Propuesta de Procesos de Explotación de Información

Paola Britos y Ramón García-Martínez

Area Informática. Sede Andina (El Bolsón). Universidad Nacional de Río Negro
Área Ingeniería del Software. Licenciatura en Sistemas. Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús.
paobritos@gmail.com, rgm1960@yahoo.com

Resumen. En este trabajo se caracterizan los procesos de explotación de información asociados a los problemas de inteligencia de negocio: descubrimiento de reglas de comportamiento, descubrimiento de grupos, descubrimiento de atributos significativos, descubrimiento de reglas de pertenencia a grupos y ponderación de reglas de comportamiento o de pertenencia a grupos. Se identifican las tecnologías de Sistemas Inteligentes (SI) que pueden utilizarse para los procesos caracterizados, validando estos procesos a través de casos aceptados por la comunidad internacional.

Palabras Claves. Procesos de explotación de información. Descubrimiento de grupos. Descubrimiento de atributos significativos. Descubrimiento de reglas. Ponderación de atributos Ponderación de reglas.

1. Introducción

La inteligencia de negocio propone un abordaje interdisciplinario (dentro del que se encuentra la Informática), que tomando todos los recursos de información disponibles y el uso de herramientas analíticas y de síntesis con capacidad de transformar la información en conocimiento, se centra en generar a partir de estos, conocimiento que contribuya con la toma de decisiones de gestión y generación de planes estratégicos en las organizaciones [1].

La Explotación de Información es la sub-disciplina Informática que aporta a la Inteligencia de Negocio [2] las herramientas para la transformación de información en conocimiento [3]. Se ha definido como la búsqueda de patrones interesantes y de regularidades importantes en grandes masas de información [4]. Al hablar de explotación de información basada en sistemas inteligentes [5] se refiere específicamente a la aplicación de métodos de sistemas inteligentes, para descubrir y enumerar patrones presentes en la información. Los métodos basados en sistemas inteligentes [6], permiten obtener resultados de análisis de la masa de información que los métodos convencionales [7] no logran tales como: los algoritmos TDIDT (Top Down Induction Decision Trees), los mapas auto organizados (SOM) y las redes bayesianas. Los algoritmos TDIDT permiten el desarrollo de descripciones simbólicas de los datos para diferenciar entre distintas clases [8]. Los mapas auto organizados pueden ser aplicados a la construcción de particiones de grandes masas de

información. Tienen la ventaja de ser tolerantes al ruido y la capacidad de extender la generalización al momento de necesitar manipular datos nuevos [9]. Las redes bayesianas pueden ser aplicadas para identificar atributos discriminantes en grandes masas de información, detectar patrones de comportamiento en análisis de series temporales. [10].

Se ha señalado la necesidad de disponer de procesos [11] que permitan obtener conocimiento [12] a partir de las grandes masas de información disponible [13], su caracterización [14] y tecnologías involucradas [15].

En este contexto en este trabajo se propone una caracterización de los procesos de explotación de información asociados a los problemas de inteligencia de negocio: descubrimiento de reglas de comportamiento, descubrimiento de grupos, descubrimiento de atributos significativos, descubrimiento de reglas de pertenencia a grupos y ponderación de reglas de comportamiento o de pertenencia a grupos y se identifican las tecnologías de SI que pueden utilizarse para los procesos caracterizados.

2. Propuesta de Técnicas para Procesos de Explotación de Información

En esta Sección se proponen los siguientes procesos de explotación de información: descubrimiento de reglas de comportamiento (Sección 2.1), descubrimiento de grupos (Sección 2.2), descubrimiento de atributos significativos (Sección 2.3), descubrimiento de reglas de pertenencia a grupos (Sección 2.4) y ponderación de reglas de comportamiento o de pertenencia (Sección 2.5).

2.1. Descubrimiento de Reglas de Comportamiento

El proceso de descubrimiento de reglas de comportamiento aplica cuando se requiere identificar cuales son las condiciones para obtener determinado resultado en el dominio del problema. Son ejemplos de problemas que requieren este proceso: identificación de características del local mas visitado por los clientes, identificación de factores que inciden en el alza las ventas de un producto dado, establecimiento de características o rasgos de los clientes con alto grado de fidelidad a la marca, establecimiento de atributos demográficos y psicográficos que distinguen a los visitantes de un website, entre otros.

Para el descubrimiento de reglas de comportamiento definidos a partir de atributos clases en un dominio de problema que representa la masa de información disponible, se propone la utilización de algoritmos de inducción TDIDT [16] para descubrir las reglas de comportamiento de cada atributos clase. Este proceso y sus subproductos pueden ser visualizados gráficamente en la Figura 1.

En primer lugar se identifican todas las fuentes de información (bases de datos, archivos planos, entre otras), se integran entre sí formando una sola fuente de información a la que se llamará datos integrados. Con base en los datos integrados se selecciona el atributo clase (atributo A en la Figura).

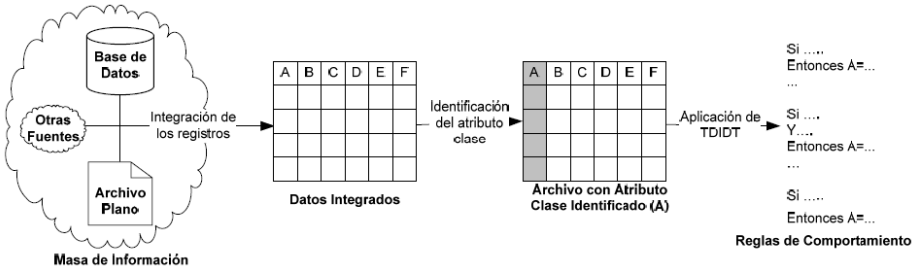

Fig. 1. Esquema y subproductos resultantes de aplicar TDIDT al descubrimiento de reglas de comportamiento

Como resultado de la aplicación del algoritmo de inducción TDIDT al atributo clase se obtiene un conjunto de reglas que definen el comportamiento de dicha clase.

2.2. Descubrimiento De Grupos

El proceso de descubrimiento de grupos aplica cuando se requiere identificar una partición en la masa de información disponible sobre el dominio de problema.

Son ejemplos de problemas que requieren este proceso: identificación de segmentos de clientes para bancos y financieras, identificación de tipos de llamadas de clientes para empresas de telecomunicación, identificación de grupos sociales con las mismas características, identificación de grupos de estudiantes con características homogéneas, entre otros.

Para el descubrimiento de grupos [17][18] a partir de masas de información del dominio de problema sobre las que no se dispone ningún criterio de agrupamiento "a priori" se propone la utilización de Mapas Auto Organizados de Kohonen o SOM por su sigla en inglés [19][20][21]. El uso de esta tecnología busca descubrir si existen grupos que permitan una partición representativa del dominio de problema que la masa de información disponible representa. Este proceso y sus subproductos pueden ser visualizados gráficamente en la Figura 2.

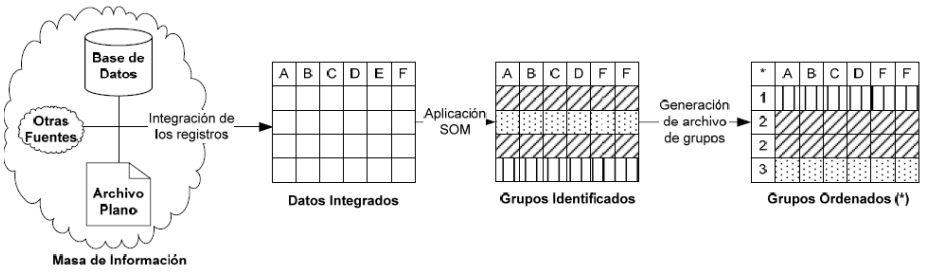

Fig. 2. Esquema y subproductos resultantes de aplicar SOM para el descubrimiento de grupos

En primer lugar se identifican todas las fuentes de información (bases de datos, archivos planos, entre otras), se integran entre sí formando una sola fuente de información a la que se llamará datos integrados. Con base en los datos integrados se

aplican mapas auto organizados (SOM). Como resultado de la aplicación de SOM se obtiene una partición del conjunto de registros en distintos grupos a los que se llamará grupos identificados. Para cada grupo identificado se generará el archivo correspondiente.

2.3. Ponderación de Interdependencia de Atributos

El proceso de ponderación de interdependencia de atributos aplica cuando se requiere identificar cuales son los factores con mayor incidencia (o frecuencia de ocurrencia) sobre un determinado resultado del problema.

Son ejemplos de problemas que requieren este proceso: factores con incidencia sobre las ventas, rasgos distintivos de clientes con alto grado de fidelidad a la marca, atributos claves que convierten en vendible a un determinado producto, características sobresalientes que tienen los visitantes de un website, entre otros.

Para ponderar en que medida la variación de los valores de un atributo incide sobre la variación del valor de un atributo clase se propone la utilización de Redes Bayesianas [22]. El uso de esta tecnología busca identificar si existe interdependencia en algún grado entre los atributos que modelan el dominio de problema que la masa de información disponible representa. Este proceso y sus subproductos pueden ser visualizados gráficamente en la Figura 3.

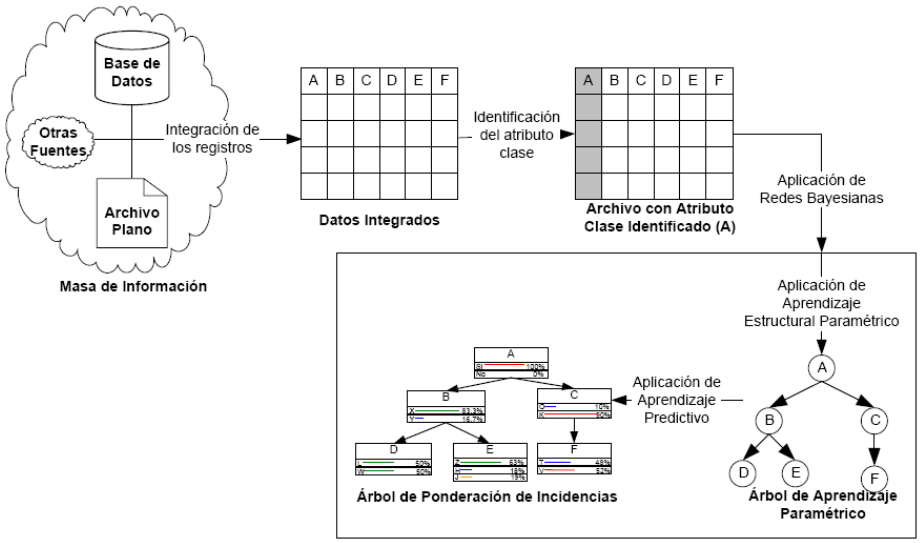

Fig. 3. Esquema y subproductos resultantes de aplicar Redes Bayesianas a la Ponderación de Interdependencia entre Atributos

En primer lugar se identifican todas las fuentes de información (bases de datos, archivos planos, entre otras), se integran entre sí formando una sola fuente de información a la que se llamará datos integrados. Con base en los datos integrados se selecciona el atributo clase (atributo A en la Figura).

Como resultado de la aplicación del aprendizaje estructural de las Redes Bayesianas al archivo con atributo clase identificado se obtiene el árbol de aprendizaje; a este se le aplica el aprendizaje predictivo Redes Bayesianas y se obtiene el árbol de ponderación de interdependencias que tiene como raíz al atributo clase y como nodos hojas a los otros atributos con la frecuencia (incidencia) sobre el atributo clase.

2.4. Descubrimiento de Reglas de Pertenencia a Grupos

El proceso de descubrimiento de reglas de pertenencia a grupos aplica cuando se requiere identificar cuales son las condiciones de pertenencia a cada una de las clases en una partición desconocida "a priori", pero presente en la masa de información disponible sobre el dominio de problema.

Son ejemplos de problemas que requieren este proceso: tipología de perfiles de clientes y caracterización de cada tipología, distribución y estructura de los datos de mi website, segmentación etaria de mis estudiantes y comportamiento de cada segmento, clases de llamadas telefónicas en una región y caracterización de cada clase, entre otros.

Para el descubrimiento de reglas de pertenencia a grupos se propone la utilización de mapas auto-organizados (SOM) para el hallazgo de los mismos y; una vez identificados los grupos, la utilización de algoritmos de inducción (TDIDT) para establecer las reglas de pertenencia a cada uno [23][24][21]. Este proceso y sus subproductos pueden ser visualizados gráficamente en la Figura 4.

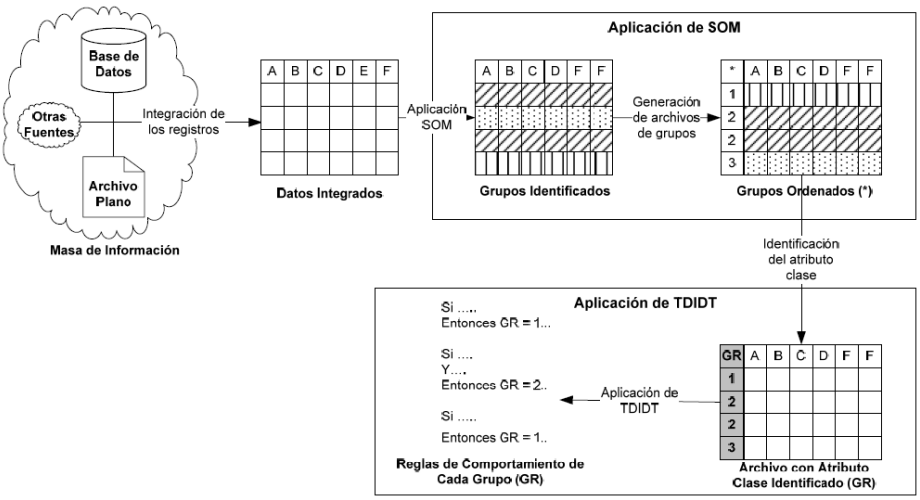

Fig. 4. Esquema y subproductos resultantes de SOM y TDIDT aplicados al descubrimiento de reglas de pertenencia a grupos

En primer lugar se identifican todas las fuentes de información (bases de datos, archivos planos, entre otras), se integran entre sí formando una sola fuente de información a la que se llamará datos integrados. Con base en los datos integrados se

aplican mapas auto-organizados (SOM). Como resultado de la aplicación de SOM se obtiene una partición del conjunto de registros en distintos grupos a los que se llama grupos identificados. Se generan los archivos asociados a cada grupo identificado. A este conjunto de archivos se lo llama grupos ordenados. El atributo "grupo" de cada grupo ordenado se identifica como el atributo clase de dicho grupo, constituyéndose este en un archivo con atributo clase identificado (GR). Se aplica el algoritmo de inducción TDIDT al atributo clase de cada grupo GR y se obtiene un conjunto de reglas que definen el comportamiento de cada grupo.

2.5. Ponderación de Reglas de Comportamiento o de Pertenencia a Grupos

El proceso de ponderación de reglas de comportamiento o de la pertenencia a grupos aplica cuando se requiere identificar cuales son las condiciones con mayor incidencia (o frecuencia de ocurrencia) sobre la obtención de un determinado resultado en el dominio del problema, sean estas las que en mayor medida inciden sobre un comportamiento o las que mejor definen la pertenencia a un grupo. Son ejemplos de problemas que requieren este proceso: identificación del factor dominante que incide en el alza las ventas de un producto dado, rasgo con mayor presencia en los clientes con alto grado de fidelidad a la marca, frecuencia de ocurrencia de cada perfil de de clientes, identificación del tipo de llamada mas frecuente en una región, entre otros.

Para la ponderación de reglas de comportamiento o de pertenencia a grupos se propone la utilización de redes bayesianas [22]. Esto puede hacerse a partir de dos procedimientos dependiendo de las características del problema a resolver: cuando no hay clases/grupos identificados; o cuando hay clases/grupos identificados.

El procedimiento a aplicar cuando hay clases/grupos identificados consiste en la utilización de algoritmos de inducción TDIDT [16] para descubrir las reglas de comportamiento de cada atributo clase y posteriormente se utiliza redes bayesianas para descubrir cual de los atributos establecidos como antecedentes de las reglas tiene mayor incidencia sobre el atributo establecido como consecuente. Este proceso y sus subproductos pueden ser visualizados gráficamente en la Figura 5.

En primer lugar se identifican todas las fuentes de información (bases de datos, archivos planos, entre otras), se integran entre sí formando una sola fuente de información a la que se llamará datos integrados. Con base en los datos integrados se selecciona el atributo clase (atributo A en la Figura 5). Como resultado de la aplicación del algoritmo de inducción TDIDT al atributo clase se obtiene un conjunto de reglas que definen el comportamiento de dicha clase. Seguidamente, se construye un archivo con los atributos antecedentes y consecuentes identificados por la aplicación del algoritmo TDIDT. Como resultado de la aplicación del aprendizaje estructural de las Redes Bayesianas al archivo con atributo clase obtenido por la utilización del algoritmo TDIDT (CL en la Figura 5), se obtiene el árbol de aprendizaje; a este se le aplica aprendizaje predictivo y se obtiene el árbol de ponderación de interdependencias que tiene como raíz al atributo clase (en este caso el atributo consecuente) y como nodos hojas a los atributos antecedentes con la frecuencia (incidencia) sobre el atributo consecuente.

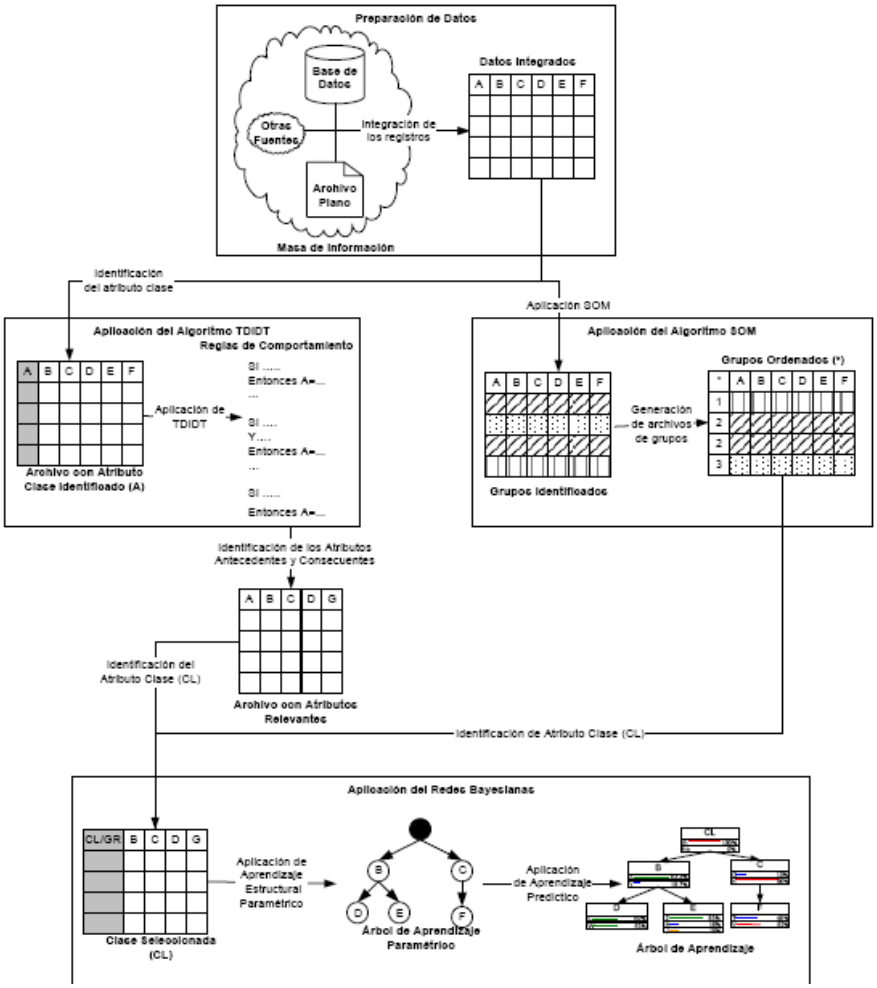

Fig. 5. Esquema y subproductos resultantes de redes bayesianas aplicadas a la ponderación de reglas de comportamiento o de pertenencia a grupos

El procedimiento a aplicar cuando no hay clases/grupos identificados consiste en identificar todas las fuentes de información (bases de datos, archivos planos, entre otras), se integran entre sí formando una sola fuente de información a la que se llamará datos integrados. Con base en los datos integrados se aplican mapas auto organizados (SOM). Como resultado de la aplicación de SOM se obtiene una partición del conjunto de registros en distintos grupos a los que se llamará grupos identificados. Para cada grupo identificado se generará el archivo correspondiente. A este conjunto de archivos se lo llama grupos ordenados. El atributo “grupo” de cada grupo ordenado se identifica como el atributo clase de dicho grupo, constituyéndose este en un archivo con atributo clase identificado (GR). Como resultado de la aplicación del aprendizaje estructural se obtiene el árbol de aprendizaje; a este se le aplica el aprendizaje predictivo y se obtiene el árbol de ponderación de

interdependencias que tiene como raíz al atributo grupo y como nodos hojas a los otros atributos con la frecuencia (incidencia) sobre el atributo grupo.

3. Validación de los Procesos de Explotación de Información Propuestos

Se validaron los procesos propuestos en tres dominios: alianzas políticas, diagnóstico médico y comportamiento de usuarios. Un detalle completo de estas validaciones puede verse en [26].

En el dominio de alianzas políticas se buscó descubrir comportamiento de los representantes demócratas y republicanos del Congreso de EE.UU en la agenda política de un período de sesiones ordinarias, identificando acuerdos y desacuerdos intrapartidarios y acuerdos entre grupos interpartidarios y entre minorías intrapartidarias. Lo primero se buscó mediante el proceso de descubrimiento de reglas de comportamiento de los representantes de cada partido y lo segundo mediante el descubrimiento de grupos de representantes que hayan votado homogéneamente (con independencia de su partido de filiación) y de las reglas que definen esa homogeneidad (reglas de pertenencia a cada grupo). Adicionalmente se buscó identificar cual ha sido la ley o leyes con mayor acuerdo dentro de los acuerdos identificados, utilizando el proceso de ponderación de reglas de comportamiento o de reglas de pertenencia a grupos.

En el dominio de diagnostico medico se buscó sintetizar el conocimiento que permite diagnosticar el tipo de linfoma a partir de determinadas características observadas en la linfografía asociada, cual es la característica o características determinantes de dicha observación para cada tipo de diagnóstico y si existen características comunes a diferentes tipos de patologías. Lo primero se buscó mediante el proceso de descubrimiento de reglas de comportamiento de los diagnósticos de cada tipo, lo segundo utilizando el proceso de ponderación de reglas de comportamiento y lo tercero mediante el descubrimiento de grupos de linfomas con características homogéneas (con independencia de la tipología) y de las reglas que definen esa homogeneidad (reglas de pertenencia a cada grupo).

En el dominio de comportamiento de usuarios se buscó dar una a descripción de las causales de alta o baja de un servicio “dial-up” de Internet provista por una compañía telefónica e identificar las causales con mayor incidencia en cada comportamiento. Lo primero se buscará mediante el proceso de descubrimiento de reglas de comportamiento de alta y baja del servicio y lo segundo mediante el proceso de ponderación de reglas de comportamiento.

4. Conclusiones

En este trabajo se proponen y describen cinco procesos de explotación de información: descubrimiento de reglas de comportamiento, descubrimiento de grupos, descubrimiento de atributos significativos, descubrimiento de reglas de pertenencia a grupos y ponderación de reglas de comportamiento o de pertenencia a grupos.

Se han asociado a cada proceso las siguientes técnicas: el uso de algoritmos TDIDT aplicados al descubrimiento de reglas de comportamiento ó reglas de pertenencia a grupos, el uso de los mapas auto organizados aplicados al descubrimiento de grupos, el uso de las redes bayesianas aplicados a la ponderación de interdependencia entre atributos, el uso de los mapas auto organizados y algoritmos TDIDT aplicados al descubrimiento de reglas de pertenencia a grupos y el uso de redes bayesianas aplicados a la ponderación de reglas de comportamiento o reglas de pertenencia a grupos.

Durante el trabajo de investigación documental se observó el uso indistinto de minería de datos y de explotación de información para referirse al mismo cuerpo de conocimiento. Sin embargo, plantear esta equivalencia es similar a plantear la equivalencia entre los sistemas informáticos y los sistemas de información. Los primeros describen la tecnología que dan soporte a los segundos y esto es lo que los hace distintos. En este contexto surge como problema abierto de interés la necesidad de un ordenamiento en el cuerpo de conocimiento en formación discriminado cuales son los procesos y las metodologías que pertenecen al campo de la explotación de información y cuales son las tecnologías de minería de datos que dan soporte a dichos procesos y metodologías. Por otra parte, en la literatura abundan los trabajos y resultados sobre la conveniencia de uso de determinados algoritmos de minería de datos frente a otros, sin embargo rara vez se plantea el proceso de explotación de información al cual estos algoritmos están asociados o la conveniencia del uso de uno algoritmo frente a otros en dicho proceso. En este contexto surge como problema abierto de interés la identificación de la correspondencia entre algoritmo de minería de datos y proceso de explotación de información.

4. Referencias

1. Thomsen, E. (2003). *BI's Promised Land*. Intelligent Enterprise, 6(4): 21-25.
2. Negash, S., Gray, P. (2008). *Business Intelligence*. En Handbook on Decision Support Systems 2, ed. F. Burstein y C. Holsapple (Heidelberg, Springer), Pág. 175-193.
3. Langseth, J., Vivatrat, N. (2003). *Why Proactive Business Intelligence is a Hallmark of the Real-Time Enterprise: Outward Bound*. Intelligent Enterprise 5(18): 34-41.
4. Grigori, D., Casati, F., Castellanos, M., Dayal, u., Sayal, M., Shan, M. (2004). *Business Process Intelligence*. Computers in Industry 53(3): 321-343.
5. Michalski, R. Bratko, I. Kubat, M. (1998). *Machine Learning and Data Mining, Methods and Applications* (Editores) John Wiley & Sons.
6. Kononenko, I. y Cestnik, B. (1986). *Lymphography Data Set*. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Lymphography. Último acceso 29 de Abril del 2008.
7. Michalski, R. (1983). *A Theory and Methodology of Inductive Learning.* Artificial Intelligence, 20: 111-161.
8. Quinlan, J. (1990). *Learning Logic Definitions from Relations.* Machine Learning, 5:239-266
9. Kohonen, T. (1995). *Self-Organizing Maps*. Springer Verlag Publishers.
10. Heckerman, D., Chickering, M., Geiger, D. (1995). *Learning bayesian networks, the combination of knowledge and statistical data.* Machine learning 20: 197-243.
11. Chen, M., Han, J., Yu, P. (1996). *Data Mining: An Overview from a Database Perspective.* IEEE Transactions on Knowledge and Data Engineering, 8(6): 866-883.

12. Chung, W., Chen, H., Nunamaker, J. (2005). *A Visual Framework for Knowledge Discovery on the Web: An Empirical Study of Business Intelligence Exploration*. Journal of Management Information Systems, 21(4): 57-84.
13. Chau, M., Shiu, B., Chan, I., Chen, H. (2007). *Redips: Backlink Search and Analysis on the Web for Business Intelligence Analysis*. Journal of the American Society for Information Science and Technology, 58(3): 351-365.
14. Golfarelli, M., Rizzi, S., Cella, L. (2004). *Beyond data warehousing: what's next in business intelligence?***.** Proceedings 7th ACM international workshop on Data warehousing and OLAP. Pág. 1-6.
15. Koubarakis, M., Plexousakis, D. (2000). A Formal Model for Business Process Modeling and Design. Lecture Notes in Computer Science, 1789: 142-156.
16. Britos, P., Jiménez Rey, E., García-Martínez, E. (2008). Work in Progress: Programming Misunderstandings Discovering Process Based On Intelligent Data Mining Tools. Proceedings 38th ASEE/IEEE Frontiers in Education Conference, en prensa.
17. Kaufmann, L. y Rousseeuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons Publishers.
18. Grabmeier, J., Rudolph, A. (2002). Techniques of Cluster Algorithms in Data Mining. Data Mining and Knowledge Discovery, 6(4): 303-360.
19. Ferrero, G., Britos, P., García-Martínez, R., (2006). *Detection of Breast Lesions in Medical Digital Imaging Using Neural Networks*. In IFIP International Federation for Information Processing, Volume 218, Professional Practice in Artificial Intelligence, eds. J. Debenham, (Boston: Springer), Pág. 1-10.
20. Britos, P., Cataldi, Z., Sierra, E., García-Martínez, R. (2008). Pedagogical Protocols Selection Automatic Assistance. Notes in Artificial Intelligence 5027: 331-336.
21. Britos, P., Grosser, H., Rodríguez, D., García-Martínez, R. (2008). Detecting Unusual Changes of Users Consumption. In Artificial Intelligence in Theory and Practice II, ed. M. Bramer, (Boston: Springer), en prensa.
22. Britos, P., Felgaer, P., García-Martínez, R. (2008). Bayesian Networks Optimization Based on Induction Learning Techniques. In Artificial Intelligence in Theory and Practice II, ed. M. Bramer, (Boston: Springer), en prensa.
23. Britos, P., Abasolo, M., García-Martínez, R. y Perales, F. (2005). *Identification of MPEG-4 Patterns in Human Faces Using Data Mining Techniques*. Proceedings 13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2005. Páginas 9-10.
24. Cogliati, M., Britos, P., García-Martínez, R. (2006a). *Patterns in Temporal Series of Meteorological Variables Using SOM & TDIDT*. In IFIP International Federation for Information Processing, Volume 217, Artificial Intelligence in Theory and Practice, ed. M. Bramer, (Boston: Springer), Pág. 305-314.
25. Britos, P., Dieste, O., García-Martínez, R. (2008b). Requirements Elicitation in Data Mining for Business Intelligence Projects. In Advances in Information Systems Research, Education, and Practice eds. George Kasper e Isabel Ramos (Boston: Springer), en prensa.
26. Britos, P. (2008). *Procesos de Explotación de Información Basados en Sistemas Inteligentes*. Tesis de Doctorado en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata. http://postgrado.info.unlp.edu.ar/Carrera/Doctorado/Tesis/Britos-Tesis%20Doctoral.pdf. Pagina vigente al 30/07/09.