Agrupamiento Conceptual Jerárquico Basado en Distancias

Ana Funes[1], Ma. José Ramírez-Quintana[2], Jose Hernández-Orallo[2], César Ferri[2]
[1]Universidad Nacional de San Luis, Ejército de los Andes 950
5700 San Luis, Argentina
afunes@unsl.edu.ar,
[2] DSIC, Universidad Politécnica de Valencia,
Camino de Vera s/n, 46022 Valencia, España
{ mramirez, jorallo, cferri}@dsic.upv.es

CONTEXTO

El presente trabajo de investigación se encuentra enmarcado en el programa internacional de cooperación Alfa-LerNet entre el grupo ELP del Departamento de Sistemas Informáticos y Computación (DSIC) de la Universidad Politécnica de Valencia, España y el Proyecto de Incentivos código 22/F822 "Ingeniería de Software: Conceptos, Métodos y Herramientas en un contexto de Ingeniería de Software en Evolución" de la Universidad Nacional de San Luis, Argentina.

RESUMEN

En este trabajo de investigación analizamos la relación existente entre el agrupamiento jerárquico basado en distancia y los conceptos que se pueden inducir por generalización a partir de una jerarquía, mostrando que pueden surgir diversas inconsistencias a partir de incompatibilidades entre las distancias subyacentes y los operadores de generalización empleados.

En este contexto, hemos definido un marco teórico genérico en el cual, por un lado, hemos propuesto un nuevo algoritmo de agrupamiento en el que integramos el agrupamiento jerárquico basado en distancias y el agrupamiento conceptual, permitiendo observar en las nuevas jerarquías obtenidas si un elemento ha sido integrado a un grupo por la distancia de enlazado o porque se encuentra cubierto por el concepto asociado al grupo.

Por otro lado, hemos definido tres niveles diferentes de consistencia entre los operadores de generalización y las distancias a partir de la similitud existente entre las nuevas jerarquías conseguidas con nuestro algoritmo y las correspondientes obtenidas por el algoritmo tradicional jerárquico.

Actualmente, nos encontramos trabajando en el análisis de diversas instanciaciones del marco teórico genérico antes mencionado.

Palabras clave: agrupamiento conceptual, agrupamiento jerárquico, generalización, distancias.

1. INTRODUCCION

Es sabido que, en la Minería de Datos y en el Aprendizaje Automático, existen dos aproximaciones diferentes que dependen de los conceptos subyacentes empleados para las tareas de aprendizaje: el concepto de similitud y el concepto de generalización.

El concepto 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. Ejemplos de estás técnicas son el algoritmo k-NN (k Nearest Neighbours) [Cover and Hart, 1967], el algoritmo de

agrupamiento k-medias [MacQueen, 1967] y los discriminantes de Fisher [Fisher, 1936]. El concepto de distancia, en cambio, no sólo formaliza la noción de similitud entre dos individuos sino que brinda propiedades adicionales del espacio métrico, las cuales pueden ser explotadas de manera beneficiosa por muchas técnicas de aprendizaje. Estas técnicas son conocidas como técnicas basadas en distancias.

Por otro lado, nos encontramos con aquellas técnicas de aprendizaje conocidas como técnicas basadas en modelos o simbólicas, que producen un modelo que puede ser interpretado por el usuario. Estas, a diferencia de las anteriores, se apoyan en la idea de que una generalización o patrón descubierto a partir de un conjunto de datos puede ser usado para describir nuevos datos que sean cubiertos por el patrón. Entre las técnicas más conocidas en aprendizaje supervisado podemos citar los árboles de decisión y las reglas de asociación; entre las de aprendizaje no supervisado se encuentra el agrupamiento conceptual de Michalski [Michalski,1980; Michalski and Stepp, 1983].

Las técnicas basadas en distancias han demostrado ser útiles en la práctica ofreciendo buenas predicciones y si bien son bastante intuitivas y flexibles, ya que basta con disponer de una función de distancia o medida de similitud adecuada al dominio de los datos para poder aplicarlas, no nos brindan patrones, generalizaciones o explicaciones que justifiquen el porqué de las decisiones tomadas para cada individuo. En este sentido, si bien es muy útil conocer que un cierto individuo pertenece a un grupo porque se encuentra cerca, de acuerdo a una cierta distancia, a los otros elementos del grupo, es de mayor utilidad poder conocer también qué propiedades comparten todos los elementos del grupo. Esta falta de comprensibilidad es un problema que aqueja tanto a las técnicas de agrupamiento como de clasificación basadas en distancia.

En este trabajo nos hemos planteado combinar para la tarea de agrupamiento lo mejor de ambas técnicas: la comprensibilidad de las técnicas basadas en modelos con la flexibilidad de las basadas en distancias. Sin embargo, un problema importante que surge al combinar ambas aproximaciones es conocer si los patrones descubiertos para cada grupo por un operador de generalización dado son consistentes con la distancia empleada para construir los grupos. Con esto queremos significar que, para un conjunto de ejemplos y una generalización de los mismos, se espera que, si distancia y generalización dadas son consistentes, aquellos ejemplos que se encuentren cercanos en un espacio métrico de acuerdo a la distancia sean cubiertos por la generalización, mientras que aquellos que estén lejos se espera que se encuentren fuera de la cobertura de la generalización.

2. LINEAS DE INVESTIGACION y DESARROLLO

Hemos ya desarrollado una aproximación general a la tarea de agrupamiento conceptual basado en distancias. La parte central de esta aproximación es un algoritmo para agrupamiento conceptual jerárquico basado en distancias, que hemos llamado HDCC.

HDCC es una modificación al algoritmo de agrupamiento jerárquico [Jain et al., 1999; Berkhin, 2006; Jain and Dubes, 1988; Johnson, 1987] el cual usando una distancia para construir la jerarquía de grupos produce al mismo tiempo patrones que describen cada uno de los grupos descubiertos. Un aspecto importante 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, 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, fue necesario poder mostrar gráficamente y de forma clara cuándo esta situación ocurría. Esto nos ha llevado al desarrollo de una nueva representación gráfica de los dendrogramas, los cuales hemos llamado dendrogramas conceptuales donde se puede visualizar en cada grupo qué elementos han sido atraídos por la distancia y cuáles por el patrón o generalización. Por otro lado, también fue necesario poder analizar a priori si dichas inconsistencias aparecería o no. Esto último dio 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, cuando un 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 será al mismo tiempo consistente con la distancia y los conceptos expresados por cada patrón en la jerarquía.

Sobre esta base, hemos llevado a cabo una instanciación de nuestro marco teórico para el caso proposicional, donde propusimos un conjunto de pares de distancia y operador de generalización, los cuales usados en forma conjunta trabajan consistentemente para datos numéricos, nominales y tuplas.

La nueva línea de investigación encarada apunta al análisis de la combinación de diversos operadores de generalización y distancias para otros tipos de datos, empleando para esto distancias ya existentes en la literatura como por ejemplo la distancia de J. Ramon [Ramon et al., 1970], la de Shan-Hwei Nienhuys-Cheng [Cheng, 1997], o la de Hutchinson [Hutchinson, 1997] para átomos; la de Hausdorff para conjuntos o sus variantes propuestas en [Eiter and Mannila, 1997]; la de Bunke [Bunke, 1997] para grafos y la de Levenshtein [Levenshtein, 1966] para secuencias entre otras a considerar. En el caso de las cláusulas, las mismas pueden ser tratadas como conjuntos de átomos por lo que las distancias definidas para conjuntos en combinación con las de átomos podrían ser empleadas. En cuanto a los operadores de generalización, se analizarán los más comúnmente usados, como por ejemplo el operador de generalización menos general (*lgg*) de Plotkin [Plotkin, 1970] para átomos o cláusulas y la unión para el caso de los conjuntos, a la vez que nuevos operadores podrían 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.

3. RESULTADOS OBTENIDOS/ESPERADOS

En una primera etapa, el trabajo giró en torno al desarrollo de una aproximación híbrida sobre la base del agrupamiento jerárquico tradicional basado en distancias y el agrupamiento conceptual, lo que dio lugar a nuestro algoritmo HDCC. En ese contexto, definimos tres niveles de consistencia entre las distancias y las generalizaciones. Este trabajo resultó en la publicación [Funes et al., 2008].

Sobre esta base, propusimos una instanciación del marco teórico genérico, llevando a cabo un análisis para un conjunto de distancias y operadores de generalización útiles para el agrupamiento proposicional. Probamos que los intervalos conjuntamente con la distancia de la diferencia absoluta para los números reales, y la unión de conjuntos junto con la distancia discreta para datos nominales son altamente consistentes en HDCC. Más importante aún, demostramos que lo mismo ocurre cuando éstos son usados para tuplas de números reales y datos nominales como operadores de generalización y distancias. Este resultado de composabilidad para tuplas, en el caso de la distancia de

enlazado completo, fue obtenido independientemente de los tipos base. Esto significa que una cierta propiedad de consistencia es heredada por las tuplas cuando los operadores de generalización asociados a los tipos de datos de las componentes cumplen con dicha propiedad de consistencia, permitiendo así la instanciación directa del marco a tuplas de tipos de datos complejos. Además del análisis teórico, se llevaron a cabo algunos experimentos para validar e ilustrar la propuesta. Esta segunda etapa del trabajo resultó en la publicación [Funes et al., 2009].

Actualmente, nos encontramos trabajando en el análisis de otras combinaciones operativas de tipos de datos y operadores de generalización, resultando de especial interés aquellos tipos de datos comunes en aplicaciones de minería de datos en la web tales como secuencias, grafos y objetos multimedia. También se pretende analizar otros tipos de datos estructurados como los conjuntos, los átomos y las cláusulas.

4. FORMACION DE RECURSOS HUMANOS

El trabajo aquí presentado ya ha dado como resultado una tesis de maestría en Ingeniería de Software y Métodos Formales la cual fue defendida en la Universidad Politécnica de Valencia por la primera autora de este trabajo. Asimismo, ha servido como puntapié inicial para la realización, por parte de la misma, de su tesis de doctorado que se encuentra actualmente en desarrollo.
Asimismo, cabe destacar que este trabajo forma parte de una colaboración internacional entre investigadores de dos grupos de investigación de dos universidades, el SEG de la Universidad Nacional de San Luis y el ELP de la Universidad Politécnica de Valencia (España).

5. BIBLIOGRAFIA

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

[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.

[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.

[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.

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

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

[Funes et al., 2008] Funes, A., Ferri, C., Hernández-Orallo, J., Ramirez-Quintana, M.J.: Hierarchical Distancebased Conceptual Clustering. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 349–364. Springer, Heidelberg (2008)

[Funes et al., 2009] Funes, A., Ferri, C., Hernández-Orallo, J. and Ramírez-Quintana, M.J.: An Instantiation of Hierarchical Distance-based Conceptual Clustering for Propositional Learning. In: Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and

Data Mining (PAKDD-09), LNAI Nº 5476, pp. 637-646, Springer(2009).

[Hernández-Orallo et al., 2004] Hernández-Orallo, J., Ramírez-Quintana, M., and Ferri, C.: 2004, Introducci´on 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.

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

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

[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.

[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.