ALGUNOS RESULTADOS EXPERIMENTALES DE LA INTEGRACIÓN DE AGRUPAMIENTO E INDUCCIÓN COMO MÉTODO DE DESCUBRIMIENTO DE CONOCIMIENTO

Kogan, A.[1], Rancan, C. [2,3], Britos, P. [3,1], Pesado, P. [2,4], García-Martínez, R.[3,1]

[1] Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. UBA
[2] Programa de Doctorado en Ciencias Informáticas. Facultad de Informática. UNLP
[3] Centro de Ingeniería de Software Ingeniería del Conocimiento. Escuela de Postgrado. ITBA
[4] Instituto de Investigaciones en Informática LIDI. Facultad de Informática. UNLP - CIC

akogan@fi.uba.ar, claudioran@yahoo.com, pbritos@itba.edu.ar, ppesado@lidi.info.unlp.edu.ar, rgm@itba.edu.ar

1. Introducción

El descubrimiento de conocimiento (KD Knowledge Discovery) consiste en la búsqueda de patrones interesantes y de regularidades importantes en grandes bases de información [Holsheimer y Siebes, 1991; Piatetski-Shapiro *et al*., 1991]. Al hablar de descubrimiento de conocimiento basado en sistemas inteligentes nos referimos específicamente a la aplicación de métodos de aprendizaje automático u otros métodos similares, para descubrir y enumerar patrones presentes en dicha información.

Un procedimiento recurrente a la hora de realizar descubrimiento de conocimiento basado en sistemas inteligentes consiste en tomar el conjunto de datos a estudiar, aplicar un algoritmo de agrupamiento [Kaski, 1997, Hall y Colmes, 2003] para separarlo en distintos grupos (clases) y sobre cada uno de ellos, intentar generar reglas que caractericen su conformación, utilizando otro algoritmo a tales efectos [Grosser *et al*., 2005; Felgaer *et al*., 2006; Cogliati *et al*., 2006].

Una de las opciones para llevar adelante el proceso de agrupamiento está dada por el uso de los mapas auto-organizados [Kohonen, 1982; 1990; 1995a, 1995b; Kohonen *et al*., 1996], los cuales consisten en un algoritmo de redes neuronales utilizado para una gran variedad de aplicaciones, principalmente para problemas de ingeniería, pero también para análisis de datos.

En cuanto a la inducción de reglas, dada la caracterización de las entidades que se utilizan comúnmente en descubrimiento de conocimiento, fuertemente basada en los valores de sus atributos y no en las relaciones establecidas entre estos, se suelen emplear métodos basados en atributos. Uno de los más claros y difundidos son los árboles de decisión o clasificación [Michalski *et al*, 1998; Grossman *et al*, 1999] en los cuales se cuenta con nodos que modelizan cada atributo, ramas que se originan en estos nodos, una por cada valor que el atributo puede tomar, y finalmente las hojas que corresponden a las clases individuales. Recorriendo un árbol desde su nodo padre hasta las distintas hojas, se pueden generar de forma muy simple las reglas a las cuales la clasificación responde. Una de las herramientas aplicadas al mencionado proceso es la familia de algoritmos TDIDT (Top Down Induction Decision Trees) [Quinlan, 1986; Servente y García-Martínez, 2002]. Sin embargo, estos pasos se realizan únicamente bajo la presunción de obtener un resultado representativo del conjunto de datos sobre el que se trabaja.

2. El problema

Trabajos recientes sobre sistemas de ayuda a la toma de decisiones centrados en tecnología de sistemas basados en conocimiento [Sierra *et al*., 2006] en áreas como el control aéreo [Ierache y García-Martínez, 2004] o el alistamiento de unidades navales [Rancán, 2004] han puesto de manifiesto que la definición de cómo se puede extraer conocimiento de las bases de datos de

registros de eventos e integrarlo a la base de conocimiento del sistema de ayuda es un problema abierto. En [Rancan *et al*, 2007] se propone un método de descubrimiento de conocimiento basado en agrupamiento e inducción de reglas en el marco de una propuesta de integración de sistemas de descubrimiento de conocimiento y sistemas basados en conocimiento.
En este contexto, resulta de interés el estudio de la integración de los algoritmos de inducción y agrupamiento al ir variando los parámetros que caracterizan el dominio en condiciones de laboratorio. Adicionalmente se intenta valorar intuitivamente la calidad de las reglas obtenidas y la degradación de dicha calidad como consecuencia de la variación de los parámetros controlados en los experimentos.

3. Algunos experimentos

El mejoramiento de una Base de Conocimiento con piezas de conocimiento descubiertas automáticamente, puede conducir a una degradación de la Base de Conocimiento original, por lo que es necesario explorar (al menos de forma teórica) cuáles son las curvas de degradación de la calidad de proceso de descubrimiento de conocimiento identificando las condiciones de borde para el modelo dentro del marco teórico desarrollado. Con el objetivo de realizar esta tarea, se ha llevado adelante un experimento que puede dividirse en tres pasos.
El primer paso consiste en la preparación del experimento. Este paso involucra: [a] generación del dominio basado en: generación de las clases y reglas que indican la pertenencia a cada clase y [b] generación de muestras para cada regla. Como salida de este paso se obtiene un conjunto de reglas de clasificación y un conjunto de muestras del dominio (ejemplos). El paso 2 consiste en la ejecución del experimento. Este paso involucra: [a] aplicar el proceso de agrupamiento al conjunto de muestras del dominio (ejemplos) para obtener el conjunto de sus clusters (grupos) y [b] aplicar a cada cluster el proceso de inducción para obtener reglas que caractericen la pertenencia a dicho cluster, obteniendo así el conjunto de reglas descubiertas. El paso 3 consiste en la comparación entre el conjunto de reglas de clasificación del paso 1 y las reglas descubiertas en el paso 2. El porcentaje de reglas descubiertas de forma correcta, define el éxito del experimento.

3.1. Variables

La experimentación hace uso de las siguientes variables independientes: [a] *attributes number*: cantidad de atributos en cada regla de clasificación (la misma en las muestras), [b] *rules per class*: cantidad de reglas de clasificación que indican la pertenencia a cada clase. [c] *class possible values*: cantidad de clases que rigen el dominio, [d] *attributes possible values*: cantidad de posibles valores que puede tomar cada atributo; y de la siguiente variable dependiente: [e] *rules correctly covered*: porcentaje de reglas pertenecientes al conjunto de reglas original que se encontró en el conjunto de reglas descubiertas.

3.2. Resultados

Los experimentos exploran el comportamiento del proceso propuesto al hacer variar una de las variables independientes sobre dominios en los que una segunda variable independiente toma ciertos valores discretos. Los resultados de los experimentos se muestran en las figuras 1 a 10.

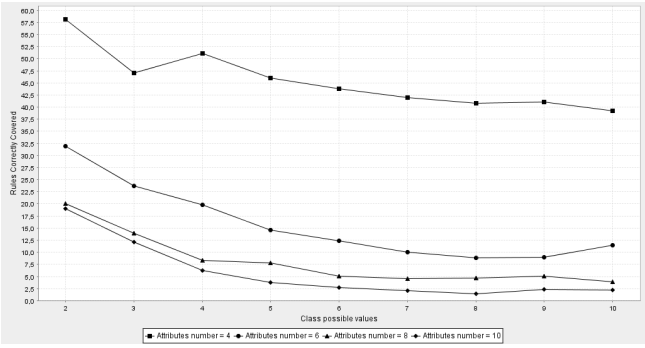

Fig. 1. Estudio de dominios variando la cantidad de clases que los rigen, para distinta cantidad de atributos que conforman las reglas

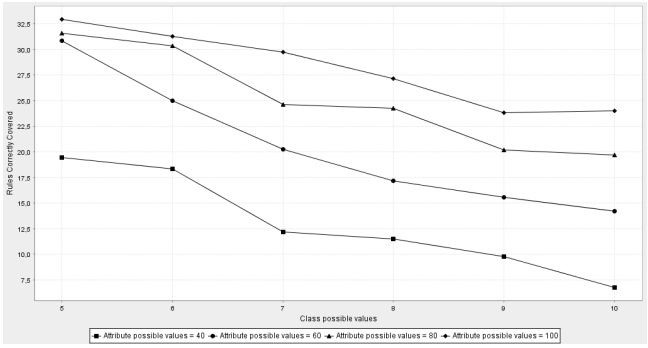

Fig. 2. Estudio de dominios variando la cantidad de clases que los rigen, para distinta cantidad de valores posibles que pueden tomar los atributos que conforman las reglas

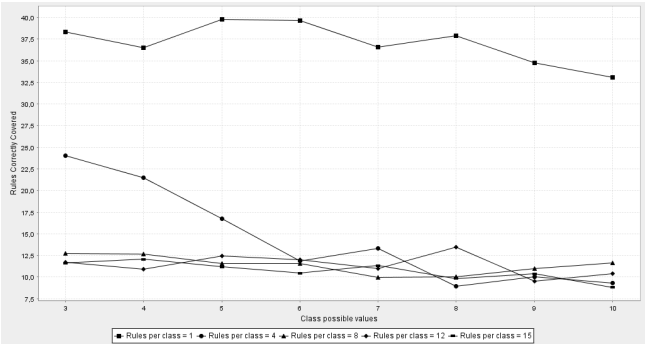

Fig. 3. Estudio de dominios variando la cantidad de clases que los rigen, para distinta cantidad de reglas que indican la pertenencia a cada clase

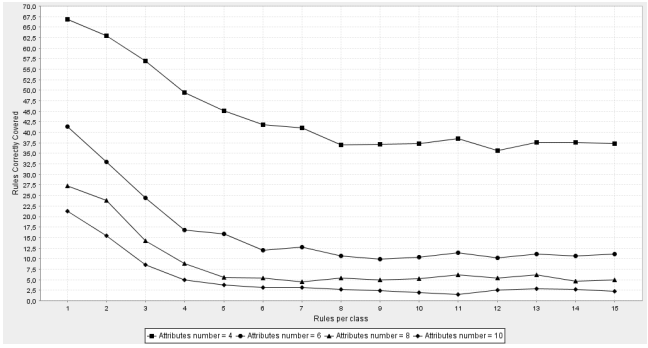

Fig. 4. Estudio de dominios variando la cantidad de reglas que indican la pertenencia a cada clase, para distinta cantidad de atributos que conforman las reglas

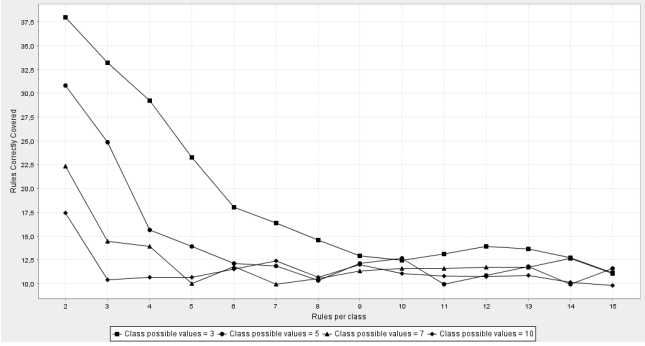

Fig. 5. Estudio de dominios variando la cantidad de reglas que indican la pertenencia a cada clase, para distinta cantidad de clases que rigen los dominios

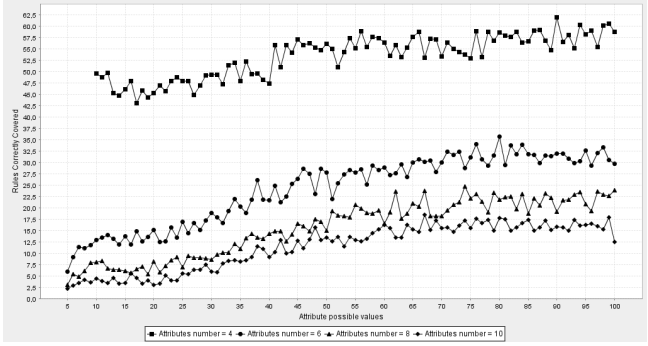

Fig. 6. Estudio de dominios variando la cantidad de valores posibles que puede tomar cada uno de los atributos que conforman las reglas, para distinta cantidad de atributos que conforman las reglas

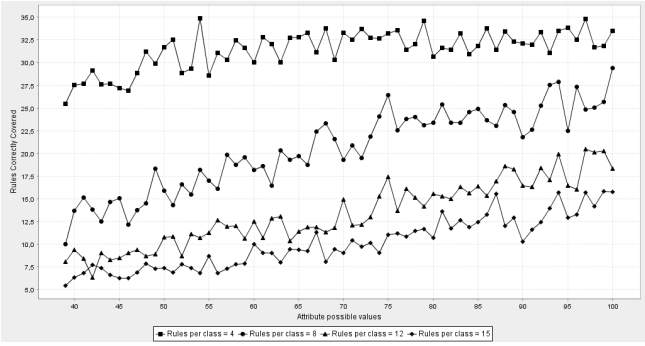

Fig. 7. Estudio de dominios variando la cantidad de valores posibles que puede tomar cada uno de los atributos que conforman las reglas, para distinta cantidad de reglas que indican la pertenencia a cada clase

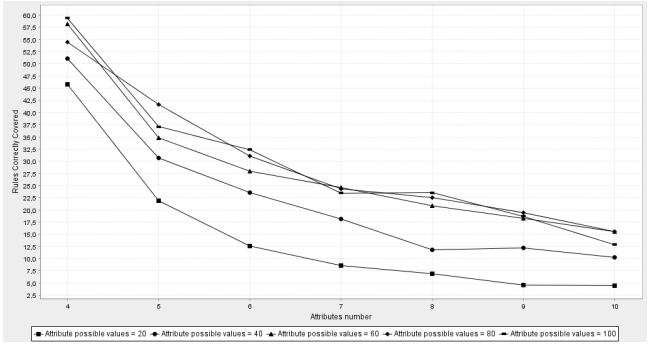

Fig. 8. Estudio de dominios variando la cantidad de atributos que conforman las reglas, para distinta cantidad de valores posibles que pueden tomar cada uno de estos atributos

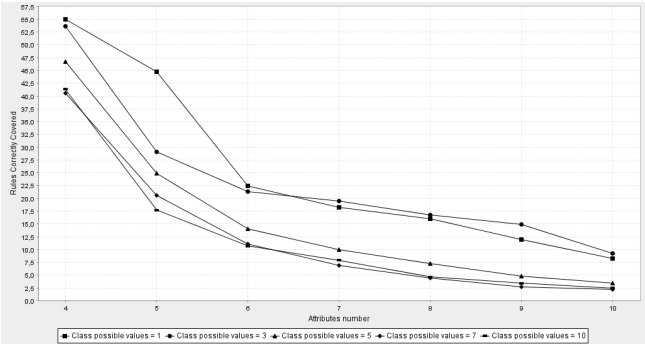

Fig. 9. Estudio de dominios variando la cantidad de atributos que conforman las reglas, para distinta cantidad clases que rigen los dominios

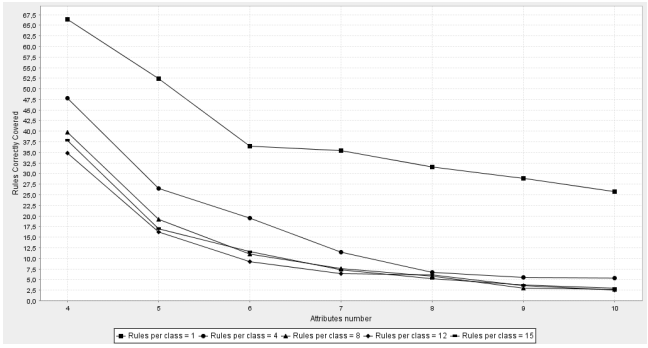

Fig. 10. Estudio de dominios variando la cantidad de atributos que conforman las reglas, para distinta cantidad de reglas que indican la pertenencia a cada clase

3.3. Interpretaciones

De los resultados experimentales se pueden extraer las siguientes proposiciones experimentales:

1. A mayor cantidad de clases que rigen el dominio, menor es la efectividad del método propuesto (figuras 1, 2, 3, 5, 9).
2. A mayor número de atributos que conforman las reglas que indican la pertenencia a cada clase, menor es la efectividad del método propuesto (figuras 1, 4, 6, 8, 9, 10).
3. A mayor cantidad de posibles valores que puede tomar cada uno de los atributos que componen las reglas, mayor es la efectividad del método propuesto (figuras 2, 6, 7, 8).
4. A mayor cantidad de reglas que indican la pertenencia a cada clase, menor es la efectividad del método propuesto (figuras 3, 4, 5, 7, 10).
5. Para un número alto de reglas que indican la pertenencia a cada clase, la efectividad del método propuesto parecería ser asintótica hacia un mínimo (figuras 4, 5).
6. A partir de determinada cantidad de posibles valores que puede tomar cada atributo, si se sigue aumentando esta cantidad, no se observan mejoras significativas en la efectividad del método propuesto (figura 8).
7. A partir de determinada cantidad de clases que rigen el dominio, el aumento de este parámetro no genera un decremento mayor en la efectividad del método (figura 9).

4. Conclusiones

Los procesos de descubrimiento de conocimiento deben lidiar con la incertidumbre vinculada a la medida de la calidad del conocimiento educido. En este contexto, la línea de trabajo en la cual se enmarca este proyecto se orienta a obtener resultados experimentales, los cuales, mediante la abducción, permiten inferir las características del dominio y en consecuencia una estimación de la calidad del conocimiento obtenido.

5. Formación de recursos humanos

En la línea de investigación cuyos resultados parciales se reportan en esta comunicación, se encuentran trabajando: un tesista de doctorado en ciencias informáticas, un tesista de grado en ingeniería informática y tres investigadores formados.

6. Bibliografía

Cogliati, M., Britos, P. García-Martínez, R. 2006. *Patterns in Temporal Series of Meteorological Variables Using SOM & TDIDT*. Lecture Notes in Artificial Intelligence (por aparecer). Springer Verlag.

Felgaer, P., Britos, P. y García-Martínez, R. 2006. *Prediction in Health Domain Using Bayesian Network Optimization Based on Induction Learning Techniques*. International Journal of Modern Physics C, 17(3): 447-455. ISSN 0129-1831.

Grosser, H., Britos, P. y García-Martínez, R. 2005. *Detecting Fraud in Mobile Telephony Using Neural Networks*. Lecture Notes in Artificial Intelligence, 3533: 613-615. Springer-Verlag.

Grossman, R., Kasif, S., Moore, R., Rocke, D. and Ullman, J. 1999. *Data Mining Research: Opportunities and Challenges* , A Report of three NSF Workshops on Mining Large, Massive, and Distributed Data, January 1999, Chicago

Hall, M. y Holmes, G. 2003. *Benchmarking Attribute Selection Techniques for Discrete Class Data Mining* , IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 6, pp. 1437-1447.

Holsheimer, M., Siebes, A. (1991). *Data Mining: The Search for Knowledge in Databases.* Report CS-R9406, ISSN 0169-118X, Amersterdam, The Netherlands.

Ierache, J. y Garcia-Martinez, R. 2004. *Sistema Experto Aplicado al Control del Espacio Aéreo.* Proceedings del IX Congreso Argentino de Ciencias de la Computación.

Kaski, S. 1997. *Data exploration using self-organizing maps.* Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82, 57 pp. Published by the Finnish Academy of Technology. ISSN 1238-9803.

Kohonen, T. (1982) *Self-organized formation of topologically correct feature maps.* Biological Cybernetics, 43:59-69.

Kohonen, T. (1990) *The Self-Organizing Map.* Proceedings of the IEEE, 78:1464-1480.

Kohonen, T. (1995a) *The adaptive-subspace SOM (ASSOM) and its use for the implementation of invariant feature detection.* In Fogelman-Soulié, F. and Gallinari, P., editors, Proceedings of ICANN'95, International Conference on Artificial Neural Networks, volume 1, pages 3-10. EC2 & Cie, Paris.

Kohonen, T. (1995c) *Self-Organizing Maps.* Springer, Berlin.

Kohonen, T., Oja, E., Simula, O., Visa, A., and Kangas, J. (1996b). *Engineering applications of the self-organizing map.* Proceedings of the IEEE, 84:1358-1384.

Michalski, R. Bratko, I. Kubat, M (eds.) 1998. *Machine Learning and Data Mining, Methods and Applications* , John Wiley & Sons Ltd, West Sussex, England

Piatetski-Shapiro, G., Frawley, W.J., Matheus, C.J. 1991. *Knowledge discovery in databases: an overview.* AAAI-MIT Press, Menlo Park, California.

Quinlan, J. R. 1986. *Induction of Decision Trees, Machine Learning* ,1:81-106

Rancán, C. 2004. *Arquitectura de Sistema Híbrido de Evaluación del Alistamiento de Unidades Navales Auxiliares*. Reportes Técnicos en Ingeniería del Software. 6(1): 45-54. ISSN 1667-5002.

Rancán, C., Pesado, P. y García-Martínez, R. (2007). *Toward Integration of Knowledge Based Systems and Knowledge Discovery Systems*. Journal of Computer Science & Technology, 7(1): 91-97.

Servente, M. y García Martínez, R. 2002. *Algoritmos TDIDT Aplicados a la Minería Inteligente*. Revista del Instituto Tecnológico de Buenos Aires, 26: 39-57. ISSN 0326-1840.

Sierra, E., García-Martínez, R., Hossian, A., Britos, P. y Balbuena, E. 2006. *Providing Intelligent User-Adapted Control Strategies in Building Environments*. Research in Computing Science Journal, 19: 235-241.