Un Protocolo de Caracterización Empírica de Dominios para Uso en Explotación de Información

Lopez-Nocera, M., Pollo-Cattaneo, F., Britos, P., García-Martínez, R.

Grupo Investigación en Sistemas de Información. Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús.
Grupo de Estudio en Metodologías de Ingeniería de Software. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional.
Grupo de Investigación en Explotación de Información. Sede Andina (El Bolsón). Universidad Nacional de Río Negro.
zappapet@yahoo.com, fpollo@posgrado.frba.utn.edu.ar, paobritos@gmail.com, rgarcia@unla.edu.ar

Abstract. En este trabajo se define un protocolo para encontrar las características del dominio que mejor es explicado por los algoritmos de descubrimiento de conocimiento seleccionados con base en procesos de explotación de información. Se da una prueba de concepto del protocolo propuesto y se dan algunas conclusiones parciales del trabajo de investigación realizado.

1. Introducción

Una de las hipótesis de trabajo subyacente al área de descubrimiento del conocimiento en procesos de explotación de información es que el funcionamiento de los algoritmos de minería de datos basados en aprendizaje automático [1] es independiente de las características del dominio que dichos datos modelan [2]. La experiencia de campo en diversos dominios [3, 4, 5, 6, 7]; muestra empíricamente que los mismos algoritmos pueden comportarse de diversa manera, dependiendo de las características del dominio a cuyos datos se estén aplicando.
En [8] se proponen cinco procesos de explotación de información: descubrimiento de reglas de comportamiento descubrimiento de grupos, ponderación de interdependencia de atributos, descubrimiento de reglas de pertenencia a grupos, y ponderación de reglas de comportamiento o de la pertenencia a grupos.
Dado que cuando no existen presunciones sobre posibles clases en el dominio el proceso estándar a utilizar es el de descubrimiento de reglas de pertenencia a grupos [9], surge el interés de estudiar los distintos algoritmos aplicables a este proceso y, de ellos, cuál es la mejor combinación en función de las características específicas del dominio.
En este contexto, en este trabajo se propone una caracterización de dominios (sección 2), se presenta un procedimiento para la generación de dominios (sección 3), se define un protocolo para encontrar las características del dominio que mejor es explicado por los algoritmos seleccionados (sección 4), se da una prueba de concepto del protocolo

propuesto (sección 5), y se ofrecen conclusiones del trabajo de investigación realizado (sección 6).

2. Caracterización de Dominios

En trabajos previos se han definido los dominios en términos de los ejemplos que los describen y de las reglas que cubren dichos ejemplos [10, 11, 12]. Las dimensiones utilizadas para esta descripción son: cantidad de atributos de cada ejemplo, cantidad de posibles valores que puede tomar cada atributo, cantidad de clases implícitas en los ejemplos, cantidad de ejemplos cubiertos por cada regla, cantidad de reglas que definen la pertenencia a cada clase, y porcentaje de reglas correctamente cubiertas si por camino inverso se indujesen a partir de los ejemplos. Esta última dimensión es dependiente de la cinco precedentes.
En este contexto, se puede postular una clasificación por complejidad de los dominios en: simples, medianos, oscilantes, complejos e hipercomplejos.

Dominios Simples: el aumento de la cantidad de ejemplos por regla, mejora el cubrimiento de reglas (independientemente de las demás dimensiones utilizadas, con lo cual los dominios en los que interviene con papel principal esta variable serán caracterizados como de *complejidad simple* o *trivial*, o directamente como *simples*).

Dominios Medianos: son aquellos con ejemplos con pocos atributos y pocas clases, ó pocos atributos y muchas clases ó pocas clases y pocas reglas por clase.

Dominios Oscilantes: son aquellos con ejemplos donde pueden variar el número de atributos por ejemplo, ó cantidad de ejemplos soportados por una regla, o valores comunes de atributos en un conjunto de ejemplos cubiertos por la misma regla.

Dominios Complejos: son aquellos con ejemplos con pocos atributos y muchos valores posibles por atributo, ó con muchos atributos y pocos valores posibles por atributo, ó con muchos atributos y muchos valores posibles por atributo.

Dominios Hipercomplejos: son aquellos con ejemplos donde pueden variar la cantidad de posibles distintos valores que pueden tomar los atributos, el número de atributos que cubren ejemplos, la cantidad de las reglas que cubren ejemplos, ó la cantidad de clases que cubren los ejemplos, ó la cantidad de reglas por clase.

3. Generación de Dominios

La generación de dominios para su caracterización empírica se realiza mediante el banco de pruebas desarrollado al efecto en [13]. El banco de pruebas genera experimentos a partir del siguiente procedimiento que se describe a continuación:

Paso 1: Preparación del Experimento. Como salida de este paso se obtiene un conjunto de reglas de clasificación y un conjunto de ejemplos del dominio que dan soporte a éstas.

Subpaso 2.1. Generación del dominio basada en la generación de clases y reglas que indican la pertenencia a cada una de éstas.

Subpaso 2.1. Generación de ejemplos que den soporte a cada una de las reglas de clasificación.

Paso 2: Ejecución del Experimento. Como salida de este paso se obtiene el conjunto de reglas descubiertas.

Subpaso 2.1. Aplicación del proceso de agrupamiento al conjunto de ejemplos del dominio para obtener el conjunto de sus clusters (grupos).

Subpaso 2.1. Aplicación a cada cluster del proceso de inducción para obtener reglas que caractericen la pertenencia a dicho cluster.

Paso 3: 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.

4. Protocolo de Identificación de Dominio

En esta sección se describe el protocolo de identificación de dominio que es mejor explicado por los algoritmos seleccionados. El protocolo se describe en términos de los siguientes pasos:

Paso 1: Elegir un algoritmo de clustering, sea K, y un algoritmo de inducción, sea I.

Paso 2: Repetir N veces:

SubP 2.1. Generar un dominio (Ejemplos y Reglas) para cada una de las combinaciones de valores posibles de las variables independientes predefinidas por el experimentador.

SP 2.1.1. Para cada dominio generado:

SP 2.1.1.1. Obtener una partición del conjunto de ejemplos del dominio en grupos

aplicando el algoritmo K a los ejemplos del dominio.

SP 2.1.1.2. Obtener reglas que caractericen la pertenencia a los grupos encontrados utilizando el algoritmo I.

SP 2.1.1.3. Obtener el porcentaje de reglas correctamente cubiertas por comparación del conjunto de reglas obtenidas en SP 2.1.1.2 con el conjunto de reglas obtenido en SubP 2.1.

SubP 2.2. Registrar los datos obtenidos (cantidad de atributos de cada ejemplo, cantidad de posibles valores que pueden tomar los atributos, cantidad de clases que rigen los ejemplos, cantidad de reglas que indican la pertenencia a cada clase, cantidad de ejemplos utilizados para cada regla, porcentaje de reglas correctamente cubiertas).

Paso 3: Obtener el Diagrama de Kiviat para los datos obtenidos.

Paso 4: Identificar usando Diagramas de Kiviat el tipo de dominio en el que la combinación (K, I) mejor descubre conocimiento.

5. Prueba de Concepto

Para la prueba de concepto se generó una muestra de los cien grupos de dominios con los correspondientes conjuntos de reglas y de ejemplos. Cada grupo se formó con todos los dominios surgidos de las combinaciones de los posibles valores de las variables independientes: cantidad de atributos de cada ejemplo, cantidad de posibles valores que pueden tomar los atributos, cantidad de clases que rigen los ejemplos, cantidad de reglas que indican la pertenencia a cada clase, y cantidad de ejemplos utilizados para cada regla.

A cada dominio generado se le aplicó el proceso de descubrimiento de reglas de pertenencia a grupos [9] utilizando el algoritmo SOM [14] como algoritmo de clustering e ID3 [15] de la familia TDIDT como algoritmo de inducción.

Se analizaron los resultados obtenidos, y se verificó el porcentaje de reglas correctamente cubiertas para cada una de las combinaciones posibles de variables, las que se muestran en la Tabla 1. Se utilizaron Diagramas de Kiviat [16] para graficar los resultados, como el que se muestra en la Figura 1.

Se observa el mejor porcentaje de cubrimiento de reglas (superior al 95%) para los casos 2 y 3 de la muestra, es decir, para aquellos dominios con menor cantidad de reglas por clase, manteniendo fijas las otras variables. En el diagrama de Kiviat de la figura 1, esto se graficó como zona “A”. Estos dominios, según la clasificación antedicha, son *medianos*. Por otra parte, dicho cubrimiento mejora, a su vez, si se aumenta adicionalmente la cantidad de ejemplos utilizados para cada clase y se mantiene constante el resto de las variables, incluyendo reglas por clase, lo cual nos pondría en presencia de dominios *simples*. Esto implica, en forma directamente

proporcional, un decrecimiento en el porcentaje de reglas cubiertas cuando la cantidad de ejemplos utilizada disminuye, como se puede apreciar en el diagrama para otros dominios catalogados como *medianos*, verbigracia el caso 1, también perteneciente a la zona "A" en la figura 1.

Por otra parte, se observa un menor pero muy importante cubrimiento (en torno al 60%) para los casos 11, 12, 19, 20 y 21 de la muestra, que corresponden a dominios con ejemplos donde pueden variar la cantidad de posibles distintos valores que pueden tomar los atributos, el número de atributos que cubren ejemplos, la cantidad de las reglas que cubren ejemplos, ó la cantidad de clases que cubren los ejemplos, ó la cantidad de reglas por clase, manteniendo fijas las otras variables. En el diagrama de Kiviat de la figura 1, estos dominios corresponden a las zonas "B". Según la clasificación brindada antes, estos dominios se catalogan como *hipercomplejos*. Una vez más, dicho cubrimiento mejora, a su vez, si se aumenta adicionalmente la cantidad de ejemplos utilizados para cada regla y se mantiene constante el resto de las variables, incluyendo la cantidad de clases que rigen los ejemplos, lo cual nos pondría nuevamente en presencia de dominios *simples* y, viceversa, decrecerá en caso contrario, como se aprecia con el caso 10, correspondiente también a una de las zonas "B" en la Figura 1.

Continuando con el análisis del diagrama, se observan zonas de cubrimiento discreto, con porcentajes que rondan el 40%, para los casos 4, 13, 17, 18, 22, 23, 24, 25, 26 y 27, que corresponden a ejemplos con pocos atributos y muchos valores posibles por atributo, ó con muchos atributos y pocos valores posibles por atributo, ó con muchos atributos y muchos valores posibles por atributo, lo que en la clasificación anterior se catalogó como un dominio del tipo *complejo*. Ésta corresponde a las zonas "C" que se ven en la Figura 1. Nuevamente, como sucede con todos los casos, dicho cubrimiento mejorará, si se aumenta adicionalmente la cantidad de ejemplos utilizados para cada regla y se mantiene constante el resto de las variables, incluyendo la cantidad de atributos de cada ejemplo y la cantidad de posibles valores que pueden tomar los atributos que rigen los ejemplos, lo cual nos pondría nuevamente en presencia de dominios *simples* y, viceversa, decrecerá en caso contrario. Esto puede apreciarse gráficamente, sobre el diagrama, puntualmente para los casos 5, 6, 14, 15 y 16, asimismo correspondientes también a las zonas "C", según lo graficado en Figura 1.

Finalmente, para la zona que corresponde a casos donde pueden variar el número de atributos por ejemplo cantidad de ejemplos soportados por una regla, o valores comunes de atributos en un conjunto de ejemplos cubiertos por la misma regla, es decir – de acuerdo con la clasificación anteriormente dada – para dominios *oscilantes*, como son los casos 7 y 8, se observa un cubrimiento inferior al 40%, más cercano al 30%, manteniéndose la salvedad de que dicho cubrimiento mejorará, si se aumenta adicionalmente la cantidad de ejemplos utilizados para cada regla y se mantiene constante el resto de las variables, incluyendo la cantidad de atributos de cada ejemplo y la cantidad de ejemplos soportados por una regla, lo cual nos pondría nuevamente en presencia de dominios *simples* y, viceversa, decrecerá en caso contrario, como puede apreciarse en el diagrama puntualmente para el caso 9. Todos estos casos se catalogaron como zona "D" en la Figura 1.

Tabla 1. Porcentaje de reglas correctamente cubiertas para combinación de variables

CANTIDAD DE ATRIBUTOS DE CADA EJEMPLO	CANTIDAD DE POSIBLES VALORES QUE PUEDEN TOMAR LOS ATRIBUTOS	CANTIDAD DE CLASES QUE RIGEN LOS EJEMPLOS	CANTIDAD DE REGLAS QUE INDICAN LA PERTENENCIA A CADA CLASE	CANTIDAD DE EJEMPLOS UTILIZADOS PARA CADA REGLA	PORCENTAJE DE REGLAS CORRECTAMENTE CUBIERTAS
3	1	1	1	1	**65**
3	1	1	1	10	**96**
3	1	1	1	20	**98**
3	1	1	5	1	**40**
3	1	1	5	10	**48**
3	1	1	5	20	**62**
3	1	1	15	1	**31**
3	1	1	15	10	**38**
3	1	1	15	20	**45**
3	1	5	1	1	**59**
3	1	5	1	10	**67**
3	1	5	1	20	**73**
3	1	5	5	1	**39**
3	1	5	5	10	**49**
3	1	5	5	20	**56**
3	1	5	15	1	**34**
3	1	5	15	10	**38**
3	1	5	15	20	**41**
3	1	10	1	1	**58**
3	1	10	1	10	**62**
3	1	10	1	20	**68**
3	1	10	5	1	**37**
3	1	10	5	10	**43**
3	1	10	5	20	**47**
3	1	10	15	1	**35**
3	1	10	15	10	**38**
3	1	10	15	20	**42**

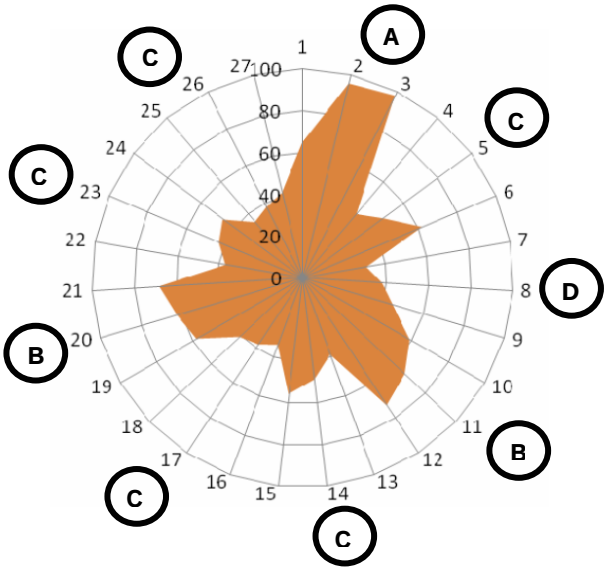

Fig. 1. Diagrama de Kiviat de los datos de la Tabla 1

En la Figura 2 puede apreciarse un gráfico comparativo del porcentual de cubrimiento de las reglas del dominio logrado por las reglas inferidas por uso de SOM e ID3 combinados que surgen de los datos de la Tabla 1 en la que A corresponde a dominios medianos, B a dominios hipercomplejos, C a dominios complejos y D = dominios simples.

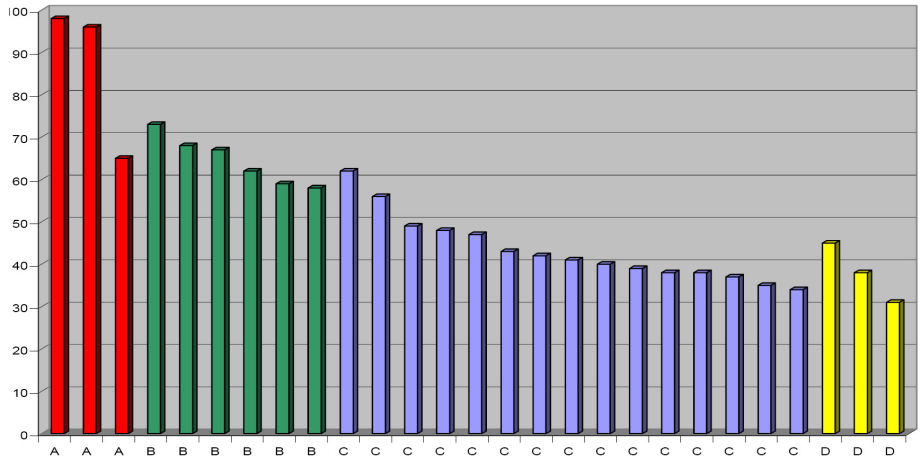

Fig. 2. Gráfico comparativo del porcentual de cubrimiento de las reglas del dominio logrado por las reglas inferidas por uso de SOM e ID3 (A = dominios medianos; B = dominios hipercomplejos; C = dominios complejos; D = dominios simples)

6. Conclusiones

Se determinó experimentalmente que las características del dominio elegido tienen influencia sobre el resultado experimental obtenido. Se observa que para combinación de algoritmos de clustering e inducción (SOM+ID3) elegida para la prueba de concepto, el mejor cubrimiento de reglas se obtiene para dominios catalogados como *medianos*, en particular para el escenario que presenta aquellos dominios con la menor cantidad posible de cada una de las variables y el mayor número utilizado de ejemplos por regla.

Se observa también que los valores obtenidos experimentalmente se ajustan a lo que se postuló a priori en la clasificación teórica realizada a los distintos dominios según su complejidad.

Como futura línea de investigación se propone aplicar este estudio a otras combinaciones de algoritmos de clustering con algoritmos de inducción, entre los que se consideran: SOM+AQ15, SOM+CN2, SOM+M5, K-means+AQ15, K-means+CN2, K-means+M5, K-means+ID3, NNC+AQ15, NNC+CN2, NNC+M5, NNC+ID3, entre otros.

7. Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por el Proyecto de Investigación 33A105 del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús, por el Proyecto de Investigación 40B065 de la Universidad Nacional de Rio Negro - Sede Andina (El Bolsón), y por el Proyecto 25C126 de la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional.

8. Referencias

1. Sammut, C., Webb, G. 2011. *Encyclopedia of Machine Learning*. Springer. ISBN 978-0-387-30768-8.
2. Cios, K., Pedrycz, W., Swiniarski, R., Kurgan, L. 2010. *Data Mining: A Knowledge Discovery Approach*. Springer. ISBN 978-0-387-33333-5.
3. Grosser, H., Britos, P., García-Martínez, R. (2005) *Detecting Fraud in Mobile Telephony Using Neural Networks*.LNAI 3533:613-615.
4. Britos, P., Grosser, H., Rodríguez, D., Garcia-Martinez, R. 2008. *Detecting Unusual Changes of Users Consumption*. En *Artificial Intelligence and Practice II*, Max Bramer Ed. (Boston: Springer), IFIP Series, 276: 297-306.
5. Felgaer, P., Britos, P., and García-Martínez, R. (2006) *Prediction in Health Domain Using Bayesian Network Optimization Based on Induction Learning Techniques*. Int. J. of Mod. Ph. C 17(3): 447-455.
6. Cogliati, M., Britos, P., García-Martínez, R. (2006) *Patterns in Temporal Series of Meteorological Variables Using SOM & TDIDT* En: *Artificial Intelligence in Theory and Practice*, Bramer M (ed), Boston, Springer, IFIP Series 217:305-314
7. Valenga, F., Fernández, E., Merlino, H., Rodríguez, D., Procopio, C., Britos. P., García Martínez, R. (2008) *Minería de Datos Aplicada a la Detección de Patrones Delictivos en Argentina*. VII JIISIC´08: 31-39
8. Britos, P. 2008. *Procesos de Explotación de Información Basados en Sistemas Inteligentes*. Tesis de Doctorado en Ciencias Informáticas. Universidad Nacional La Plata. http://post grado.info.unlp.edu.ar/Carrera/Doctorado/Tesis/Britos-Tesis%20Doctoral.pdf Página vigente al 11/07/11.
9. Pollo-Cattaneo, F., Britos, P., Pesado, P., García-Martínez, R. 2010. *Ingeniería de Procesos de Explotación de Información*. En *Ingeniería de Software e Ingeniería del Conocimiento: Tendencias de Investigación e Innovación Tecnológica en Iberoamérica* (Editores: R. Aguilar, J. Díaz, G. Gómez, E- León). Pág. 252-263. Alfaomega Grupo Editor. ISBN 978-607-707-096-2.
10. 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. ISSN 1666-6038.
11. Rancan, C., Kogan, A., Pesado, P., García-Martínez, R. 2007. *Knowledge Discovery for Knowledge Based Systems. Some Experimental Results*. Research in Computing Science Journal, 27: 3-13. ISSN 1665-9899.
12. Rancan, C., Pesado, P., García-Martínez, R. 2010. *Issues in Rule Based Knowledge Discovering Process*. Advances and Applications in Statistical Sciences Journal, 2(2): 303-314. ISSN 0974-6811.
13. Kogan, A. 2007. *Integración de Algoritmos de Inducción y Agrupamiento. Estudio del Comportamiento*. Tesis de Ingeniería Informática. Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires.
14. Kohonen, T. 1995. *Self-Organizing Maps*. Springer Verlag Publishers.

15. Quinlan, J. 1986. *Induction of decision trees.* Machine Learning, 1(1): 81-106. ISSN 0885-6125.
16. Soman, K. P., Diwakar, S., Ajay, V. 2006. *Insight into Data Mining: Theory and Practice*, 2nd Edition, Prentice Hall, New Delhi, India.