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.