

Investigación en el grupo de Sistemas Inteligentes y Minería de Datos de la UCLM

José A. Gámez, José M. Puerta

Departamento de Informática
Universidad de Castilla-La Mancha
EPSA - Campus Universitario s/n
Albacete, 02071

Resumen En este trabajo se presentan las principales actividades de investigación del grupo de Sistemas Inteligentes y Minería de Datos de la Universidad de Castilla-La Mancha. Se hace especial hincapié en aquellas actividades más relacionadas con el aprendizaje automático y la minería de datos, incluyendo proyectos de investigación en curso, otras líneas de trabajo abiertas, actividades organizadas y docencia estrechamente relacionada con la investigación.

1 Introducción

El grupo de investigación de *Sistemas Inteligentes y Minería de Datos*¹ es un grupo de investigación de reciente constitución (Septiembre - 2000) en el Departamento de Informática de la Universidad de Castilla-La Mancha. Hasta esa fecha la mayor parte de la actividad investigadora se realizó dentro del grupo de investigación de *Tratamiento de la incertidumbre en sistemas inteligentes (UTAI)* de la Universidad de Granada.

Actualmente el grupo está compuesto por 6 profesores a tiempo completo (tabla 1) a los que se unen varios alumnos colaboradores y en fechas próximas (15 de Octubre) un becario de investigación adscrito a uno de los proyectos de investigación en los que el grupo participa.

Nombre	Formación
José Antonio Gámez Martín	Doctor en Informática
José Miguel Puerta Callejón	Doctor en Informática
María Julia Flores Gallego	Ingeniero en Informática
Ismael García Varea	Ingeniero en Informática
Francisco José Vigo Bustos	Ingeniero en Informática
Juan Antonio Guerrero Abenza	Licenciado en Matemáticas (EIO)

Cuadro1. Personal del grupo

¹ <http://www.info-ab.uclm.es/is1>

Indicaremos además que el grupo dispondrá de una sección (de idéntico nombre) en el Instituto de Investigación en Informática de Albacete² (*i³A*), centro concedido hace algo más de un año en convocatoria FEDER y que se está construyendo en la actualidad.

Las principales líneas de trabajo en que se enfoca la investigación (básica y aplicada) son las siguientes:

- *Sistemas de toma de decisiones bajo condiciones de incertidumbre*: Sistemas expertos probabilísticos (Redes Bayesianas), sistemas expertos basados en reglas difusas.
- *Metaheurísticas y algoritmos evolutivos*: algoritmos genéticos, algoritmos de estimación de distribuciones, colonias de hormigas, otras metaheurísticas (búsqueda en entorno variable, enfriamiento estocástico, búsqueda tabú, ...)
- *Minería de datos y aprendizaje automático*: descubrimiento del conocimiento y su representación como: redes Bayesianas (estáticas y dinámicas), árboles de decisión, sistemas basados en reglas difusas descriptivas, ...
- *Traducción automática del habla basada en técnicas estadísticas*.
- Aplicaciones

En las siguientes secciones se detallarán los proyectos de investigación en los que participa el grupo, otras líneas de trabajo y/o aplicaciones, la labor docente relacionada con la investigación, las actividades organizadas y los principales grupos de investigación con los que mantiene continua colaboración.

2 Proyectos en desarrollo

En la actualidad se participa en tres proyectos de investigación financiados por distintas instituciones. En dos de ellos el equipo está formado íntegramente por personal del Grupo de Sistemas Inteligentes y Minería de Datos, mientras que en el tercero sólo dos de los miembros del grupo participan en el equipo investigador.

2.1 TIC2001-2973-C05-05

Se trata de un subproyecto financiado por el Ministerio de Ciencia y Tecnología, perteneciente al proyecto coordinado *Elvira II: Aplicaciones de los modelos gráficos probabilísticos*. Este proyecto es una continuación de otro proyecto coordinado (*Elvira*³) aprobado en la convocatoria TIC97 y del que formaban parte cuatro equipos: Universidad de Granada (coordinador), Universidad de Almería, UNED y Universidad del País Vasco. Nuestra participación en *Elvira* se realizó mediante la inclusión de José A. Gámez y José M. Puerta en el equipo de la Universidad de Granada. Fruto de *Elvira* fue un entorno integrado de desarrollo de modelos gráficos probabilísticos (figura 1) en el que se dispone de

² <http://www.info-ab.uclm.es/iiia/index.htm>

³ leo.ugr.es/~elvira

capacidades de edición, explicación, inferencia y aprendizaje. En esta etapa inicial de Elvira el objetivo era definir estructuras de datos comunes e implementar algoritmos clásicos de inferencia y aprendizaje para poder realizar comparaciones frente a nuevos diseños. Por nuestra parte se añadieron nuevos algoritmos de *razonamiento abductivo* [6, 7] y de *aprendizaje de redes Bayesianas* [10].

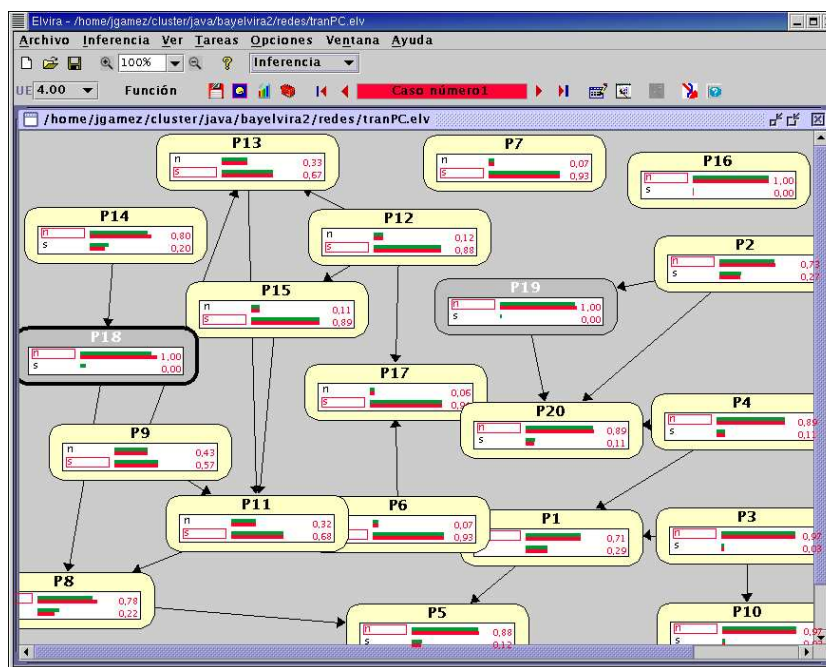


Figura1. Una captura de pantalla de Elvira en el modo inferencia

Tras la creación de nuestro grupo y de cara a la petición de Elvira II, ésta se realiza con 5 equipos: los cuatro anteriores más uno de la Universidad de Castilla-La Mancha, formado íntegramente por personal del grupo de investigación y actuando José A. Gámez como investigador principal. El período de realización es desde Diciembre de 2001 a Diciembre de 2004, y contiene dos objetivos principales:

- Completar el desarrollo del entorno integrado y dotarlo de nuevos algoritmos de inferencia y aprendizaje.
- Abordar una aplicación para cada equipo investigador (agricultura, bio-informática, medicina, filtrado colaborativo y/o comercio electrónico).

En lo que respecta a nuestro subproyecto (05), al margen de colaborar en tareas comunes como son la mejora o redefinición del formato (migrar a XML) y del entorno gráfico, participamos en las siguientes tareas:

■ **A2. Algoritmos de inferencia.**

- *Desarrollo e implementación de algoritmos de triangulación incremental (UCLM,ugr).* Todo algoritmo de propagación tiene implícita una triangulación determinada del grafo original, dependiendo la complejidad de la propagación de la calidad de la triangulación realizada. En este proyecto pretendemos desarrollar nuevos algoritmos de triangulación (directa [3] e incremental) basados en la técnica de la descomposición del grafo en sus subgrafos primos maximales.
- *Desarrollo e implementación de algoritmos de abducción parcial (UCLM, ugr).* El objetivo es continuar con el desarrollo de algoritmos que permitan realizar razonamiento diagnóstico para un conjunto reducido de variables en una red Bayesiana. En Elvira II nos centraremos en los algoritmos aproximados y en la búsqueda de explicaciones minimales.
- *Algoritmos para diagramas de influencia basados en árboles de probabilidad (UGR,uclm,uned).* Se trata de investigar los beneficios que puede aportar a estos algoritmos el uso de una representación más compacta que las tablas de probabilidad, como son los árboles.

Las tareas involucradas en este apartado constituyen el grueso de la línea de tesis doctoral de M. Julia Flores.

■ **A3. Preprocesamiento (EHU,ucl).** Selección de variables, transformación de variables, imputación en BBDD incompletas, categorización.

Aunque en principio no tenemos asignada ninguna de estas tareas, como se verá en las próximas secciones sí que estamos trabajando en alguna de ellas.

■ **A4. Aprendizaje.**

- *Aprendizaje de modelos gráficos dinámicos (UGR,uclm).* En ocasiones los procesos a modelar requieren la inclusión de relaciones temporales entre las variables, lo que nos lleva a hablar de modelos dinámicos. Estudiaremos algoritmos de aprendizaje de estos modelos basados en asumir que existe una red básica que modela un estado de tiempo, la cual es replicada para cada instante que se quiera modelar, y que existen relaciones temporales entre los distintos instantes de tiempo, pero sólo desde t_i hasta t_{i+1} [9].
- *Aprendizaje con conocimiento parcial (UGR,uclm).* En ocasiones se dispone de conocimiento parcial sobre el problema a modelar, relaciones directas entre variables (arcos), preórdenes entre algunas variables, etc, ... La integración de este conocimiento en los algoritmos de aprendizaje ayudará a limitar el espacio de búsqueda y, por tanto, a mejorar la eficiencia/eficacia del proceso.

■ **A5. Aprendizaje e Interfaz gráfica.**

- *Integración de los alg. de aprendizaje en la IG (UCLM,uned).*
- *Especificación de la información parcial inicial para alg. de aprendizaje en la IG (UCLM,ugr)*

■ **B. Aplicaciones.** La aplicación que corresponde a nuestro equipo dentro de Elvira II está relacionada con las Aplicaciones a Comercio Inteligente.

La idea es analizar datos de la empresa ONO mediante los denominados

"Sistemas de recomendaciones": sistemas en los que se aplican técnicas de descubrimiento del conocimiento para realizar recomendaciones a usuarios sobre los posibles productos que les pueden interesar. Estos sistemas están obteniendo especial éxito en el dominio del comercio electrónico. Una exitosa tecnología en los sistemas de recomendaciones es el filtrado cooperativo, el cual formula recomendaciones en base a las efectuadas anteriormente por usuarios de "similares" preferencias. El filtrado cooperativo se ha mostrado especialmente eficaz produciendo recomendaciones de alta calidad. Los sistemas tradicionales de filtrado cooperativo degradan su eficacia y eficiencia cuando el número de productos y usuarios crece, esto es debido al tamaño del conjunto de datos que se deben manejar para producir las recomendaciones. Es por ello, que es interesante poder modelar un sistema a partir de los datos previos, para representar el conocimiento que disponemos del problema de forma compacta y eficiente. Para ello trataremos de aprender un modelo basado en redes Bayesianas a partir de los datos disponibles.

2.2 PBC-02-002

Se trata de un proyecto concedido por la Consejería de Ciencia y Tecnología de la Junta de Comunidades de Castilla-La Mancha, dentro del plan de Desarrollo Tecnológico e Innovación. El título es Librería de programación de técnicas modernas de búsqueda y optimización. Aplicaciones al campo del aprendizaje automático.

Se trata de un proyecto coordinado, formado por dos subproyectos: Universidad de Castilla-La Mancha (coordinador) y Universidad de Granada. El equipo de la UCLM está formado íntegramente por personal del grupo de investigación, y tiene como investigador principal a José A. Gámez. El período de tiempo para la realización del proyecto es del 1/1/2002 al 31/12/2004.

El proyecto está dividido en dos grandes bloques de tareas:

- Dedicadas a implementar una librería de programación de técnicas heurísticas modernas para resolver problemas de optimización combinatoria: algoritmos genéticos, algoritmos de estimación de distribuciones, algoritmos basados en colonias de hormigas, búsqueda tabú, ...
- Dedicadas a resolver problemas concretos usando las técnicas anteriores. De este conjunto de tareas la mayoría están relacionadas con problemas o aplicaciones de aprendizaje automático.

Dado el ámbito de este *workshop* nos centraremos en comentar sólo algunas de las tareas incluidas en este segundo grupo:

- **C1. Aplicación al problema del aprendizaje de redes de creencia (UGR,uclm).**

El problema del aprendizaje de redes de creencia puede ser formulado como un problema de optimización combinatoria, en el que la búsqueda se realiza normalmente en el espacio de grafos dirigidos acíclicos. Sin embargo, otros

espacios como el de órdenes (permutaciones de las variables) o híbridos han sido recientemente considerados.

El objetivo en esta tarea es la adaptación y aportación de conocimiento específico del problema a diversas técnicas de optimización combinatoria, tales como las colonias de hormigas [4], algoritmos de estimación de distribuciones, etc, ...

- **C2. Aplicación al problema de la selección de atributos en aprendizaje automático (UCLM,ugr).**

En la mayoría de los problemas abordados por las técnicas de aprendizaje automático y minería de datos el número de variables o atributos suele ser grande o muy grande. En general, parece probado que sólo un subconjunto de las variables consideradas es relevante y, por tanto, el modelo puede construirse usando este conjunto reducido.

El objetivo en este proyecto es estudiar la aplicación de algoritmos basados en colonias de hormigas y en búsqueda en entorno variable al problema de la selección de variables en aprendizaje automático.

- **C3. Aplicación al problema del aprendizaje de sistemas basados en reglas difusas descriptivas (UCLM,ugr).**

Los sistemas difusos descriptivos o lingüísticos tienen en la interpretabilidad del modelo su fuerte frente a otras técnicas. Sin embargo, también es muy interesante que la exactitud conseguida sea lo más alta posible.

Recientemente se ha propuesto la metodología COR [1] de aprendizaje de estos sistemas, consistente en plantear el problema como la optimización del conjunto de reglas resultante de entre las distintas posibilidades. En este proyecto proponemos abordar el problema haciendo uso de los algoritmos de estimación de distribuciones [2], con el objetivo de explotar las dependencias que puedan existir entre los distintos subespacios de reglas.

- **C4. Aplicación a la clasificación automática (UGR,uclm).** Las técnicas de aprendizaje de redes Bayesianas, que son de tipo genérico, son susceptibles de ser adaptadas o particularizadas para el caso de la clasificación. El objetivo de esta tarea será partir de los métodos desarrollados en la tarea C1 y adaptarlos para el caso específico de la clasificación.

2.3 INDOOR

Se trata del proyecto *Automatización del proceso de producción de compost mediante sistema INDOOR*, financiado por la Consejería de Agricultura y Medio Ambiente de la Junta de Comunidades de Castilla-La Mancha (protocolo de investigación). El proyecto se realiza entre la Universidad de Castilla-La Mancha (6 investigadores del Dpto. de Informática) y la empresa CHAMPINIESTA, actuando el Dr. Antonio Garrido (UCLM) como investigador principal. La duración del proyecto va de Junio de 2000 hasta Mayo de 2003.

El objetivo del proyecto es la automatización del proceso de fabricación del *compost*, para lo que se realizará una monitorización del proceso seguido de unas acciones de control. Actualmente se está en fase de pruebas de la automatización

realizada inicialmente, en la que se toman como entradas tres sensores de temperatura y se ejecuta como acción de control la regulación de la apertura de una compuerta de ventilación. La figura 2 muestra la evolución de una carga de compost.

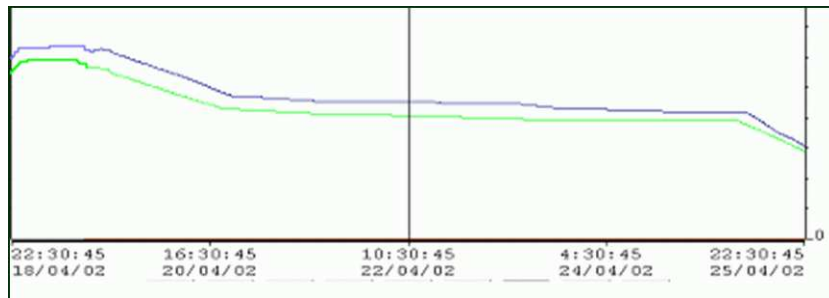


Figura2. Evolución deseable vs evolución actual para una carga de compost

El objetivo de nuestro grupo en este proyecto es tomar como punto de partida la automatización actual y usar el proceso de monitorización para la adquisición de datos que nos permitan aprender un sistema de control basado en reglas difusas descriptivas. Por otra parte, está previsto ampliar las entradas del sistema con un sensor de amoníaco y otro de CO₂, así como sustituir la compuerta de ventilación por un ventilador automático. Indudablemente, también se plantea como objetivo el aprendizaje de un sistema de reglas difusas que modele este nuevo sistema.

3 Otras líneas de trabajo en curso

En esta sección comentaremos otras líneas de trabajo en desarrollo en la actualidad, pero que no han sido incluidas como tareas específicas de ninguno de los proyectos anteriores.

- **Agrupamiento de variables para el aprendizaje de redes Bayesianas.**

Uno de los principales problemas del aprendizaje de redes Bayesianas es la dificultad de los métodos existentes de trabajar con un gran número de variables, lo que implica que los resultados se degraden tanto en eficacia como en eficiencia. El objetivo de esta línea de trabajo es, a partir de la base de datos o conjunto de entrenamiento, identificar grupos de variables formados por variables con estrecha relación entre ellas y con escasa relación con las de los demás grupos. Evidentemente se trata de un problema de *clustering* o agrupamiento y para medir la similaridad/disimilaridad entre variables nos proponemos usar medidas de dependencia.

Una vez obtenidos los grupos de variables, podremos aprender una red para cada grupo y, o bien plantear la multired existente como una aproximación al modelo global; o bien considerar la unión de estas redes en un único modelo mediante la búsqueda de relaciones entre ellos.

- **Aprendizaje de redes Bayesianas mediante selección/promediado de modelos.**
Cuando la cantidad de datos disponibles es pequeña respecto al tamaño del problema, diferentes redes Bayesianas pueden explicar dichos datos de forma similar. En estos casos la técnica de promediado de modelos ofrece un marco de trabajo que permite realizar mejores predicciones que la típica selección de modelos.
La idea de esta técnica es seleccionar un subconjunto de modelos representativos explorados durante el proceso de aprendizaje y usar estos modelos para calcular las probabilidades de distintas características estructurales, como puede ser la presencia o ausencia de un arco.
En esta línea de trabajo lo que nos proponemos es usar el conjunto de modelos representativos seleccionado para generar una base de datos de más alto nivel, en la que cada variable representará una característica estructural (arco). Evidentemente el número de variables en esta nueva base de datos es mucho mayor que en la inicial ($O(n^2)$). La idea es usar esta nueva base de datos para aprender una nueva red Bayesiana que actuará como meta-modelo del problema inicial, y que permitirá razonar acerca de las probabilidades de las distintas características estructurales presentes en el modelo correspondiente al dominio del problema.
Esta tarea presenta interesantes complicaciones, como son la mayor cardinalidad del espacio de búsqueda, que obligará, en general, a usar métodos heurísticos durante el proceso de búsqueda y a aproximar la red Bayesiana mediante un modelo restringido (árboles, poli-árboles, ...).
Una primera aproximación a este problema se encuentra actualmente sometida [8].
- **Minería de datos procedentes de ganadería.**
Actualmente se mantienen contactos con la asociación de ganaderos AGRAMA, la cuál posee numerosos datos sobre animales (ganadería ovina) censados. A partir de contactos iniciales se ha establecido como futura línea de trabajo el intentar estimar la valoración genética de un animal y su potencial producción lechera a partir de los datos disponibles.

4 Actividades organizadas

Las principales actividades organizadas por nuestro grupo han sido:

- *First European Workshop on Probabilistic Graphical Models*, que tendrá lugar en Cuenca del 6 al 8 de Noviembre de 2002.
Para este workshop se ha configurado un comité de programa internacional presidido por José A. Gámez (UCLM) y Antonio Salmerón (UAL); y un comité organizador en el que se incluye todo nuestro grupo de investigación.

- *Sesiones especiales sobre Modelos Gráficos Probabilísticos* en la Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA-2001). Estas dos sesiones fueron organizadas por los profesores José A. Gámez (UCLM) y Antonio Salmerón (UAL) y reunieron investigadores nacionales e internacionales con un total de 13 ponencias.
- *VIII Escuela de Verano de Informática: Sistemas Expertos Probabilísticos*, celebrada en Albacete (Julio, 1998). En este evento se reunieron los principales expertos nacionales en el tema, impartándose un total de 11 conferencias. Actuaron como coordinadores los profesores Serafín Moral (UGR) y Luis M. de Campos (UGR); siendo el comité organizador José A. Gámez, José M. Puerta y Francisco J. Vigo.

5 Asignaturas relacionadas

Los miembros del grupo de Sistemas Inteligentes y Minería de Datos imparten en la Escuela Politécnica Superior de Albacete, las siguientes asignaturas relacionadas con sus temas de investigación:

- En la Ingeniería en Informática:
 - Ingeniería del conocimiento (4º curso)
 - Bio-Informática (5º curso)
 - Minería de datos (5º curso)
- En el Programa de Doctorado:
 - Aprendizaje automático en sistemas expertos probabilísticos
 - Sistemas expertos probabilísticos: modelización e inferencia

6 Colaboración con otros grupos de investigación

El grupo de Sistemas Inteligentes y Minería de Datos colabora intensamente en la actualidad con los siguientes grupos de investigación:

- *Grupo de Tratamiento de la Incertidumbre en Sistemas Inteligentes (UTAI)*⁴ de la Universidad de Granada. Se trata del grupo en el que se han formado los dos únicos doctores del grupo. Además, con este grupo se trabaja de forma conjunta en los dos proyectos coordinados expuestos en la sección 2. Los principales temas de investigación comunes son el aprendizaje de modelos gráficos y los algoritmos de abducción parcial en redes Bayesianas.
- *Grupo de Análisis de Datos (GAD)*⁵ de la Universidad de Almería. Al margen de la colaboración en el proyecto Elvira II y de diversas visitas entre los dos grupos con motivo de charlas y seminarios, con este grupo se ha colaborado y se colabora en la actualidad en la organización de eventos como las Sesiones especiales sobre Modelos Gráficos Probabilísticos en CAEPIA'01 y el Workshop PGM'02. También se trabaja de forma conjunta en algoritmos aproximados para abducción parcial.

⁴ http://decsai.ugr.es/gte/index_gte.html

⁵ <http://www.ual.es/personal/asalmero/elvira/>

- *Grupo de Sistemas Inteligentes (ISG)*⁶ de la Universidad del País Vasco. Con este grupo se colabora en el proyecto **Elvira II** y además se han realizado trabajos comunes en la comparación de algoritmos genéticos y algoritmos de estimación de distribuciones [5]. Actualmente se está iniciando la colaboración en temas de selección de variables para clasificación.
- *Grupo de investigación en redes Bayesianas* de la UNED. Con este grupo se colabora principalmente en el desarrollo del proyecto **Elvira II**.
- *Decision Support Systems Group (DSS)*⁷ de la Universidad de Aalborg. La relación con este grupo se ha consolidado mediante distintas estancias realizadas por miembros de nuestro grupo en Aalborg, así como por la presencia de miembros del grupo DSS en las sesiones organizadas en CAEPIA'01. Actualmente se está trabajando conjuntamente con el profesor Kristian Olesen en métodos de triangulación incremental. Además, el grupo DSS colabora en el proyecto **Elvira II**.
- *Grupo de Reconocimiento de Formas y Tecnología del Lenguaje (PRHLT)*⁸ de la Universidad Politécnica de Valencia. Con este grupo se colabora en temas de traducción automática de lenguajes basada en técnicas estadísticas, tema éste que constituye la línea de tesis doctoral de Ismael García Varea.

Agradecimientos

En la actualidad la mayor parte de la investigación realizada por el grupo de Sistemas Inteligentes y Minería de Datos es financiada por el Ministerio de Ciencia y Tecnología (proyecto TIC2001-2973-C05-05) y por la Consejería de Ciencia y Tecnología de la Junta de Comunidades de Castilla-La Mancha (proyecto PBC-02-002).

References

1. Casillas J., Cordon, O., Herrera, F.: Improving the Wang and Mendel's Fuzzy Rule Learning Method by Inducing Cooperation Among Rules. Eighth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU). 2000.
2. Flores, M.J., Gámez, J.A.: Applicability of Estimation of Distribution Algorithms to the Fuzzy Rule Learning problem: A preliminary study. Proceedings of the 4th International Conference On Enterprise Information Systems (ICEIS'02). 2002.
3. Flores, M.J., Gámez, J.A.: Triangulation of Bayesian networks by retriangulation. International Journal of Intelligent Systems (2002, en prensa).
4. De Campos, L.M., Fernández, J.M., Gámez, J.A., Puerta, J.M.: Ant colony optimization for learning Bayesian networks. International Journal of Approximate Reasoning. (2002, en prensa).

⁶ <http://www.sc.ehu.es/ccwbayes/main.html>

⁷ <http://www.cs.auc.dk/research/DSS/>

⁸ <http://prhlt.iti.es>

5. De Campos, L.M., Gámez, J.A., Larrañaga, P., Moral, S., Romero, T.: Partial Abductive Inference in Bayesian Networks: An empirical comparison between Gas and EDAs. En *Estimation of Distribution Algorithms. A new tool for evolutionary computation*, pp. 319–338 (2001). Kluwer Academic Publishers, Boston.
6. De Campos, L.M., Gámez, J.A., Moral, S.: Partial Abductive Inference in Bayesian Belief Networks by Simulated Annealing. *International Journal of Approximate Reasoning*. Vol. 27:3 (2001), pp. 263–283.
7. De Campos, L.M., Gámez, J.A., Moral, S.: Partial Abductive Inference in Bayesian Belief Networks - An Evolutionary Computation Approach by Using Problem-Specific Genetic Operators. *IEEE Transactions on Evolutionary Computation*. Vol. 6:2 (2002), pp. 105–131
8. De Campos, L.M., Gámez, J.A., Puerta, J.M.: Graphical Models to causal discovery from data. *Sometido*.
9. De Campos, L.M., Puerta, J.M.: Learning dynamic belief networks using conditional independence tests. In *proceedings of the 8th International Conference on Information Processing and Management of Uncertainty (IPMU'00)*. Vol I (2000), pp. 325-332.
10. De Campos, L.M., Puerta, J.M.: Stochastic Local Algorithms for Learning Belief Networks: Searching in the Space of Orderings. *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'2001)*. LNAI, Vol. 2143 (2001), pp. 228–239.