Proyecto BMF2003-05695 (MTM2006-15671)

Redes funcionales, neuronales y bayesianas en problemas de clasificación y predicción

María Asunción Beamonte[1], Beatriz Lacruz[1], David Lahoz[1], Rosa Eva Pruneda[2] and Cristina Solares[2]

[1]Universidad de Zaragoza

asunbea@unizar.es, lacruz@unizar.es, davidla@unizar.es

[2]Universidad de Castilla-La Mancha

rosa.pruneda@uclm.es, cristina.solares@uclm.es

1 Problemas que tratamos de resolver

Nuestro objetivo principal es el desarrollo de técnicas matemáticas para el tratamiento de problemas de clasificación y predicción basadas tanto en métodos estadísticos clásicos como es la regresión logística, como en técnicas más novedosas como son las redes funcionales, neuronales y probabilísticas así como los modelos bayesianos espacio-temporales.

Trabajamos tanto los aspectos teóricos de los modelos como su implementación, que es donde aparece nuestro interés en los aspectos computacionales. Además, utilizamos tanto datos simulados como reales.

En lo que se refiere a los datos reales utilizamos, por una parte, bases de datos extensamente estudiadas como son las que están disponibles en las páginas electrónicas StatLib, UCI KDD Archives y UCI ML Repository, entre otras. Y, por otra, disponemos de datos medioambientales proporcionados por Instituto Nacional de Meteorología como las bases TEMP y SYNOP, datos de pacientes de una Unidad de Tabaquismo sobre características de individuos que se someten a tratamiento para dejar de fumar y datos del mercado inmobiliario en Zaragoza proporcionados por el Colegio Nacional de Registradores de la Propiedad, por la Diputación General de Aragón, el Ayuntamiento de Zaragoza y la Cámara de Comercio de Zaragoza.

2 Aspectos computacionales de nuestra investigación

Nuestra investigación requiere, por una parte, la programación de algoritmos propios para lo que utilizamos:

1. Mathematica por su versatilidad en el cálculo simbólico, que permite analizar la sensibilidad de las soluciones, y
2. MatLab por su eficacia en el cálculo vectorial y la cantidad de funciones, en particular estadísticas, que tiene implementadas.

Por otra parte, para el análisis de datos y la comparación de técnicas usamos paquetes estándar, entre los que se encuentran:

1. SPSS, para el análisis estadístico de datos con técnicas clásicas y para el trabajo con redes neuronales, y
2. SPLUS para el análisis estadístico de datos con técnicas avanzadas, en particular, el módulo SpatialStats para técnicas de estadística espacial.

Para la obtención de datos simulados utilizamos las funciones de variables aleatorias implementadas en MatLab.

Para la implementación de modelos gráficos probabilísticos hemos utilizado Bayes Net Toolbox que es una herramienta libre para MatLab que proporciona Kevin Murphy.

Cuando se trata de implementar la selección de modelos utilizamos métodos MCMC y estamos particularmente interesados en los algoritmos genéticos, aunque todavía no hemos trabajado con ellos.

3 Dificultades

La principal dificultad con la que nos encontramos es que al no ser programadores expertos, nuestras implementaciones no suelen ser eficientes al máximo. Sin embargo, en algunos casos nos preocupa el desarrollo de algoritmos de búsqueda y selección de modelos que sean capaces de, en un tiempo computacional razonable, encontrar buenas soluciones.

Muchas de las funciones que implementamos, sabemos que están programadas por otros investigadores y muchas veces las encontramos en la red, pero resulta casi siempre más cómodo (que no más rápido ni más fácil) pensar una nueva implementación que tratar de comprender lo que han hecho otros.

Además, nos gustaría y es de hecho uno de nuestros objetivos, que las personas que leen nuestros artículos tuvieran disponibles las implementaciones y les fueran útiles, pero para ello deberían disponer de un interface amigable y esto nos resulta muy costoso de conseguir en tiempo.

Por otra parte, trabajar con grandes bases de datos reales requiere mucho más esfuerzo que trabajar con datos simulados o con las bases usuales que están disponibles en la red (StatLib, UCI KDD Archives y UCI ML Repository) ya que sabes de antemano a qué tipo de datos te enfrentas, por razones obvias en el primer caso y por tratarse de bases de datos muy estudiadas en el segundo.