable ya que cualquier elemento cubierto por la generalización de p_1 y p_2 (en este caso $\Delta^*_{nom}(p_1, p_2) = p_1 \cup p_2$) es también cubierto por p_1 o por p_2 y, por lo tanto, se encuentra siempre a una distancia menor o igual a la máxima distancia entre los elementos de $Set(p_1) \cup Set(p_2) = p_1 \cup p_2$.

□

Ejemplo 16 *La figura 5.1 ilustra el uso de HDCC para la evidencia* $E = \{XXS, S, M, L\}$ *con la distancia discreta y los operadores de generalización* Δ^*_{nom} *y* Δ_{nom} *propuestos anteriormente.*

El dendrograma conceptual resultante es el mismo para cualquiera de las distancias de enlazado estudiadas ya que d_L *depende de* d*, y* d *es una función particular ya que todos los elementos se encuentran a distancia 1 de cualquier otro excepto de sí mismo, por lo que la máxima distancia entre dos elementos es la misma que la mínima y que la media.*

Además, dado que Δ^*_{nom} *es fuertemente acotado por* d_L*, el dendrograma conceptual es, en consecuencia, equivalente al correspondiente dendrograma tradicional.*

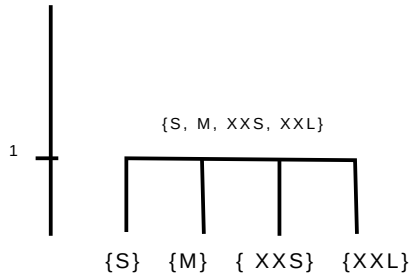
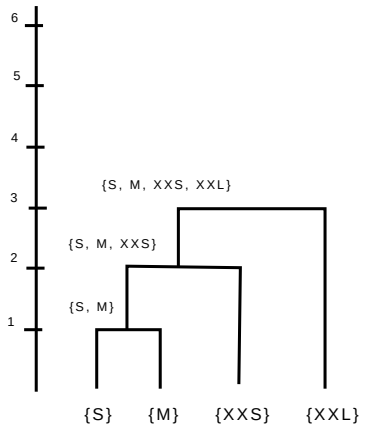
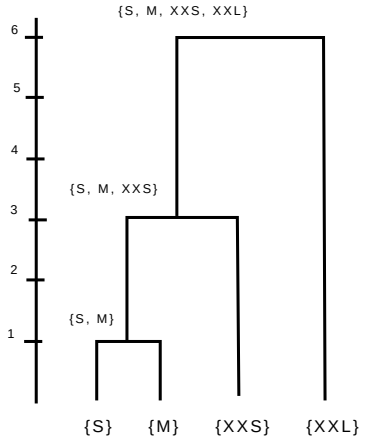


Figura 5.1: Una instanciación de HDCC para datos nominales, con los operadores Δ^*_{nom} y Δ_{nom}, y la distancia discreta.

Notemos que si cambiamos a la distancia definida por el usuario dada en el ejemplo 15, el dendrograma conceptual cambia al mostrado en la figura 5.2(a) para el caso de la distancia de enlace simple y al de la figura 5.2(b) en el caso de la distancia de enlace completo. Sin embargo, en ambos casos podemos afirmar por las proposiciones 1 y 6(i) que los dendrogramas conceptuales son equivalentes a los correspondientes dendrogramas resultantes de la versión tradicional del algoritmo.



(a) Enlace simple d_L^s.



(b) Enlace completo d_L^c.

Figura 5.2: Instanciacines de HDCC para datos nominales, con los operadores Δ_{nom}^* y Δ_{nom}, la distancia definida por el usuario, usando (a) d_L^s y (b) d_L^c.

5.2. Datos numéricos

Los datos numéricos son ampliamente empleados para expresar cantidades y medidas y muchos otros atributos de objetos del mundo real.

Un espacio métrico bien conocido para los datos numéricos es $(\mathbb{R}, d)$ donde d es la distancia definida como la diferencia absoluta de dos números reales, es decir $d(e_i, e_j) = |e_i - e_j|$.

Una generalización usual para un conjunto de números es el mínimo intervalo cuyos extremos son los valores mínimos y máximos en el conjunto. En este caso el lenguaje de patrones $\mathcal{L}$ corresponde al conjunto de los intervalos finitos cerrados en $\mathbb{R}$. Usando $\mathcal{L}$ como lenguaje de patrones, podemos definir la generalización de dos elementos en $\mathbb{R}$ como el mínimo intervalo que incluye a ambos elementos:

Proposición 7 *Sea $\mathcal{L}$ el conjunto de los intervalos finitos cerrados en $\mathbb{R}$.*

$\forall e_1, e_2 \in \mathbb{R}$ tal que $e_1 \leq e_2$, la función $\Delta_{num} : \mathbb{R} \times \mathbb{R} \to \mathcal{L}$ definida por $\Delta_{num}(e_1, e_2) = [e_1, e_2]$ es un operador binario de generalización para números reales.

Demostración. Es trivial ver que Δ_{num} satisface la condición para ser un operador binario de generalización de acuerdo a la definición 3 ya que ambos elementos (e_1 y e_2) pertenecen a su generalización (el mínimo intervalo que los contiene). □

De igual manera, extendemos la idea en la proposición 8 para operadores de generalización de patrones, en donde proponemos como la generalización de dos intervalos el mínimo intervalo que cubre a ambos.

Proposición 8 *Sea $\mathcal{L}$ el conjunto de los intervalos finitos cerrados en $\mathbb{R}$.*

La función $\Delta^*_{num} : \mathcal{L} \times \mathcal{L} \to \mathcal{L}$ definida por $\Delta^*_{num}([e_{i1}, e_{f1}], [e_{i2}, e_{f2}]) = [e_i, e_f]$, donde e_i es el menor valor en $\{e_{i1}, e_{i2}\}$ y e_f es el mayor valor en $\{e_{f1}, e_{f2}\}$, es un operador binario de generalización de patrones.

Demostración. Es trivial ver que Δ^*_{num} satisface la condición para ser un operador binario de generalización de patrones de acuerdo a la definición 4 ya que ambos intervalos se encuentran completamente incluidos en su generalización. □

En la proposición 9 proponemos y demostramos las propiedades que son satisfechas por Δ_{num} y Δ^*_{num}.

Proposición 9 *Sean Δ_{num} y Δ^*_{num} los operadores de generalización propuestos en las Proposiciones 7 y 8, d la distancia de la diferencia absoluta y d_L una distancia de enlazado.*

(i) *Δ_{num} y Δ^*_{num} son fuertemente acotados por d y d_L, respectivamente.*

(ii) *Δ_{num} y Δ^*_{num} son débilmente acotados por d y d_L, respectivamente.*

(iii) *Δ^*_{num} es aceptable.*

Demostración.

(i) La generalización de dos intervalos p_1 y p_2 es un nuevo intervalo p que sólo cubre los elementos cubiertos por p_1 y p_2 más aquellos elementos que se encuentran entre los valores extremos de ambos intervalos.

Si e_1 y e_2 son los puntos de enlace en C_1 y C_2 que determinan la distancia de enlazado $l = d_L(C_1, C_2, d)$, para ser fuertemente acotado por la distancia de enlace los nuevos elementos en el intervalo p deben estar incluidos en la bola $B(e_1, l)$ o $B(e_2, l)$, es decir, pertenece al intervalo $[e_1 - l, e_1 + l]$ o $[e_2 - l, e_2 + l]$. En el caso del enlazado simple, e_1 y e_2 corresponden a los dos elementos más cercanos en C_1 y C_2, en consecuencia $p_1 = [a, e_1]$ y $p_2 = [e_2, b]$ y $p = [a, b]$. Claramente los nuevos elementos en p, es decir los elementos en el intervalo $]e_1, e_2[$, están incluidos en $[e_1 - l, e_1 + l]$ y también en $[e_2 - l, e_2 + l]$ dado que $l = |e_2 - e_1|$.

El enlazado simple es el caso peor. Para las otras distancias de enlazado tenemos que e_1 y e_2 estarán más distantes y en consecuencia el radio de las bolas será mayor. Esto también es verdad para la distancia de enlazado a la media ya que podemos pensar en e_1 y e_2 como dos puntos en p_1 y p_2 cuya distancia l será siempre mayor que la distancia de enlace simple.

Por otra parte, es trivial ver que Δ_{num} es fuertemente acotado por d ya que para cualquier par de elementos e_1, e_2 y cualquier elemento e en $[e_1, e_2]$ cumple que $d(e, e_1) \leq d(e_1, e_2) \wedge d(e, e_2) \leq d(e_1, e_2)$.

(ii) Dado que Δ_{num} y Δ^*_{num} son fuertemente acotados, por las partes (i) y (ii) de la proposición 2, tenemos que Δ^*_{num} y Δ_{num} son débilmente acotados.

(iii) Para demostrar que Δ^*_{num} es aceptable debemos demostrar que para todo $p_1, p_2 \in \mathcal{L}, e \in Set(\Delta^*_{num}(p_1, p_2))$ existe $e' \in Set(p_1) \cup Set(p_2)$ tal que $d(e, e') \leq d^c_L(Set(p_1), Set(p_2), d)$.

Supongamos que $p_1 = [a, b]$ y $p_2 = [c, d]$.

$\Delta^*_{num}(p_1, p_2) = [min(a, c), max(b, d)] = p$.

Δ^*_{num} es aceptable ya que para cualquier elemento e en p siempre es posible encontrar un elemento e' en $Set(p_1) \cup Set(p_2)$, es decir en $[a, b]$ o $[c, d]$ tal que $d(e, e')$ sea menor o igual que la máxima distancia entre los elementos de $[a, b]$ y $[c, d]$, es decir $|min(a, c) - max(b, d)|$.

□

Ejemplo 17 *La figura 5.3 muestra un ejemplo muy sencillo de una aplicación de HDCC bajo enlace simple usando los operadores Δ^*_{num} y Δ_{num} propuestos en las proposiciones 7 y 8 para números reales.*

Por las proposiciones 1 y 9 parte (i), podemos asegurar que el dendrograma conceptual resultante es equivalente al dendrograma tradicional.

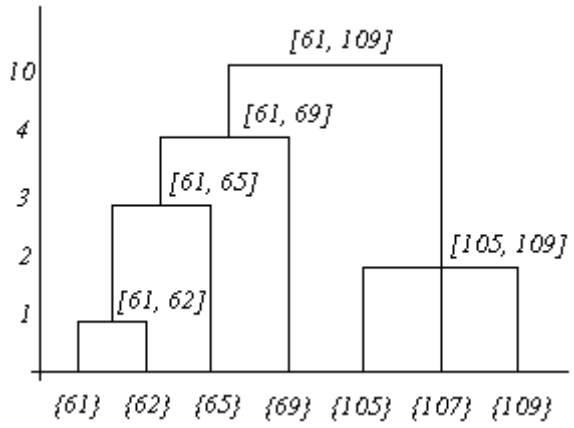


Figura 5.3: Una instanciación de HDCC para datos numéricos bajo enlace simple, con los operadores Δ^*_{num} y Δ_{num}, y la distancia de la diferencia absoluta.

5.3. Tuplas

Una tupla es una estructura ampliamente usada en el aprendizaje proposicional para la representación de datos, ya que los ejemplos pueden ser representados en forma de tuplas de datos numéricos y nominales.

Para definir un operador de generalización de tuplas, a diferencia de los dos tipos de datos previos, en este caso nos basamos en las propiedades de los tipos básicos a partir de los cuales el tipo tupla es construido. Asumimos que estos tipos básicos se encuentran inmersos en espacios métricos y que, por lo tanto, podemos usar las distancias definidas sobre cada uno de ellos para definir distancias entre tuplas.

Análogamente, para definir el lenguaje de patrones para tuplas $\mathcal{L}$, usamos también los lenguajes de patrones $\mathcal{L}_i$ definidos para cada uno de los espacios.

Sea (X_i, d_i) una colección de espacios métricos, donde $d_i(.,.)$ son funciones de distancia definidas sobre los espacios $X_i (i = 1, ..., n)$, y $\mathcal{L}_i$ $(i = 1, ..., n)$ una colección de lenguajes de patrones que corresponden a cada una de las n dimensiones de una tupla.

Denotamos (X, d) al espacio métrico de tuplas donde $X = X_1 \times X_2 \times ... \times X_n$. Por lo tanto, si $x \in X$ entonces x es una n-tupla $(x_1, ..., x_n)$, donde $x_i \in X_i$. En cuanto a la distancia d definida en X se han propuesto diversas funciones como por ejemplo las mostradas en la tabla 5.1. En lo que sigue de esta sección, denotaremos como d_T cualquiera de estas distancias para tuplas.

Puesto que definimos el lenguaje de patrones $\mathcal{L}$ para tuplas como $\mathcal{L} = (\mathcal{L}_1, ..., \mathcal{L}_n)$, usando los lenguajes de patrones básicos $\mathcal{L}_i$, la generalización Δ de dos tuplas x e y puede ser definida como la tupla cuyas componentes son las generalizaciones de las respectivas componentes en x e y. Este concepto es formalizado por la proposición 10, luego de que en la definición 12 hayamos formalizado el concepto de cobertura de un patrón en $\mathcal{L}$.

Definición 12 *Sea $X = X_1 \times ... \times X_n$ el espacio de tuplas, $\mathcal{L} = \mathcal{L}_1 \times \mathcal{L}_2 \times ... \times \mathcal{L}_n$ $(i = 1, ..., n)$ el lenguaje de patrones de tuplas definido sobre los lenguajes de patrones $\mathcal{L}_i$ y $p = (p_1, ..., p_n)$ un patrón en $\mathcal{L}$.*

La cobertura $Set(p)$ del patrón p se define como

$$\{(x_1, ..., x_n) \in X | x_i \in Set(p_i), i = 1, ..., n\}.$$

Tabla 5.1: Algunas distancias para tuplas.

Distancia de Manhattan	$d(x,y) = \sum_{i=1}^{n} d_i(x_i, y_i)$
Distancia de Manhattan pesada	$d(x,y) = \sum_{i=1}^{n} \alpha_i d_i(x_i, y_i)$
Distancia Euclídea	$d(x,y) = \sqrt{\sum_{i=1}^{n} d_i(x_i, y_i)^2}$
Distancia Euclídea pesada	$d(x,y) = \sqrt{\sum_{i=1}^{n} \alpha_i d_i(x_i, y_i)^2}$
Distancia de Chebyshev	$d(x,y) = max_{1 \leq i \leq n} d_i(x_i, y_i)$
Distancia de Chebyshev pesada	$d(x,y) = max_{1 \leq i \leq n} \alpha_i d_i(x_i, y_i)$

Proposición 10 *Sea $X = X_1 \times ... \times X_n$ el espacio de tuplas, $\mathcal{L}_i$ un lenguaje de patrones sobre el tipo básico X_i, Δ_i un operador binario de generalización de elementos en X_i con $(i = 1, ..., n)$ y $\mathcal{L} = (\mathcal{L}_1, ..., \mathcal{L}_n)$ el lenguaje de patrones de tuplas.*

La función $\Delta : X \times X \to \mathcal{L}$ definida por

$$\Delta((x_1, ..., x_n), (y_1, ..., y_n)) = (\Delta_1(x_1, y_1), ..., \Delta_n(x_n, y_n))$$

es un operador binario de generalización de tuplas.

Demostración. Debemos probar que:

$$(x_1, ..., x_n) \in Set((\Delta_1(x_1, y_1), ..., \Delta_n(x_n, y_n))) \tag{5.1}$$

y

$$(y_1, ..., y_n) \in Set((\Delta_1(x_1, y_1), ..., \Delta_n(x_n, y_n))). \tag{5.2}$$

Por definición 12,

$$Set((\Delta_1(x_1, y_1), ..., \Delta_n(x_n, y_n))) = \{(z_1, ..., z_n) \in X | z_i \in Set(\Delta_i(x_i, y_i)), i = 1, ..., n\}.$$

Dado que Δ_i es un operador binario de generalización , $x_i \in Set(\Delta_i(x_i, y_i)) \wedge y_i \in Set(\Delta_i(x_i, y_i)) \forall i = 1, ..., n$. En consecuencia se verifican las ecuaciones 5.1 y 5.2. □

Dado que los patrones en $\mathcal{L}$ son tuplas cuyos elementos son patrones en los lenguajes de patrones $\mathcal{L}_i$, la generalización de dos tuplas p y q de patrones puede ser definida como la tupla cuyas componentes son las generalizaciones de las respectivas componentes en p y q. Esto se formaliza en la proposición 11.

Proposición 11 *Sea $\mathcal{L} = (\mathcal{L}_1, ..., \mathcal{L}_n)$ un lenguaje de patrones de tuplas, con $\mathcal{L}_i$ un lenguaje de patrones sobre el tipo básico X_i y $\Delta_i^* : \mathcal{L}_i \times \mathcal{L}_i \to \mathcal{L}_i$ un operador binario de generalización de patrones en $\mathcal{L}_i$ con $(i = 1, ..., n)$.*

La función $\Delta^* : \mathcal{L} \times \mathcal{L} \to \mathcal{L}$ definida por

$$\Delta^*((p_1, ..., p_n), (q_1, ..., q_n)) = (\Delta_1^*(p_1, q_1), ..., \Delta_n^*(p_n, q_n))$$

es un operador binario de generalización de patrones en $\mathcal{L}$.

Demostración. Debemos probar que el operador de generalización dado en la proposición 11 satisface las condiciones necesarias para ser un operador binario de generalización de patrones, es decir

$$Set((p_1, ..., p_n)) \subseteq Set((\Delta_1^*(p_1, q_1), ..., \Delta_n^*(p_n, q_n))) \tag{5.3}$$

y

$$Set((q_1, ..., q_n)) \subseteq Set((\Delta_1^*(p_1, q_1), ..., \Delta_n^*(p_n, q_n))) \tag{5.4}$$

Podemos ver que 5.3 se cumple dado que por la definición 12

$$Set((p_1, ..., p_n)) = \{(z_1, ..., z_n) \in X | z_i \in Set(p_i), i = 1, ..., n\}$$

y

$$Set((\Delta_1^*(p_1, q_1), ..., \Delta_n^*(p_n, q_n))) = \{(z_1, ..., z_n) \in X | z_i \in Set(\Delta_i^*(p_i, q_i)), i = 1, ..., n\}.$$

Dado que Δ_i^* es un operador binario de generalización,

$$Set(p_i) \subseteq Set(\Delta_i^*(p_i, q_i)) \; y \; Set(q_i) \subseteq Set(\Delta_i^*(p_i, q_i)) \; \forall i = 1, ..., n.$$

En consecuencia 5.3 y 5.4 se cumplen. □

En HDCC, las generalizaciones de conjuntos unitarios son computadas como la generalización del elemento consigo mismo. Por lo tanto, los patrones asociados a los grupos iniciales con una única tupla está dado por

$$\Delta((x_1, ..., x_n)), (x_1, ..., x_n)) = ((\Delta(x_1, x_1), ..., \Delta(x_n, x_n))).$$

El siguiente teorema establece la propiedad de composabilidad del operador Δ.

Teorema 1 (Composabilidad de Δ) *El operador binario de generalización Δ para tuplas, dado en la proposición 10, cuando se aplica a tuplas del espacio $X = X_1 \times ... \times X_n$, donde $(X_i, d_i)(i = 1, ..., n))$ son espacios métricos equipados con operadores binarios de generalización Δ_i, es:*

(i) *Fuertemente acotado por d_T si Δ_i es fuertemente acotado por d_i, $\forall i : i = 1, ..., n$.*

(ii) *Débilmente acotado por d_T si Δ_i es fuertemente acotado por d_i, $\forall i : i = 1, ..., n$.*

Demostración.

(i) Sean $\Delta_i (i = 1, ..., n)$ operadores binarios de generalización para X_i fuertemente acotados por d_i. Por definición 7 tenemos que
$\forall e, e_1, e_2 \in X_i$: si $e \in Set(\Delta_i(e_1, e_2))$, entonces $d_i(e, e_1) \leq d_i(e_1, e_2) \vee d_i(e, e_2) \leq d_i(e_1, e_2)$.

Debemos probar que $\forall z, x, y \in X : si\, z \in Set(\Delta(x, y))$, entonces $d_T(z, x) \leq d_T(x, y) \vee d_T(z, y) \leq d_T(x, y)$

Supongamos que Δ no es fuertemente acotado, entonces

$$\exists z \in Set(\Delta(x, y)) :\sim (d_T(z, x) \leq d_T(x, y) \vee d_T(z, y) \leq d_T(x, y)) \Leftrightarrow$$

$$\Leftrightarrow \exists z \in Set(\Delta(x, y)) : d_T(z, x) > d_T(x, y) \wedge d_T(z, y) > d_T(x, y)$$

En consecuencia,

$$\exists i, i = 1, ..., n : (d_i(z_i, x_i) > d_i(x_i, y_i) \wedge \forall j, j \neq i : d_j(z_j, x_j) \geq d_j(x_j, y_j)) \wedge$$

$$\exists i, i = 1, ..., n : (d_i(z_i, y_i) > d_i(x_i, y_i) \wedge \forall j, j \neq i : d_j(z_j, y_j) \geq d_j(x_j, y_j))$$

Por lo tanto, debe existir al menos un Δ_i que no es fuertemente acotado por d_i, lo que resulta en una contradicción.

(ii) Por el apartado (ii) de la proposición 2 y la parte (i) del teorema 1, es trivial ver que Δ está débilmente acotado por d_T.

□

Ejemplo 18 *Los dendrogramas mostrados en las figuras 4.1(a) y 4.1(c) del ejemplo 5 pueden verse como instanciaciones de tuplas en $X = \mathbb{R} \times \mathbb{R}$.*

En dicho ejemplo usamos como lenguaje de patrones los rectángulos mínimos que, visto desde el punto de vista de las tuplas, es equivalente a decir que usamos el lenguaje de los mínimos intervalos cerrados en cada una de las dimensiones de X ya que una tupla patrón, en este caso, describirá un rectángulo mínimo.

En cada dimensión, a su vez, empleamos la distancia de la diferencia absoluta entre números reales.

Como se puede ver en las figuras, el dendrograma conceptual bajo enlace simple no es equivalente al tradicional. Esto ocurre porque, si bien el operador binario de generalización Δ para tuplas dado en la definición 10 es, de acuerdo al teorema 1(i), fuertemente acotado por la distancia d_T ya que los operadores Δ_i^* se encuentran instanciados Δ_{num}^* que es fuertemente acotado, tenemos que Δ^* no es fuertemente acotado por d_L^s ya que la generalización de dos rectángulos p_1 y p_2 asociados a dos grupos C_1 y C_2 es un rectángulo p que cubre puntos que pueden caer fuera de las bolas con centro en los puntos de enlace de C_1 y de C_2 y radio $d_L^s(C_1, C_2, d)$ como ocurre por ejemplo con el grupo $\{i\}$ el cual es cubierto por p_4 (ver figura 3.7).

Notemos que esta misma situación podría presentarse para tuplas en $X_1 \times ... \times X_n$ cuando al menos dos de los dominios X_i son instanciados en $\mathbb{R}$. Para ello consideremos el siguiente ejemplo:

$$\begin{aligned} C_1 &= \{(0,0,x_3,...,x_n),(1,1,x_3,...,x_n),(2,2,x_3,...,x_n),(4,4,x_3,...,x_n)\} \\ C_2 &= \{(\texttt{5.1, 5.1}, x_3, ..., x_n)\} \end{aligned}$$

con patrones

$$\begin{aligned} p_{C_1} &= ([0,4],[0,4],p_3,...,p_n) \\ p_{C_2} &= ([\texttt{5.1},\texttt{5.1}],[\texttt{5.1},\texttt{5.1}],p_3,...,p_n). \end{aligned}$$

Tenemos que:

$$\begin{aligned} \Delta^*(p_{C_1}, p_{C_2}) &= ([0,\texttt{5.1}],[0,\texttt{5.1}],p_3,...,p_n) = p \\ d_L^s(C_1, C_2, d_T) &= \texttt{1.55} \qquad \textit{con } d_T \textit{ la distancia Euclidea.} \end{aligned}$$

Sin embargo, existe la tupla $x = (\texttt{4.5}, \texttt{0.5}, x_3, ..., x_n) \in Set(p)$ tal que

$$\begin{aligned} d_L^s(x, C_1, d_T) &= \texttt{2.91} > \texttt{1.55} \quad y \\ d_L^s(x, C_2, d_T) &= \texttt{4.63} > \texttt{1.55}. \end{aligned}$$

En efecto la propiedad de composabilidad de Δ^* puede solamente ser probada con respecto a la distancia de enlace completo d_L^c, como lo establece el teorema 2.

Teorema 2 (Composabilidad de Δ^*) *El operador binario de generalización de patrones Δ^* para tuplas en el espacio $X = X_1 \times ... \times X_n$ dado en la proposición*

11, aplicado a patrones en el espacio $\mathcal{L} = \mathcal{L}_1 \times ... \times \mathcal{L}_n$ *donde* $\mathcal{L}_i (i = 1, ..., n)$ *son lenguajes de patrones para elementos en* X_i *y* (X_i, d_i) *son espacios métricos equipados con operadores binarios de generalización de patrones* Δ_i^**, es:*

(i) *Fuertemente acotado por la distancia de enlace completo* d_L^c *si* Δ_i^* *es fuertemente acotado por* $d_L^c, \forall i : i = 1, ..., n$.

(ii) *Débilmente acotado por la distancia de enlace completo* d_L^c *si* Δ_i^* *es fuertemente acotado por* $d_L^c, \forall i : i = 1, ..., n$.

(iii) *Aceptable si* Δ_i^* *es aceptable,* $\forall i : i = 1, ..., n$.

Demostración.

(i) Sabemos que, por definición 6, Δ^* es fuertemente acotado por d_L^c sii:

$\forall C \subseteq Set(\Delta^*(p_1, p_2)) - (Set(p_1) \cup Set(p_2)), p_1 \in \mathcal{L}, p_2 \in \mathcal{L},$
$C_1 \subseteq Set(p_1), C_2 \subseteq Set(p_2)$:

$$d_L^c(C, C_1, d_T) \leq d_L^c(C_1, C_2, d_T) \vee d_L^c(C, C_2, d_T) \leq d_L^c(C_1, C_2, d_T).$$

Supongamos que Δ^* no es fuertemente acotado por d_L^c, por lo tanto

$\exists C_3 \subseteq Set(\Delta^*(p_1, p_2)) - (Set(p_1) \cup Set(p_2)), p_1 \in \mathcal{L}, p_2 \in \mathcal{L},$
$C_1 \subseteq Set(p_1), C_2 \subseteq Set(p_2)$:

$$d_L^c(C_3, C_1, d_T) > d_L^c(C_1, C_2, d_T) \wedge d_L^c(C_3, C_2, d_T) > d_L^c(C_1, C_2, d_T).$$

En consecuencia los puntos de enlace:

- $z \in C_3 \wedge w \in C_1$ entre C_3 y C_1,
- $x \in C_1 \wedge y \in C_2$ entre C_1 y C_2,
- $v \in C_3 \wedge u \in C_2$ entre C_3 y C_2,

satisfacen

$$d_T(z, w) > d_T(x, y) \wedge d_T(v, u) > d_T(x, y).$$

$\Rightarrow$ por las definiciones de d_T

$\exists i(i = 1, ..., n)$:

$$(d_i(z_i, w_i) > d_i(x_i, y_i) \wedge \forall j(j \neq i) : d_j(z_j, w_j) \geq d_j(x_j, y_j)) \tag{5.5}$$

$\wedge$
$\exists i(i = 1, ..., n)$:

$$(d_i(v_i, u_i) > d_i(x_i, y_i) \wedge \forall j(j \neq i) : d_j(v_j, u_j) \geq d_j(x_j, y_j)) \tag{5.6}$$

Dado que los operadores $\Delta_i^*(i = 1, ..., n)$ son fuertemente acotados por d_L^c,
$\forall C_k \subseteq Set(\Delta_i^*(p_{1i}, p_{2i})) - (Set(p_{1i}) \cup Set(p_{2i})), p_{1i} \in \mathcal{L}_i, p_{2i} \in \mathcal{L}_i$,
$C_{1i} \subseteq Set(p_{1i}), C_{2i} \subseteq Set(p_{2i})$:

$$d_L^c(C_k, C_{1i}, d_i) \leq d_L^c(C_{1i}, C_{2i}, d_i) \vee d_L^c(C_k, C_{2i}, d_i) \leq d_L^c(C_{1i}, C_{2i}, d_i).$$

En consecuencia los puntos de enlace:

- $z_i \in C_k \wedge w_i \in C_{1i}$ entre C_k y C_{1i},
- $x_i \in C_{1i} \wedge y_i \in C_{2i}$ entre C_{1i} y C_{2i},
- $v_i \in C_k \wedge u_i \in C_{2i}$ entre C_k y C_{2i},

satisfacen

$$d_i(z_i, w_i) \leq d_i(x_i, y_i) \tag{5.7}$$

$\vee$

$$d_i(v_i, u_i) \leq d_i(x_i, y_i) \tag{5.8}$$

Supongamos que 5.8 no se cumple, por lo que 5.7 debe ser cierta. Dado que $d_i(z_i, w_i)$ es la máxima distancia entre dos puntos en C_k y C_{1i}, y $d_i(x_i, y_i)$ es la máxima distancia entre dos puntos en C_{1i} y C_{2i}, entonces no es posible hacer cierta la ecuación 5.5. Un razonamiento análogo podemos aplicar si suponemos que 5.7 no se cumple, contradiciendo 5.6.

En consecuencia, Δ^* está fuertemente acotado por d_L^c.

(ii) Por el apartado (i) de la proposición 2 y el teorema 2 (i), es trivial ver que Δ^* está débilmente acotado por d_L^c.

(iii) Δ^* es aceptable sii $\forall p_1, p_2 \in \mathcal{L}, x \in Set(\Delta^*(p_1, p_2)), \exists x' \in Set(p_1) \cup Set(p_2) : d_T(x, x') \leq d_L^c(Set(p_1), Set(p_2), d_T)$.

Supongamos que Δ^* no es aceptable, entonces $\exists x \in Set(\Delta^*(p_1, p_2)) : (\forall x' \in Set(p_1) \cup Set(p_2) : d_T(x, x') > d_L^c(Set(p_1), Set(p_2), d_T))$.

En consecuencia,

$$\exists x \in Set(\Delta^*(p_1, p_2)) : (\forall x' \in Set(p_1) \cup Set(p_2) : d_T(x, x') > d_T(y_1, y_2)) \tag{5.9}$$

donde $y_1 \in Set(p_1)$ y $y_2 \in Set(p_2)$ son los puntos más distantes entre $Set(p_1)$ y $Set(p_2)$.

De acuerdo a las definiciones de las distancias entre tuplas d_T consideradas, para que la ecuación 5.9 se satisfaga debe cumplirse que

$$\exists i(i = 1, ..., n) : (d_i(x_i, x_i') > d_i(y_{1i}, y_{2i}) \wedge \forall j(j \neq i) : d_j(x_j, x_j') \geq d_j(y_{1j}, y_{2j})) \tag{5.10}$$

Siendo que los operadores $\Delta_i^*(i = 1, ..., n)$ son aceptables, entonces

$\forall p_{1i}, p_{2i} \in \mathcal{L}_i, x_i \in Set(\Delta^*(p_{1i}, p_{2i})), \exists x_i' \in Set(p_{1i}) \cup Set(p_{2i}) :$

$$d_i(x_i, x_i') \leq d_L^c(Set(p_{1i}), Set(p_{2i}), d_i).$$

$\Rightarrow$

$\forall p_{1i}, p_{2i} \in \mathcal{L}_i, x_i \in Set(\Delta^*(p_{1i}, p_{2i})), \exists x_i' \in Set(p_{1i}) \cup Set(p_{2i}) :$

$$d_i(x_i, x_i') \leq d_i(y_{1i}, y_{2i}),$$

donde y_{1i}, y_{2i} son los puntos más distantes entre $Set(p_{1i})$ y $Set(p_{2i})$, lo que contradice la ecuación 5.10

□

Ejemplo 19 *Podemos ver que la aplicación de HDCC bajo enlace completo del ejemplo 5 (página 33) produce un dendrograma que es equivalente al tradicional, tal como lo establece la proposición 1 dado que el apartado (i) de los teoremas 1 y 2 se cumplen para los operadores Δ y Δ^* usados en el ejemplo.*

6

Experimentos

En el capítulo 5 propusimos pares de operadores de generalización-distancias para tuplas de datos numéricos y nominales, los cuales, aplicados a HDCC bajo la distancia de enlace completo d_L^c producen dendrogramas conceptuales equivalentes al dendrograma tradicional, con la ventaja adicional de proveer descripciones para cada uno de los grupos de la jerarquía, solucionando de esta manera el problema de la interpretabilidad asociado al agrupamiento jerárquico además de brindarnos generalizaciones que son totalmente consistentes con las distancias subyacentes.

Hemos visto también que, los mismos operadores, cuando son usados bajo la distancia de enlace simple d_L^s pueden producir dendrogramas que no son equivalentes.

Teniendo en cuenta las consideraciones previas, los experimentos descritos en este capítulo están orientados a:

(i) Ilustrar empíricamente el resultado de composabilidad de los operadores Δ y Δ^* para tuplas, instanciando HDCC para aprendizaje proposicional bajo enlace completo con los operadores propuestos en el capítulo 5 y para un dataset real. Este experimento es presentado en la sección 6.1.

(ii) Mostrar que los nuevos dendrogramas (dendrogramas conceptuales), resultantes del proceso de cobertura y reorganización que lleva a cabo HDCC, si bien pueden no ser equivalentes a la correspondiente versión producida por el algoritmo tradicional de agrupamiento jerárquico, no degradan la calidad del mismo. Estos experimentos son presentados en la sección 6.2.

6.1. Experimento 1: HDCC aplicado al dataset Iris

El primero de los experimentos fue llevado a cabo sobre el conjunto de datos *Iris* [Black and Merz, 1998]. Este es un conjunto de datos muy conocido que consiste de la clase más cuatro atributos numéricos. Cada clase corresponde a un tipo de planta: Iris Setosa, Iris Versicolor e Iris Virginica. Los atributos numéricos corresponden a las longitudes y anchos de los sépalos y de los pétalos medidos en centímetros. Los datos que conforman el dataset se encuentran listados en la tabla A.1 del apéndice A. El dataset cuenta con 150 instancias (50 instancias de cada clase).

Para evaluar la calidad del agrupamiento empleamos dos medidas diferentes:

(i) Una medida interna [1] llamada S que refleja la dispersión media sobre k grupos con n_i $(i = 1, .., k)$ instancias cada uno.

Esta medida viene dada por la ecuación 6.1 donde, en nuestro caso, d denota la distancia Euclídea.

$$S = \frac{1}{k} \sum_{i=1}^{k} \sqrt{\sum_{j=1}^{n_i} \sum_{l=j+1}^{n_i} d(\vec{x}_j, \vec{x}_l)^2} \tag{6.1}$$

Cuánto más bajo sea el valor de S mejor será la calidad del agrupamiento.

(ii) Una medida externa, la pureza P dada por la ecuación 6.2, donde k es el número de grupos, n es el número total de instancias y n_{ij} el número de instancias en el grupo i de clase j.

$$P = \frac{1}{n} \sum_{i=1}^{k} max_j(n_{ij}) \tag{6.2}$$

La pureza puede ser interpretada como la precisión (accuracy) en clasificación, bajo la asunción de que todos los objetos de un grupo son clasificados como miembros de la clase dominante en el grupo.

[1] Una medida interna es una función de los datos y/o similitudes, mientras que una externa usa información externa tal como la clase.

Aunque la clase fue considerada para obtener las medidas de pureza de cada grupo, ésta fue eliminada del dataset a la hora de construir los grupos con HDCC.

La tabla 6.1 muestra los patrones retornados por HDCC considerando las distancias de enlace simple y completo para cada uno de los tres grupos.

Tabla 6.1: Patrones computados para los tres grupos por HDCC usando las distancias de enlace d_L^s y d_L^c.

		Patrones en $\mathcal{L}$
C_1	Enlace Simple d_L^s	([4.3, 5.8], [2.3, 4.4], [1.0, 1.9], [0.1, 0.6])
	Enlace Completo d_L^c	([4.3, 5.8], [2.3, 4.4], [1.0, 1.9], [0.1, 0.6])
C_2	Enlace Simple d_L^s	([4.9, 7.7], [2.0, 3.6], [3.0, 6.9], [1.0, 2.5])
	Enlace Completo d_L^c	([4.9, 6.1], [2.0, 3.0], [3.0, 4.5], [1.0, 1.7])
C_3	Enlace Simple d_L^s	([7.7, 7.9], [3.8, 3.8], [6.4, 6.7], [2.0, 2.2])
	Enlace Completo d_L^c	([5.6, 7.9], [2.2, 3.8], [4.3, 6.9], [1.2, 2.5])

Cada patrón es una 4-upla donde la componente i es también un patrón que provee una descripción del atributo i común a todos los elementos en el grupo. En el grupo C_1 la clase dominante fue *Iris Setosa*, en C_2 fue *Iris Versicolor* y en C_3 fue *Iris Virginica*.

En efecto, cada uno de esos patrones puede ser visto como una regla. Por ejemplo, el patrón descubierto para C_1 tanto bajo enlace simple como completo corresponde a la tupla

$$([4.3, 5.8], [2.3, 4.4], [1.0, 1.9], [0.1, 0.6])$$

la cual puede ser interpretada como la regla

`(sepallength` $\leq$ `4.3 AND sepallength` $\leq$ `5.8 AND sepalwidth` $\geq$ `2.3 AND sepalwidth` $\leq$ `4.4 AND petallength` $\geq$ `1.0 AND petallength` $\leq$ `1.9 AND petalwidth` $\geq$ `0.1 AND petalwidth` $\leq$ `0.6)`

donde `sepallength`, `sepalwidth`, `petallength` y `petalwidth` son los cuatro atributos numéricos del dataset.

La tabla 6.2 muestra los valores de S y P para HDCC así como para la versión tradicional del algoritmo de agrupamiento jerárquico. En ambos casos las medidas consideradas fueron calculadas usando la distancia de enlace simple y completo. El valor de k, que corresponde al número de grupos, fue fijado igual al número de clases en el dataset (en este caso $k = 3$).

Tabla 6.2: Valores de S y P para k grupos ($k = 3$) descubiertos en el agrupamiento tradicional y en el conceptual, usando las distancias de enlace simple d_L^s y completo d_L^c.

Distancia de enlace	$S_{Tradicional}$	$S_{Conceptual}$	$P_{Tradicional}$	$P_{Conceptual}$
Simple d_L^s	46.56	46.56	0.68	0.68
Completo d_L^c	37.44	37.44	0.84	0.84

Como se puede ver en los resultados de la tabla 6.2, la calidad de los grupos conceptuales no difieren del agrupamiento jerárquico tradicional (en este dataset en particular, incluso coinciden también para el caso de la distancia de enlace simple, cosa que podría no suceder con otros datasets), además de proveernos con útiles descripciones que permiten la interpretación del significado de cada grupo de instancias.

6.2. Experimento 2: HDCC aplicado a n distribuciones Gaussianas

Los experimentos presentados en esta sección fueron llevados a cabo a efectos de comprobar que los dendrogramas conceptuales no degradan la calidad de los agrupamientos obtenidos con respecto al agrupamiento jerárquico tradicional.

Para ello, construimos 100 datasets artificiales generando puntos a partir de una mezcla finita de k distribuciones gaussianas en $\mathbb{R}^2$ cuyas medias fueron ubicadas de forma aleatoria en $[0, 100] \times [0, 100]$ y usando desviaciones estándar de 1. Aunque k representa el número real de distribuciones gaussianas en cada uno de los datasets, notemos que puede existir solapamiento entre los puntos de las gaussianas y por lo tanto existir un número menor de grupos.

El valor de k fue instanciado a 3, y para cada dataset generamos 600 puntos (3 grupos de 200 puntos obtenidos a partir de cada una de las 3 distribuciones gaussianas).

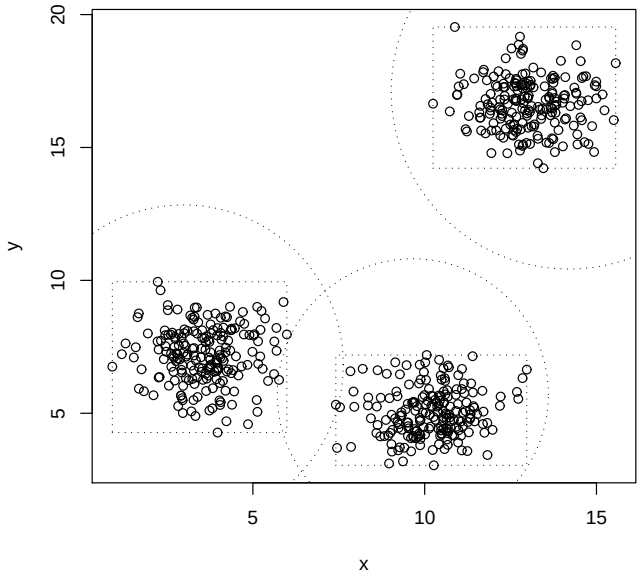
Los experimentos fueron realizados bajo enlace simple y enlace completo. En ambos casos se emplearon dos lenguajes de patrones $\mathcal{L}_R$ y $\mathcal{L}_C$ diferentes. $\mathcal{L}_R$ denota el lenguaje de los rectángulos de ejes paralelos y $\mathcal{L}_C$ el del los círculos. En el caso de $\mathcal{L}_R$, usamos los operadores de generalización Δ y Δ^* propuestos para tuplas en las proposiciones 10 y 11, conjuntamente con los operadores Δ_{num} y Δ^*_{num} propuestos para números reales en las proposiciones 7 y 8. En el caso de $\mathcal{L}_C$, $\Delta(e, e)$ fue calculado como $circ(e, 0)$ es decir como el círculo con centro en el punto e y radio 0, y $\Delta^*(circ(e_1, r_1), circ(e_2, r_2)) = circ(\frac{(e_1+e_2)}{2}, \frac{d(e_1,e_2)}{2} + max(r_1, r_2))$, o sea que la generalización de dos círculos es el círculo con centro en el centroide de ambos círculos y cuyo radio es igual a la mitad de la distancia entre los centros de ambos círculos mas el mayor de los radios de los mismos.

Las figuras 6.1(a) y 6.1(b) muestran los patrones descubiertos en $\mathcal{L}_R$ y $\mathcal{L}_C$ para los 600 puntos obtenidos de las distribuciones gaussianas para uno de los 100 datasets (en la primera usando enlace simple y en la segunda bajo enlace completo). Notemos que los rectángulos obtenidos incrementalmente por HDCC se ajustan a los puntos igual que un operador de generalización Δ_X aplicado sobre cada uno de los grupos. Esto no ocurre en $\mathcal{L}_C$ donde los patrones descubiertos son mas generales que si se hubiese usado un operadorΔ_X que retorne el mínimo círculo, con centro en el centroide del grupo. Sin embargo, como podemos ver en la tabla 6.3, esto no afecta a la calidad del agrupamiento ya que los grupos son construidos incrementalmente y HDCC en cada paso solamente une aquellos grupos que se encuentran completamente cubiertos por el patrón.

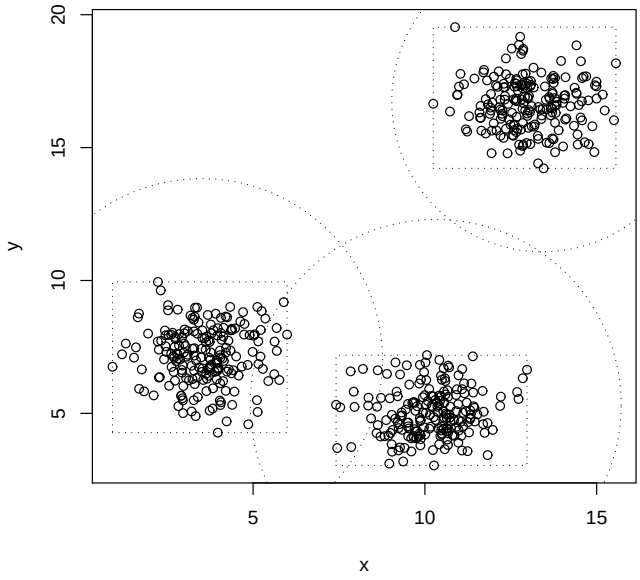
Para evaluar la calidad de los agrupamientos, empleamos la media de la medida S dada en la ecuación 6.1 la cual mide la dispersión de cada agrupamiento.

Notemos que n (en este caso, n es el número de datasets) puede tomar valores menores a 100 en el cálculo de la media de S para HDCC ya que no siempre las 100 jerarquías resultantes tendrán k grupos dado que varios grupos pueden ser unidos por un patrón en un mismo nivel de la jerarquía.

Los experimentos muestran que no sólo la calidad no es degradada por HDCC sino que, para el caso del lenguaje $\mathcal{L}_C$, HDCC mejora al dendrograma tradicional



(a) Patrones bajo d_L^s.



(b) Patrones bajo d_L^c

Figura 6.1: Patrones en $\mathcal{L}_R$ y $\mathcal{L}_C$ bajo las distancias de enlace d_L^s y d_L^c.

Tabla 6.3: Valores medios de S para los dendrogramas tradicionales y conceptuales obtenidos de los 100 datasets con $k = 3$.

		Dendrograma Tradicional		**Dendrograma Conceptual**		**Relación de Equivalencia**
		S	n	S	n	
$\mathcal{L}_R$	d_L^s	292.29	(100)	292.29	(100)	No equivalentes
	d_L^c	281.39	(100)	281.39	(100)	Equivalentes
$\mathcal{L}_C$	d_L^s	292.29	(100)	281.54	(98)	No equivalentes
	d_L^c	281.39	(100)	281.72	(100)	No equivalentes

bajo la distancia de enlace simple.

El mismo resultado fue confirmado por otros cuatro experimentos llevados a cabo, en donde variamos los valores de las medias y las desviaciones estándar en las distribuciones Gaussianas. Las medias y desviaciones fueron instanciadas a los siguientes valores:

(i) $\sigma = 1$ y $\mu \in [0, 10] \times [0, 10]$

(ii) $\sigma = 1$ y $\mu \in [0, 200] \times [0, 200]$

(iii) $\sigma = 5$ y $\mu \in [0, 100] \times [0, 100]$

(iv) $\sigma = 5$ y $\mu \in [0, 200] \times [0, 200]$

Los valores medios de S obtenidos de los experimentos llevados a cabo sobre 100 datasets con puntos extraídos de 3 distribuciones Gaussianas con medias y desviaciones de acuerdo a (i), (ii), (iii) y (iv) se muestran en la tabla 6.4.

A partir de los resultados de los experimentos mostrados en las tablas 6.3 y 6.4 se pude observar, además, que:

1. Tanto en HDCC como con el algoritmo tradicional, bajo la distancia d_L^s se obtuvieron calidades iguales o mayores al usar el lenguaje de patrones $\mathcal{L}_C$, mientras que bajo la distancia d_L^c se mejoró al usar el lenguaje de patrones $\mathcal{L}_R$.

2. En la mayor parte de los casos, la calidad fue mejor bajo la distancia de enlace completo d_L^c que bajo la de enlace simple d_L^s.

3. En la mayor parte de los casos, la calidad de HDCC fue similar o igual a la del algoritmo tradicional.

De lo expuesto anteriormente podemos concluir que la mejor calidad para HDCC se obtiene al usar la distancia d_L^c con el lenguaje $\mathcal{L}_R$, la cual coincide con la del tradicional (por la proposición 1 y la parte (i) de los teoremas 2 y 1). Al usar enlace simple, la mejor calidad de HDCC es lograda en combinación con el lenguaje $\mathcal{L}_C$ que es, en general, mejor que la del algoritmo tradicional (aunque no mejor que la obtenida por HDCC bajo d_L^c y $\mathcal{L}_R$). Esto no parece ser un problema sino, por el contrario, un resultado positivo ya que, según [Jain and Dubes, 1988], se ha observado en la práctica que, en muchas aplicaciones , el enlace completo produce jerarquías más utiles que el enlace simple.

Tabla 6.4: Valores medios de S sobre 100 datasets para HDCC (Conc.) y el algoritmo tradicional de agrupamiento jerárquico (Trad.) para 3 distribuciones Gaussianas con (i) $\sigma = 1$ y $\mu \in [0, 10] \times [0, 10]$; (ii) $\sigma = 1$ y $\mu \in [0, 200] \times [0, 200]$; (iii) $\sigma = 5$ y $\mu \in [0, 100] \times [0, 100]$; (iv)$\sigma = 5$ y $\mu \in [0, 200] \times [0, 200]$.

		Trad. (i)	**Conc. (i)**	**Trad. (ii)**	**Conc. (ii)**	**Trad. (iii)**	**Conc. (iii)**	**Trad. (iv)**	**Conc. (iv)**
$\mathcal{L}_R$	d_L^s	524.82	514.41	282.60	282.60	1830.42	1851.40	1607.84	1595.19
	d_L^c	285.62	285.62	282.60	282.60	1401.35	1401.35	1410.49	1410.49
$\mathcal{L}_C$	d_L^s	524.82	387.98	282.60	282.60	1830.42	1561.27	1607.84	1508.09
	d_L^c	285.62	286.82	282.60	282.60	1401.35	1412.50	1410.49	1409.28

7

Conclusiones y trabajos futuros

En este capítulo se describen las conclusiones de esta tesis de máster así como posibles líneas de trabajo a abordar en el futuro.

7.1. Conclusiones

En esta tesis hemos presentado una aproximación general al agrupamiento jerárquico conceptual basado en distancias y operadores de generalización. Esta aproximación añade a la flexibilidad inherente al agrupamiento jerárquico basado en distancias la interpretabilidad proporcionada por el agrupamiento conceptual, permitiendole al usuario elegir cualquier grupo de un dendrograma conceptual, obtener una descripción conceptual expresada en algún lenguaje de patrones a la vez de conocer si todos los elementos cubiertos por el concepto se encuentran próximos con respecto a la distancia subyacente.

Fácilmente hemos podido comprobar que para tipos de datos complejos (secuencias, grafos, etc.) los dendrogramas originales son usualmente diferentes a los obtenidos por generalización, por lo que hemos propuesto tres propiedades de consistencia que deben ser analizadas para cada par de distancia y operador de generalización. Algunos pares de distancias y operadores de generalización son compatibles a un cierto grado, resultando en dendrogramas conceptuales que son equivalentes, preservan el orden o son aceptables, mientras que otros pares pueden no alcanzar ninguno de los tres niveles, mostrando que algunas distancias y operadores de generalización no deberían ser usados en forma conjunta.

Hemos visto que la instanciación de HDCC para el caso proposicional es directa cuando los tipos de datos son numéricos y nominales. Demostramos que los

operadores de generalización para datos nominales (conjuntos por extensión) y datos numéricos (intervalos) son fuertemente acotados en los espacios métricos definidos por las funciones de distancia comúnmente usadas para esos tipos de datos (la distancia discreta y la distancia de la diferencia absoluta). Por lo que, de acuerdo a uno de los resultado teóricos obtenidos, pudimos aseverar además que en este caso los dendrogramas conceptuales basados en distancia son equivalentes a los dendrogramas basados en distancias clásicos independientemente de la distancia de enlazado empleada.

Las cosas son mas variadas e interesantes cuando aplicamos la propuesta a tipos de datos estructurados. Hemos visto a través de varios ejemplos en esta tesis que las condiciones se mantienen algunas veces para un tipo de distancia de enlazado pero no para otras, o que sólo uno de los niveles de consistencia es alcanzado (el más débil, de aceptabilidad). En particular, analizamos el problema en profundidad para un tipo de dato estructurado: las tuplas, las cuales son comúnmente usadas en aprendizaje proposicional. En este caso empleamos una extensión de los operadores de generalización, definidos sobre la base de los operadores para los tipos de las componentes y distancias definidas sobre las distancias para los tipos de las componentes, obteniendo, en este caso, resultados mucho mas positivos. En efecto, encontramos que la más fuerte de las propiedades se mantiene para este tipo de datos para aplicaciones de HDCC bajo la distancia de enlazado completo. Asimismo, llevamos a cabo experimentos para tuplas de números reales donde pudimos observar que la calidad de los clusters obtenidos en HDCC no se ve degradada con respecto al algortimo tradicional sino que, por el contrario, en algunos caso la mejora. A partir de estos resultados teóricos y experimentales, podemos afirmar que la integración del agrupamiento jerárquico basado en distancias y el conceptual para datos proposicionales, es decir, tablas (las cuales aun conforman el grueso de las aplicaciones de minería de datos) es posible, congruente y relativamente directo.

Adicionalmente, es importante destacar que el resultado de composabilidad encontrado para el tipo de datos tupla y varias distancias de tuplas, permite el tratamiento de información más elaborada en la forma de tabla, donde algunos de los atributos pueden tener estructura, siempre que contemos con operadores y distancias para cada atributo adecuados al nivel de consistencia deseado.

Finalmente, como ya hemos dicho, HDCC puede ser visto como un operador n-ario construido sobre operadores de generalización binarios aplicados a lo sumo n veces, siendo n el número de ejemplos. Esta es una propiedad interesante para áreas de aprendizaje automático donde existen operadores de generalización binarios bien establecidos, como es el caso de ILP.

7.2. Trabajos futuros

A continuación, se pretende describir posibles líneas de investigación que pudieran dar origen a trabajos futuros.

El trabajo futuro está focalizado, fundamentalmente, en encontrar pares consistentes de operadores de generalización y distancias para otros tipos de datos de interés.

Por un lado, nos interesa abordar el estudio de aquellos tipos de datos comunes en aplicaciones de minería de datos en la web tales como las secuencias, los grafos y los objetos multimedia. También se pretende analizar otros tipos de datos estructurados como los conjuntos, los átomos y las cláusulas.

En todos los casos se estudiarán diversas distancias ya existentes en la literatura, como por ejemplo la distancia de Hausdorff para conjuntos o las variantes de ésta, propuestas en [Eiter and Mannila, 1997]; la de Bunke [Bunke, 1997] para grafos; la de J. Ramon [Ramon et al., 1970] o la de Hutchinson [Hutchinson, 1997] para átomos. En el caso de las cláusulas, las mismas pueden ser vistas como conjuntos de átomos por lo que las distancias definidas para conjuntos en combinación con las de átomos podrán ser analizadas. En cuanto a los operadores de generalización, se analizarán los más comúnmente usados, como por ejemplo el *lgg* para átomos o cláusulas, o la unión para el caso de los conjuntos, a la vez que nuevos operadores pueden ser propuestos y analizados.

Otra posible línea a investigar consiste en extender el análisis de consistencia entre distancias y generalizaciones a otras técnicas de agrupamiento basadas en distancias, como por ejemplo el algoritmo de agrupamiento k-medias.

Bibliografía

[Berkhin, 2006] Berkhin, P.: 2006, A survey of clustering data mining techniques, in *Grouping Multidimensional Data*, pp. 25–71.

[Bisson, 1992] Bisson, G.: 1992, Conceptual clustering in a first order logic representation, in *European Conference on Artificial Intelligence*, pp. 458–462.

[Black and Merz, 1998] Black, C. and Merz, C. J.: 1998, *UCI Repository of Machine Learning Databases*.

[Blockeel and De Raedt, 1998] Blockeel, H. and De Raedt, L.: 1998, Top-down induction of first order logical decision trees, in *Artificial Intelligence*, Vol. 101, pp. 285–297.

[Blockeel et al., 1998] Blockeel, H., De Raedt, L., and Ramon, J.: 1998, Top-down induction of clustering trees, in *Proc. of the 15th International Conference on Machine Learning*, pp. 55–63.

[Bunke, 1997] Bunke, H.: 1997, On a relation between graph edit distance and maximum common subgraph, in *Pattern Recognition Letters*, Vol. 18, pp. 689–694.

[Cheng, 1997] Cheng, S.: 1997, Distance between herbrand interpretations: A measure for approximations to a target concept, in *LNCS*, Vol. 1297, pp. 213–226, Springer.

[Cover and Hart, 1967] Cover, T. and Hart, P.: 1967, Nearest neighbour pattern classification, in *IEEE Transactions on Information Theory*, pp. 13–27.

[De Raedt and Blockeel, 1997] De Raedt, L. and Blockeel, H.: 1997, Using logical decision trees for clustering, in Springer (ed.), *Proc. 7th Intl Workshop on ILP*, Vol. LNCS 1297, pp. 133–140.

[Eiter and Mannila, 1997] Eiter, T. and Mannila, H.: 1997, Distance measures for point of sets and their computation, in *Acta Informatica*, Vol. 34.

[Emde, 1994a] Emde, W.: 1994a, Inductive learning of characteristic concept descriptions, in *Proc. 4th Intl Workshop on Inductive Logic Programming (ILP-94)*.

[Emde, 1994b] Emde, W.: 1994b, Inductive learning of characteristic concept descriptions from small sets of classified examples, in F. Bergadano and L. De Raedt (eds.), *Proc. of the European Conference on Machine Learning 1994*, Vol. 784 of *LNAI*, pp. 103–121, Springer.

[Emde and Wettschereck, 1996] Emde, W. and Wettschereck, D.: 1996, Relational instance-base learning, in *Proc. of the Thirteen International Conference on Machine Learning (ICML'96)*, pp. 122–130, Morgan Kaufmann.

[Estruch, 2008] Estruch, V.: 2008, *Ph.D. thesis*, DSIC-UPV, http://www.dsic.upv.es/ vestruch/thesis.pdf.

[Fisher, 1987] Fisher, D.: 1987, Knowledge acquisition via incremental conceptual clustering, in *Machine Learning*, pp. 139–172.

[Fisher, 1936] Fisher, R.: 1936, The use of multiple measurements in taxonomic problems, in *Ann. Eurgenics*, Vol. 7, Part II, pp. 179–188.

[Gluck and Corter, 1985] Gluck, M. A. and Corter, J. E.: 1985, Information, uncertainty and the utility of categories, in *Proc. of the 7th Annual Conference of the Cognitive Science Society*, pp. 283–287.

[Haykin, 1998] Haykin, S.: 1998, *Neural Networks -A Comprehensive Foundation*, Prentice-Hall, 2nd edition.

[Hernández-Orallo et al., 2004] Hernández-Orallo, J., Ramírez-Quintana, M., and Ferri, C.: 2004, *Introducción a la Minería de Datos*, Pearson Prentice-Hall.

[Hutchinson, 1997] Hutchinson, A.: 1997, Metrics on terms and clauses, in Springer-Verlag (ed.), *Proc. of the 9th European Conference on Machine Learning (ECML'1997)*, pp. 138–145.

[Hutchinson, 2002] Hutchinson, A.: 2002, *A Catalogue of Metrics*, Technical Report 02-01, Department of Computer Science, King's College London.

[Isasi and Galván, 2003] Isasi, P. and Galván, I.: 2003, *Las redes neuronales artificiales y sus aplicaciones prácticas*, Prentice-Hall.

[Jain and Dubes, 1988] Jain, A. K. and Dubes, R. C.: 1988, Algorithms for clustering data, in *Prentice-Hall advanced reference series*, Prentice-Hall.

[Jain et al., 1999] Jain, A. K., Murty, M. N., and Flynn, P. J.: 1999, Data clustering: a review, in *ACM Comput. Survey*, Vol. 31, pp. 264–323, ACM, New York, NY, USA.

[Johnson, 1987] Johnson, S. C.: 1987, Hierarchical clustering schemes, in *Psychometrika*, Vol. 2, pp. 241–254.

[Kirsten and Wrobel, 1998] Kirsten, M. and Wrobel, S.: 1998, Relational distance-based clustering, in *In D. Page (Ed.) Proc. Eighth Int. Conference on Inductive Logic Programming*, pp. 261–270, Springer, LNAI.

[Levenshtein, 1966] Levenshtein: 1966, Binary codes capable of correcting deletions, insertions, and reversals, in *Soviet Physics Doklady*, Vol. 10, pp. 707–710.

[MacQueen, 1967] MacQueen, J. B.: 1967, 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.

[Michalski, 1980] Michalski, R. S.: 1980, Knowledge acquisition through conceptual clustering, in *Policy Analysis and Information Systems*, Vol. 4, pp. 219–244.

[Michalski and Stepp, 1983] Michalski, R. S. and Stepp, R. E.: 1983, *Machine Learning: An Artificial Intelligence Approach*, Chapt. Learning from Observation: Conceptual Clustering, pp. 331–363, TIOGA Publishing Co.

[Muggleton and De Raedt, 1994] Muggleton, S. and De Raedt, L.: 1994, Inductive logic programming: Theory and methods, in *Journal of Logic Programming*, Vol. 19, pp. 629–679.

[Plotkin, 1970] Plotkin, G.: 1970, A note on inductive generalization, in *Machine Intelligence*, Vol. 5, pp. 153–163, Edimburgh University Press.

[Ramon et al., 1970] Ramon, J., Bruynooghe, M., and VanLaer, W.: 1970, Distance measures between atoms, in *CompulogNet Meeting ComputingLogic and Machine Learning*, pp. 35–41.

[Stanfill and Waltz, 1986] Stanfill, A. and Waltz, D.: 1986, Toward memory-based reasoning, in *Comm. of the ACM*, Vol. 22, pp. 1213–1228.

[Sycara et al., 1992] Sycara, K., Guttal, R., Koning, J., Narasimhan, S., and Navinchandra, D.: 1992, Cadet: A case-based synthesis tool for engineering design, in *International Journal of Expert Systems*, Vol. 4, pp. 157–188.

A

Conjunto de datos Iris

Tabla A.1: Instancias en el dataset Iris

SepalLength	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa
5.4	3.4	1.7	0.2	setosa

Tabla A.1: (continuación)

SepalLength	Sepal Width	Petal Length	Petal Width	Species
5.1	3.7	1.5	0.4	setosa
4.6	3.6	1.0	0.2	setosa
5.1	3.3	1.7	0.5	setosa
4.8	3.4	1.9	0.2	setosa
5.0	3.0	1.6	0.2	setosa
5.0	3.4	1.6	0.4	setosa
5.2	3.5	1.5	0.2	setosa
5.2	3.4	1.4	0.2	setosa
4.7	3.2	1.6	0.2	setosa
4.8	3.1	1.6	0.2	setosa
5.4	3.4	1.5	0.4	setosa
5.2	4.1	1.5	0.1	setosa
5.5	4.2	1.4	0.2	setosa
4.9	3.1	1.5	0.2	setosa
5.0	3.2	1.2	0.2	setosa
5.5	3.5	1.3	0.2	setosa
4.9	3.6	1.4	0.1	setosa
4.4	3.0	1.3	0.2	setosa
5.1	3.4	1.5	0.2	setosa
5.0	3.5	1.3	0.3	setosa
4.5	2.3	1.3	0.3	setosa
4.4	3.2	1.3	0.2	setosa
5.0	3.5	1.6	0.6	setosa
5.1	3.8	1.9	0.4	setosa
4.8	3.0	1.4	0.3	setosa
5.1	3.8	1.6	0.2	setosa
4.6	3.2	1.4	0.2	setosa
5.3	3.7	1.5	0.2	setosa
5.0	3.3	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor

Tabla A.1: (continuación)

SepalLength	Sepal Width	Petal Length	Petal Width	Species
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.5	2.8	4.6	1.5	versicolor
5.7	2.8	4.5	1.3	versicolor
6.3	3.3	4.7	1.6	versicolor
4.9	2.4	3.3	1.0	versicolor
6.6	2.9	4.6	1.3	versicolor
5.2	2.7	3.9	1.4	versicolor
5.0	2.0	3.5	1.0	versicolor
5.9	3.0	4.2	1.5	versicolor
6.0	2.2	4.0	1.0	versicolor
6.1	2.9	4.7	1.4	versicolor
5.6	2.9	3.6	1.3	versicolor
6.7	3.1	4.4	1.4	versicolor
5.6	3.0	4.5	1.5	versicolor
5.8	2.7	4.1	1.0	versicolor
6.2	2.2	4.5	1.5	versicolor
5.6	2.5	3.9	1.1	versicolor
5.9	3.2	4.8	1.8	versicolor
6.1	2.8	4.0	1.3	versicolor
6.3	2.5	4.9	1.5	versicolor
6.1	2.8	4.7	1.2	versicolor
6.4	2.9	4.3	1.3	versicolor
6.6	3.0	4.4	1.4	versicolor
6.8	2.8	4.8	1.4	versicolor
6.7	3.0	5.0	1.7	versicolor
6.0	2.9	4.5	1.5	versicolor
5.7	2.6	3.5	1.0	versicolor
5.5	2.4	3.8	1.1	versicolor

Tabla A.1: (continuación)

SepalLength	Sepal Width	Petal Length	Petal Width	Species
5.5	2.4	3.7	1.0	versicolor
5.8	2.7	3.9	1.2	versicolor
6.0	2.7	5.1	1.6	versicolor
5.4	3.0	4.5	1.5	versicolor
6.0	3.4	4.5	1.6	versicolor
6.7	3.1	4.7	1.5	versicolor
6.3	2.3	4.4	1.3	versicolor
5.6	3.0	4.1	1.3	versicolor
5.5	2.5	4.0	1.3	versicolor
5.5	2.6	4.4	1.2	versicolor
6.1	3.0	4.6	1.4	versicolor
5.8	2.6	4.0	1.2	versicolor
5.0	2.3	3.3	1.0	versicolor
5.6	2.7	4.2	1.3	versicolor
5.7	3.0	4.2	1.2	versicolor
5.7	2.9	4.2	1.3	versicolor
6.2	2.9	4.3	1.3	versicolor
5.1	2.5	3.0	1.1	versicolor
5.7	2.8	4.1	1.3	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica
6.5	3.0	5.8	2.2	virginica
7.6	3.0	6.6	2.1	virginica
4.9	2.5	4.5	1.7	virginica
7.3	2.9	6.3	1.8	virginica
6.7	2.5	5.8	1.8	virginica
7.2	3.6	6.1	2.5	virginica
6.5	3.2	5.1	2.0	virginica

Tabla A.1: (continuación)

SepalLength	Sepal Width	Petal Length	Petal Width	Species
6.4	2.7	5.3	1.9	virginica
6.8	3.0	5.5	2.1	virginica
5.7	2.5	5.0	2.0	virginica
5.8	2.8	5.1	2.4	virginica
6.4	3.2	5.3	2.3	virginica
6.5	3.0	5.5	1.8	virginica
7.7	3.8	6.7	2.2	virginica
7.7	2.6	6.9	2.3	virginica
6.0	2.2	5.0	1.5	virginica
6.9	3.2	5.7	2.3	virginica
5.6	2.8	4.9	2.0	virginica
7.7	2.8	6.7	2.0	virginica
6.3	2.7	4.9	1.8	virginica
6.7	3.3	5.7	2.1	virginica
7.2	3.2	6.0	1.8	virginica
6.2	2.8	4.8	1.8	virginica
6.1	3.0	4.9	1.8	virginica
6.4	2.8	5.6	2.1	virginica
7.2	3.0	5.8	1.6	virginica
7.4	2.8	6.1	1.9	virginica
7.9	3.8	6.4	2.0	virginica
6.4	2.8	5.6	2.2	virginica
6.3	2.8	5.1	1.5	virginica
6.1	2.6	5.6	1.4	virginica
7.7	3.0	6.1	2.3	virginica
6.3	3.4	5.6	2.4	virginica
6.4	3.1	5.5	1.8	virginica
6.0	3.0	4.8	1.8	virginica
6.9	3.1	5.4	2.1	virginica
6.7	3.1	5.6	2.4	virginica

Tabla A.1: (continuación)

SepalLength	Sepal Width	Petal Length	Petal Width	Species
6.9	3.1	5.1	2.3	virginica
5.8	2.7	5.1	1.9	virginica
6.8	3.2	5.9	2.3	virginica
6.7	3.3	5.7	2.5	virginica
6.7	3.0	5.2	2.3	virginica
6.3	2.5	5.0	1.9	virginica
6.5	3.0	5.2	2.0	virginica
6.2	3.4	5.4	2.3	virginica
5.9	3.0	5.1	1.8	virginica

B

Trabajos desarrollados en el marco de esta tesis

- **Internacionales**

  - A. Funes, C. Ferri, J. Hernández-Orallo, M. J. Ramírez-Quintana. *Hierarchical Distance-based Conceptual Clustering*. Proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2008, Part II, LNAI 5212, pp. 349 - 364. Springer (2008).

  - A. Funes, C. Ferri, J. Hernández-Orallo, M. J. Ramírez-Quintana. *An Instantiation of Hierarchical Distance-based Conceptual Clustering for Propositional Learning*. Sometido a la 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-09), 27 al 30 de Abril de 2009, Bangkok, Thailand.