

## ANÁLISIS INTELIGENTE DE DATOS COMO HERRAMIENTA DE INTEGRACIÓN EN LA GESTIÓN INTEGRADA DEL AGUA

Joaquín Izquierdo<sup>1\*</sup>, Manuel Herrera<sup>1</sup>, Rafael Pérez<sup>1</sup>, P. Amparo López<sup>1</sup>

**Resumen** - Durante las dos o tres últimas décadas, el campo del agua ha evolucionado pasando de ser una disciplina fundamentalmente concentrada en el diseño y construcción de grandes estructuras (presas, trasvases,...) a otra en la que se consideran conceptos como conservación, ahorro, precio del agua, salud humana, derechos fronterizos, participación social,... Así se está haciendo imprescindible cada día más que Ingenieros Hidráulicos, ecologistas, economistas, abogados, sociólogos y otros profesionales no relacionados con las Ingenierías trabajen conjuntamente. Actualmente la mayor parte de proyectos de Ingeniería son multidisciplinarios y requieren de la integración de destrezas diversas. El reto está en conseguir una integración adecuada y eficiente. En este artículo se presenta una propuesta para una posibilidad de integración basada en el denominado Análisis Inteligente de Datos. Se trata de un enfoque en el que la teoría cede su protagonismo clásico a los datos. En multitud de campos y también en la Gestión Integrada del Agua (IWM, integrated water management) y en el campo Medioambiental la potencia de las TICs (Tecnologías de Información y Comunicación) avanzadas ha posibilitado la realización de campañas de obtención de datos a gran escala. El número de parámetros que deben ser medidos para controlar un (eco)sistema es potencialmente alto. La medición sistemática de esos parámetros genera enormes cantidades de datos que deben ser utilizados e interpretados adecuadamente. Para obtener la mejor información se debe utilizar un planteamiento pragmático, tratando de extraer el máximo conocimiento de todos estos datos. Dentro de toda esta cantidad de datos se encuentra mucha información oculta, en términos de modelos, patrones, tendencias, etc. Pero resulta difícil extraer dicha información debido a la heterogeneidad de los datos en calidad y cantidad. Como consecuencia, la extracción semiautomática de conocimiento a partir de datos ha cobrado gran importancia dentro de la comunidad científica y económica. El denominado Descubrimiento del Conocimiento a partir de Bases de Datos (KDD, Knowledge Discovery from Databases) ha surgido como un marco donde ha florecido una plétora de técnicas de identificación de patrones útiles y comprensibles. Muchas de estas técnicas pueden ser usadas en el campo Medioambiental y, en particular, en IWM. En este artículo intentamos ofrecer una visión de conjunto de lo que es KDD y mencionar algunas de sus aplicaciones en estas áreas.

**Abstract** - Over the past two or three decades, water issues have evolved from a discipline focused primarily on design and construction of large structures (dams, waterways, transfers,...) to one that includes concepts like conservation, preservation, water price, human health, border rights, social participation,... Thus, more and more, there is a need for Hydraulic Engineers, ecologists, economists, lawyers, sociologists and other non-engineering professionals to work jointly. Today most hydraulic engineering projects are multidisciplinary and require integration of different skill. The challenge is to prepare for integration. In this

---

<sup>1</sup>GMMF – Grupo Multidisciplinar de Modelación de Fluidos. Universidad Politécnica de Valencia.

\*E-mail: [jizquier@gmmf.upv.es](mailto:jizquier@gmmf.upv.es) Tel. +34 963879890 Fax. +34 963877981

paper a proposal for a possibility of integration based on the so-called Intelligent Data Analysis is presented. It consists on an approach in which theory gives up its classical stardom to data. In many fields, and IWM (integrated water management), including the Environment field, the power of advanced ICTs (Information and communication technologies) has allowed for large-scale data collection campaigns. The number of parameters that must be measured to monitor an (eco)system is potentially high. Systematic measurement of those parameters generates huge amounts of data that should be suitably interpreted and used. A pragmatic approach has to be used to get the best information = knowledge from all this data. Within such amounts of data there is a lot of hidden information, in terms of models, patterns, trends, etc. But information is difficult to be extracted since data is of varying quantity and quality. As a consequence, semi-automatic knowledge extraction from data has gained great importance within the economic and scientific community. Knowledge Discovery from Databases (KDD) has emerged as a framework where a plethora of techniques for identifying useful and understandable patterns in data have flourished. Most of those techniques can be used with success in the Environmental field and, in particular, in IWM. In this paper we attempt to give an overview of what KDD is and mention some applications in these areas.

**Palabras clave** - Multidisciplinaridad, Minería de Datos, Algoritmos genéticos, Redes Neuronales, Lógica Borrosa, Árboles de Decision, Aprendizaje Automático, Análisis Inteligente de Datos, Gestión Integrada del Agua.

## INTRODUCCIÓN

*Motivación: 1) La multidisciplinaridad del campo del agua*

No es preciso ya invertir ni mucho tiempo ni mucho esfuerzo para motivar la dimensión multidisciplinar del campo del agua. Así que todos, incluidos los Ingenieros, debemos cambiar decididamente nuestra mentalidad. El mundo del agua no es ya solo un campo para expertos científicos. Hasta no hace mucho, el Ingeniero era un acerbo de conocimientos, es decir, un conocedor. Pero actualmente el Ingeniero debe ser un acerbo de la suma de todos los medios necesarios para acceder al conocimiento, es decir, un consumidor de conocimiento. Los Ingenieros utilizaban el ordenador como una herramienta para realizar cálculos. Actualmente deben utilizar procesadores de conocimiento como mecanismos para manipular el conocimiento encapsulado. El desarrollo matemático y científico los llevó a una separación gradual del entorno social inmediato, concentrando su atención los aspectos técnicos y científicos de un problema, dejando los aspectos sociales e incluso los constructivos a otras personas. Se confinaron en su torre de marfil completamente convencidos de que lo que era bueno para la ciencia era bueno para la humanidad y la naturaleza. Pensaban que los resultados técnicos de sus estudios encontrarían de alguna manera y en algún lugar aplicaciones sociales, mientras que los procesos por lo cuales tales aplicaciones sociales se plasmasen no eran de su incumbencia directa y, por tanto, no merecían tiempo de su investigación. En el nuevo paradigma, por el contrario, es prácticamente imposible investigar en el campo del agua sin investigar simultáneamente en cómo los resultados de tal investigación producirán las aplicaciones sociales. En el nuevo paradigma, es preciso desarrollar nuevos medios tecnológicos que permitan esa aplicación social. Así, el mundo del agua es un campo esencialmente socio-tecnológico. El campo del agua ya no es homogéneo

sino heterogéneo. Se debe entender que las acciones a realizar deben proporcionar de manera indivisible medios técnicos y sociales, o socio-tecnológicos, de modo que la mayor cantidad de personal involucrado en el mundo del agua, normalmente no técnico, pueda seguir una línea de acción u otra, apoyado por un verdadero arsenal de técnicas.

En este terreno, resulta de crucial importancia el papel decisivo de la gestión del conocimiento. Por ejemplo, la integración de medidas de campo, modelos numéricos y datos proporcionados por sensores remotos proporcionan nuevas posibilidades que mejorarán grandemente el valor de la actividad conjunta. Así que debemos comprometernos seriamente en aquel tipo de actividades tales como técnicas para encapsular conocimiento latente, con la intención de acelerar el acceso al conocimiento. Algunas de tales técnicas se utilizan ya en el campo del agua y creemos que pueden jugar un papel decisivo como herramienta de integración en la gestión integrada del agua. Se las clasifica, a veces, como técnicas de análisis inteligente de datos. En este artículo se presentan brevemente algunas de tales técnicas y se referencian aplicaciones a través de las que puede ser reconocido su potencial.

#### *Motivación: 2) ¡cantidades de datos aplastantes!*

De acuerdo con la guía para la implementación de la Directiva Marco del Agua de la UE (WFD, Water Framework Directive), (Anexo V), la lista de elementos sobre calidad contiene parámetros biológicos (prioritarios), hidromorfológicos (secundarios) y físico-químicos (secundarios). Actualmente, es más que conocido que las Directivas dan recomendaciones sobre los parámetros que deben ser medidos, especialmente sobre sustancias prioritarias y “relevantes”, por ejemplo, las sustancias de la Lista 1 de la Directiva 76/464EEC y la Lista 2 de sustancias prioritarias de acuerdo con el Artículo 16 de la WFD. Obviamente el número de parámetros susceptibles de medida es potencialmente alto, por lo que debe ser utilizado un planteamiento pragmático a la hora de seleccionar los parámetros a medir, [Lavkulich, 2004]. Las mediciones sistemáticas de todos esos parámetros generarán grandes cantidades de datos que bajo la hipótesis favorable de que estén bien organizados producirán bases de datos de tamaño considerable que han de ser adecuadamente asimiladas e interpretadas para sacar provecho de su contenido.

Ésto es sólo un ejemplo de lo que ocurre en muchas compañías, industrias y grupos de investigación en casi cualquier campo. Extensas cantidades de datos están hoy en día disponibles gracias a que recopilar datos es fácil y normalmente no es caro. Conforme las preocupaciones medioambientales crecen y la tecnología de la información avanza, se recogen más y más datos de aspectos muy diferentes del medioambiente (físicos, químicos, biológicos...). La monitorización medioambiental es una importante fuente de datos. Típicamente, se toman muestras de aire, suelo o agua y se analizan sus propiedades. Los tests de laboratorio, diseñados para estimar la influencia de algunas sustancias sobre el medioambiente, y el propio hombre añaden nueva información al conocimiento. La captación remota es otra fuente de datos: distintos satélites proveen continuamente de imágenes de la Tierra que contienen información sobre la atmósfera, geología, vegetación... Finalmente, evitando ser exhaustivos, los Sistemas de Información Geográfica (GISs), por su rápida divulgación, son una valiosa fuente de información medioambiental de relevancia, proporcionando modelos de elevación digital, de insolación, de usos del suelo, etc.

La mayoría de las personas hace uso de toda esta información para observar la evolución de ciertas variables, controlar valores anormales y escribir informes simplificados. Esto es un objetivo por sí mismo. Pero, sin duda, dentro de toda esta cantidad de datos se encuentra mucha información, en términos de modelos, patrones, tendencias, etc. Siendo capaces de extraer esta información, debiéramos ofrecer perspectivas más globales, elementos de juicio más fundados, indicadores de salud

humana y medioambiental más eficientes y realistas, capacidad de predicción, etc. Resulta difícil extraer esta información dado que los datos son variables en cantidad y calidad. Por una parte, los datos son abundantes y de distribución variable espacial y temporalmente. En muchas ocasiones es demasiado amplia para ser procesada por el humano de modo que resulte útil para el problema que tiene entre manos. Por otra, los casos primarios, tal como son recogidos por personas o sensores incluyen típicamente algún tipo de dato erróneo. Los valores ausentes, los datos con ruido y los datos no homogéneos son los datos erróneos típicos encontrados en los ficheros, aunque, en algunos casos, se encuentran también inconsistencias o datos distorsionados por razones de seguridad [Doyle, 2001].

Es, de este modo, comprensible que la extracción semi-automática del conocimiento de los datos haya ganado importancia dentro de la comunidad científica y económica. El hecho de que toda la información esté en formato digital ha sido también definitivo. Llegados a este punto, es posible formularse la siguiente pregunta. ¿Qué es lo nuevo en este enfoque si la modelación matemática y la Estadística han sido utilizadas durante décadas o incluso siglos? La respuesta es diversa. Pero la razón principal emana del hecho de que esas técnicas son principalmente de validación y es el usuario el que debe dar el modelo para su validación, dejando la calibración y el ajuste de los parámetros a la Estadística. Más aún, para realizar esta tarea el usuario debe tener algunos conocimientos de Matemáticas y Estadística [Izquierdo, 2004a]. Este es el paradigma clásico para la modelación fundamentado en la teoría. Sin embargo, las necesidades en la toma de decisiones requieren de modelos nuevos, no convencionales y no obtenibles de manera manual para extraer conocimiento de tan gran número de datos como el que se tiene. Esta forma de conocimiento, por otra parte, no debiera requerir habilidades especializadas. El nuevo paradigma está basado en los datos [Abbot, 2001]. Estamos hablando sobre Descubrimiento del Conocimiento a partir de Bases de Datos (KDD), definido [Fayyad, 1996] como “el proceso no trivial de identificar patrones en los datos, que sean válidos, novedosos, potencialmente útiles y, finalmente, comprensibles”.

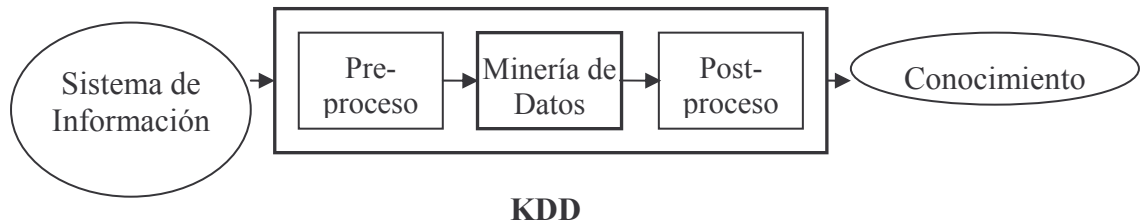
Dada la enorme aplicabilidad de las ciencias medioambientales cualquier repaso de las aplicaciones de los métodos de KDD tiende a ser incompleta. En este artículo intentamos dar una visión de conjunto de lo que es KDD y mencionar algunas aplicaciones en ciencias medioambientales, especialmente en IWM.

## **El proceso del KDD**

Algunas veces el KDD se confunde con la Minería de Datos (DM). Pero, estrictamente hablando, la DM se dedica a generar y contrastar modelos y, como tal, es solo una parte del complejo proceso del KDD, dedicada a la aplicación de técnicas computacionales [Dzeroski, 2001] (ver Figura 1). Con anterioridad a este proceso se deben identificar las fuentes de información, diseñar almacenes de datos adecuados y suministrar las técnicas de navegación y visualización para identificar los aspectos que sean de interés. De esta manera el proceso del KDD abarca tres etapas claramente definidas [Torra, 2003]:

1. Pre-proceso: dedicado a incrementar la calidad de los datos mediante procedimientos de limpiado de datos y de re-identificación.
2. Construcción de modelos (esto es propiamente DM): utiliza los llamados operadores de agregación o fusiona información a través de la combinación de varios modelos de datos.

3. Post-proceso (o extracción de la información): para interpretar, transformar y representar los patrones extraídos, mediante la construcción de resúmenes y reduciendo dimensionalidades. Esta fase también incluye la divulgación y el uso del nuevo conocimiento.



**Figura 1. Diagrama del KDD**

Obviaremos aquí el pre-proceso y el post-proceso, campos en los que se están invirtiendo enormes esfuerzos (ver [Tsumoto, 2003], [Torra, 2003], [Yager, 2003], [Keim, 2003] y las citas en ellos), y consideramos algunas de las técnicas más populares de DM utilizadas en aplicaciones Medioambientales.

## Minería de datos

Una vez que los datos han sido recogidos y almacenados, es importante decidir en qué tipo de patrón se está interesado. El conocimiento objetivo condiciona la técnica de DM a utilizar. Afortunadamente, muchos sistemas de DM permiten seleccionar la técnica en el supuesto de que el usuario haya escogido el patrón. A continuación presentamos los patrones más comunes.

- Asociación de dos atributos que ocurre cuando la frecuencia de ambos presentándose a la vez es relativamente alta.
- Dependencia mostrada por un atributo cuando su valor (absoluto o relativo) está determinado por valores de otros atributos. Hay que observar que, algunas veces, las dependencias son tan obvias que no son interesantes.
- Clasificación es la identificación de un conjunto de dependencias que permiten que se puedan asignar un valor o una categoría entre varias posibles a un parámetro determinado.
- Clustering es la identificación de grupos de individuos. Mientras que en la clasificación las categorías se conocen a priori, en clustering los grupos deben ser identificados. Así, la clasificación es un proceso supervisado mientras que el clustering no lo es.
- Tendencias, que permiten la predicción del valor de una variable continua dependiente usualmente del tiempo.
- Reglas generales que son patrones más generales que no se ajustan a ninguna de las anteriores categorías.

Una vez que el patrón de interés ha sido decidido el sistema o el usuario selecciona las técnicas más adecuadas. La representación de patrones puede ser realizada tanto por técnicas simbólicas como no simbólicas. Las técnicas no simbólicas son más adecuadas para variables continuas, pero necesitan una percepción más clara del conocimiento buscado. Entre éstas tenemos:

- Técnicas estadísticas. Varias técnicas estadísticas pueden ser utilizadas para confirmar asociación y dependencia. Regresión lineal (y no lineal) y las redes de regresión son quizá las más importantes de entre las utilizadas para establecer tendencias.
- Métodos de vecindad y sus variantes, junto con el Aprendizaje Basado en Ejemplos. Son usados para la clasificación y el clustering mediante el uso de distancias y similitudes con el prototipo u otros miembros del grupo.
- Redes Neuronales Artificiales, Lógica Borrosa y Algoritmos Genéticos y sus combinaciones. Son técnicas muy populares (y ya tradicionales) de aprendizaje automático con importantes aplicaciones en la clasificación y el clustering. Se espera (razonablemente) que, aunque son capaces de modelar con precisión muchos fenómenos, no ofrecen interpretabilidad al modelo. Sin embargo, algunas combinaciones permiten, utilizando reglas, obtener modelos más fáciles de entender.

Por el contrario, las técnicas simbólicas ofrecen modelos más entendibles y pueden trabajar con una amplia variedad de variables y estructuras de datos. Entre otras tenemos:

- Árboles de decisión. Se utilizan para la clasificación y clustering mediante un test en cascada que genera una estructura jerárquica donde cada nodo interno contiene una pregunta sobre un atributo, cada rama corresponde a un resultado de la pregunta y cada hoja (nodo final) ofrece una predicción del valor de la variable de clase. Los árboles de regresión son similares a los árboles de decisión pero son utilizados para el análisis de regresión en vez de para la clasificación.
- La Programación Inductiva y otras Técnicas de Inducción Simbólica de Alto Nivel. Son usadas para obtener patrones más generales. Sin embargo, la denominada Programación Lógica Inductiva (ILP) proporciona algunas aproximaciones sencillas e interesantes basadas en la inducción de reglas. Una regla es un patrón de la forma 'SI Conjunción de condiciones ENTONCES Conclusión'. Las condiciones individuales en la Conjunción contrastarán los valores de los atributos individuales. La conclusión da una predicción del valor de la variable (clase) objetivo. Un conjunto de reglas ordenadas se denomina lista de decisión. Las reglas en la lista son consideradas de arriba a abajo de la misma. La primera regla a la que se ajusta un ejemplo dado es usada para predecir el valor de su clase. Un árbol de decisión siempre puede ser transformado en una lista de decisión con cada hoja del árbol de decisión correspondiendo a una regla de clasificación de la lista. Por el contrario, las listas de decisión no siempre pueden ser transformadas en árboles de decisión.

Existe toda una constelación de técnicas y combinaciones de ellas, y los sistemas de KDD no dejan de incorporar tantas como sea preciso junto con técnicas heurísticas con el propósito de aconsejar al usuario sobre los mejores métodos para su problema. Dado el amplio espectro tanto de las ciencias medioambientales como en las técnicas del KDD, cualquier revisión de aplicaciones resultará incompleta. En los apartados siguientes se introduce de manera breve tan sólo un número reducido de técnicas de construcción de modelos con DM, se referencian diferentes aplicaciones en el campo del Medioambiente, relacionadas principalmente con aspectos del agua, y se proporciona información sobre implementaciones de estas técnicas en paquetes de software.



## KDD, Algunas posibilidades

### a) *Redes Neuronales Artificiales (RNA).*

La aplicación de las RNAs a las Ciencias Medioambientales crece de manera rápida. Una RNA es un modelo inspirado en la estructura del cerebro que trabaja muy bien con tareas complicadas como el reconocimiento de patrones, la compresión de datos, la optimización y el clustering. Las RNAs constituyen una familia de modelos, más que una técnica simple. Cada clase de RNA se optimiza para un conjunto específico de condiciones. Una de los paradigmas de RNAs utiliza un conjunto de entrenamiento de datos etiquetados para realizar el aprendizaje. Los datos son los patrones de entrada junto con su respuesta correcta. Este tipo de aprendizaje se denomina supervisado. Sin embargo, algunas veces, aunque tengamos disponible un enorme conjunto de datos no hay indicios de cómo esos datos están organizados, ni siquiera para un pequeño subconjunto que pudiera servir como conjunto de entrenamiento. En este caso sólo los parámetros de entrada están disponibles. Sin una respuesta objetivo de la que aprender, la red neuronal sólo puede descubrir por sí misma patrones típicos, regularidades, grupos, o cualquier otra relación de interés dentro del conjunto de entrenamiento. Es el aprendizaje no supervisado. De todas las RNAs pertenecientes al primer paradigma el Perceptrón Multicapa (MLP, multi-layer perceptron) y las Funciones de Base Radial (RBF, Radial Basis Functions) adiestradas mediante los denominados algoritmos de retropropagación (BP, back propagation) son las más populares (ver, por ejemplo, [Cichocki, 1993]). Entre las RNAs no supervisadas los mapas Auto-Organizativos de Kohonen (SOM, Self-Organizing Maps) son los más populares, [Kohonen, 2001]. Para tratar de ser breves, nos restringiremos aquí a dar una somera descripción de las Redes Neuronales Unidireccionales (Feedforward NN), una clase a la que pertenecen MLPs y RBFs.

Una red neuronal unidireccional es capaz de generalizar datos no vistos previamente cuando es entrenada por un algoritmo de aprendizaje llamado de retro-propagación. Esta red es un modelo de transferencia, en el sentido de que las entradas (input) simplemente permiten calcular las salidas (output) de acuerdo con el aprendizaje al que han estado sujetas. Un MLP consiste en un conjunto de unidades de proceso (neuronas) de input distribuidas, llamado primera capa o capa de entrada, un conjunto de neuronas de salida, llamado capa de salida y una o más capas intermedias (ocultas). Las neuronas de una capa están interconectadas con las neuronas de la siguiente capa, por medio de conexiones caracterizadas por ciertos pesos. Las neuronas reciben entradas o bien de los datos de entrada o de las interconexiones. La capa de entrada pre-procesa el espacio de entrada y las otras capas, basándose en el output de la capa anterior, construyen las superficies de discriminación necesarias para resolver el problema. Finalmente, la capa de salida ofrece la respuesta o output.

En la fase de aprendizaje el output es comparado con el objetivo, el cual es proporcionado simultáneamente con los datos de entrada correspondientes. Los errores (discrepancia entre el output y el objetivo) de la capa de salida son transmitidos progresivamente hacia atrás a las capas ocultas. Esa retro-propagación de errores es usada para ajustar los pesos de las conexiones entre capas. El ajuste de los pesos entre las capas y el cálculo de las nuevas salidas constituyen un proceso iterativo que se lleva a cabo hasta que los errores son menores que una cierta tolerancia. El proceso de aprendizaje se lleva a cabo mediante técnicas de optimización no lineal basadas en métodos tales como el gradiente y el gradiente conjugado. Algunos parámetros controlan la velocidad de aprendizaje ejerciendo cierta influencia sobre las magnitudes de los pesos después de cada evaluación de los errores. De este modo, al mostrar a la red parejas de vectores formados por los datos de entrada y la respuesta correcta (objetivo), se lleva a cabo el aprendizaje. Cuando se comparan los valores objetivo con los que el estado actual de la red proporciona, ésta puede evaluar el error, lo que permite reajustar los pesos hasta que el error decrece hasta un valor aceptable.

Las RNAs han sido utilizadas para abordar diversos problemas en las Ciencias Medioambientales. Aquí referenciamos sólo unos pocos ejemplos.

Monitorización Medioambiental. La protección medioambiental implica ‘la vigilancia o comprobación periódica o continua que permita determinar el nivel de conformidad con los requerimientos estatutarios y/o niveles de contaminación en distintos medios, o en personas, plantas y animales’ [EPA, 2000]. De este modo, la protección medioambiental implica la monitorización de diferentes variables, lo que además incluye el muestreo regular, el análisis y la interpretación de las muestras en términos de calidad de agua, por ejemplo. También implica el estudio de los efectos de la contaminación en la salud humana, etc. Las RNAs son utilizadas para interpretar y clasificar esas muestras. En [Ruck, 1993] se utilizan RNAs para clasificar la calidad del agua de los ríos. [Walley, 1996] compara las RNAs con otras técnicas de KDD para clasificar muestras biológicas tomadas de ríos británicos. [Walley, 2000] utiliza redes neuronales no supervisadas para diagnosticar la calidad del río a partir de datos biológicos y medioambientales. [Rowland, 2004], después de reconocer que la determinación de indicadores para controlar el entorno natural, que asegure la sostenibilidad ecológica de zonas sometidas a estrés, es problemática y susceptible de errores y que las técnicas estadísticas clásicas presentan determinados inconvenientes derivados básicamente de la calidad de los datos, utiliza RNAs para examinar los datos del ecosistema en un espacio multidimensional y seleccionar el mínimo número de medidas indicadoras que tengan el mayor peso en el mantenimiento de un ecosistema sostenible.

Predicción y Análisis de Series Temporales. Las RNAs no sólo son capaces de resolver problemas estáticos (el output es consecuencia directa de un cierto input estático), sino que también son capaces de generar un patrón de salida o una sucesión entera de patrones de salida a partir de una sucesión de patrones de entrada. El objetivo es entrenar una red neuronal apropiada por un lado para reconocer sucesiones de vectores de entrada y por otro para generar la correspondiente salida, posiblemente secuencial. Estos tipos de RNAs son denominados dinámicos. Para obtener un proceso de evolución tal del estado de la red, se añaden a la estructura unidireccional consideradas conexiones hacia atrás de una neurona en una capa a otra neurona en una capa previa, conexiones cruzadas entre las neuronas de una misma capa y auto-conexiones de una neurona dada consigo misma. Las arquitecturas neuronales, que presentan autoconexiones y/o conexiones cruzadas y/o conexiones hacia atrás, junto a las tradicionales conexiones hacia adelante, son llamadas Redes Neuronales Recurrentes (RNR). Las RNRs han mostrado ser adecuadas para tareas de aprendizaje dinámico y, en algunos casos, para caracterizar diferentes evoluciones dinámicas de los patrones de entrada. Además de los artículos de investigación dedicados al Análisis de Series Temporales en general, como [Tronchi, 2003] (por citar uno), se pueden encontrar abundantes artículos considerando el Análisis de Series Temporales en cuestiones Medioambientales. En [Anctil, 2004] una RNA combinada con descomposición wavelet es utilizada en la predicción de la escorrentía. La predicción del caudal de un río para la gestión de depósitos a través de RNAs se considera en [Baratti, 2003]. En [Panella, 2003] se utiliza un tipo específico de red neuronal para abordar el problema de predecir valores futuros de sucesiones de datos medioambientales. Los autores sostienen que su enfoque permite mejorar la precisión de la predicción de sucesiones de datos reales, lo que permitirá una gestión de coste optimizado de los recursos disponibles.

RNAs en Agua Potable y Sistemas de Aguas Residuales. Una plétora de artículos de investigación también trata estas aplicaciones. La monitorización, el control y los test a escala en sistemas de distribución de agua se consideran, por ejemplo, en [Izquierdo, 2004b], [Bhattacharya, 2003] y [Baxter, 2004]. La calidad del agua es también considerada ampliamente, como en [Millot, 2002]. Finalmente, los aspectos relativos a aguas residuales pueden también beneficiarse de las RNAs, [El-Din, 2004].



## *b) Lógica Borrosa*

En muchos, si no en todos los escenarios del mundo real, no disponemos de mediciones exactas. Por ejemplo, la complejidad del proceso de envejecimiento de las tuberías y la dificultad, el coste y el tiempo de realizar medidas precisas hacen que el diámetro efectivo y la rugosidad no sean valores fácilmente predecibles. Mientras que los datos deterministas pueden ser fácilmente incluidos en modelos y fórmulas, la información imprecisa representa un gran problema durante la modelación. El uso de valores simples (medias, por ejemplo, o valores máximo y mínimo), aunque resulta una forma de cálculo rápida y a veces conveniente, es a menudo simplista y debe ser corregido con valores de seguridad apropiados. Por otro lado, atribuir a un problema un carácter artificialmente estadístico mediante probabilidades asignadas subjetivamente conlleva imprecisiones y un alto grado de aleatoriedad y corre un alto riesgo de estar inventando información o distribuciones que no existen. En el contexto de la incertidumbre el interés se centra en el rango en el que caen nuestras mediciones. Es importante subrayar que ésto es diferente de la probabilidad, donde trabajamos con la verosimilitud de que una cierta medida exacta sea obtenida [Zadeh, 1995]. Para tratar con la incertidumbre han sido propuestas diversas aproximaciones. Por ejemplo, los intervalos aritméticos nos permiten trabajar y hacer cálculos con intervalos en vez de con números exactos [Moore, 1979]. Una herramienta conceptual apropiada para este tipo de datos es la teoría de los conjuntos borrosos [Kaufmann, 1991], [Bardosy, 1995]. El concepto de conjunto borroso fue introducido por Lotfi A. Zadeh en 1965, [Zadeh, 1965]. Es un enfoque para trabajar con conceptos imprecisos basados en la Lógica Borrosa. Este tipo de lógica nos permite manejar la incertidumbre de una manera muy natural e intuitiva. Permite formalizar los números imprecisos, y también habilita las operaciones aritméticas con tales números borrosos. La teoría de conjuntos clásica puede ser extendida para manejar pertenencias parciales, lo que hace posible expresar conceptos humanos vagos usando conjuntos borrosos y a la vez describir los sistemas de inferencia correspondientes basados en reglas. Otra característica interesante del uso de sistemas borrosos es la capacidad de granular información, es decir, considerar la información a los niveles adecuados de especificidad. Mediante el uso de clusters borrosos de similitud podemos ocultar información no deseada o no utilizada, obteniendo así sistemas en los que la granulación puede ser usada para centrar el análisis en los aspectos de interés para el usuario [Bargiela, 2003]. Los sistemas borrosos pueden ser, entonces, usados para el análisis de conjuntos de datos. La teoría borrosa puede ser combinada con otras técnicas de Minería de Datos. Por ejemplo, en [Jacob, 2003] se aborda el refinamiento de las reglas borrosas de inferencia utilizando algoritmos genéticos. También, las reglas borrosas pueden ser utilizadas para inyectar conocimiento experto en las redes neuronales y los algoritmos de entrenamiento obtenidos pueden entonces ser utilizados para ajustar las funciones de pertenencia de las correspondientes reglas. En [Nauck, 1997] se describen con detalle tales sistemas Neuro-Borrosos.

Por ejemplo, [Bazartseren, 2003] utiliza una aproximación neuro-borrosa para llevar a cabo un estudio comparativo sobre una predicción a corto plazo del nivel del agua usando RNAs, que ha mostrado capacidad ante situaciones con datos escasos, en las que las condiciones están basadas sólo en condiciones hidrológicas agua arriba. En [Juang, 2003] se diseña una red dinámica borrosa basada en algoritmos genéticos con cromosomas de tamaño variable para abordar problemas temporales. Un artículo de investigación realmente avanzado, en el que también se trabaja con análisis de series temporales, es [Oh, 2003] donde se hace uso de redes neuronales polinómicas basadas en neuronas polinomiales borrosas. Aplicaciones más específicas son [Kuncheva, 2000], donde se construye un modelo borroso de concentraciones de metales pesados en la bahía de Liverpool, y [Faye, 2003], quien propone un enfoque completo de la gestión a largo plazo de un modelo de almacenamiento/traslado/distribución en un sistema de recursos hídricos utilizando teoría borrosa.

### c) Algoritmos Genéticos

El aprendizaje automático moderno y el análisis de datos dependen de sofisticadas técnicas de búsqueda. Los sistemas de computación que son capaces de extraer información de grandes cantidades de datos, es decir, de realizar tareas de Minería de Datos, tales como reconocimiento de patrones, clasificación, diagnóstico, etc. y, en general, los sistemas adaptativos y que muestran capacidades de aprendizaje, descansan fundamentalmente en técnicas de búsqueda eficientes y efectivas. Cualquier sistema adaptativo necesita algún mecanismo de búsqueda a la hora de explorar un espacio de características describiendo todos los estados posibles del sistema. Usualmente, el interés está en la búsqueda de estados o configuraciones óptimas o casi óptimas para la aplicación de interés. Por ejemplo, los parámetros que definen (calibran) un modelo particular, las preferencias de hábitat de ciertas especies, los parámetros que describen un ecosistema sostenible y eficientemente controlado, los pesos de una red neuronal para la correcta clasificación de algunos datos, los diámetros y el diseño de trazado de una red de distribución de agua económicamente óptima (y fiable), son unos pocos ejemplos. En general, es necesaria la exploración en espacios multimodales y de alta dimensionalidad. La alta dimensionalidad hace del diseño de técnicas de búsqueda apropiadas un problema realmente complejo. La multimodalidad significa que no hay un solo óptimo global, sino que se pueden encontrar diversos óptimos locales a veces interesantes. En general, encontrar el óptimo global de una función objetivo que tiene muchos grados de libertad y está sujeto a restricciones contradictorias (y frecuentemente subjetivas) es un problema NP-completo, ya que es probable que la función objetivo tenga muchos óptimos locales. Muchos problemas interesantes son de este tipo de problemas de optimización fuerte. Los procedimientos para explorar el espacio de estados deben muestrear este espacio multimodal de forma que haya alguna manera de descubrir soluciones cercanas al óptimo con una alta probabilidad. Además, la técnica asociada debe mostrar una implementación eficiente. Durante las últimas dos décadas, algunos algoritmos que imitan ciertos principios naturales han sido utilizados en diferentes entornos de la ciencia. Entre ellos, los algoritmos genéticos (AG), una subclase de métodos de búsqueda a través de una evolución artificial basada en la selección natural y en mecanismos de genética de poblaciones denominados *programas evolutivos*, [Michalewicz, 1992], son los más populares. Este tipo de búsqueda evoluciona a través de generaciones, mejorando las características de las soluciones potenciales por medio de mecanismos inspirados en la Biología.

La calibración de los parámetros utilizados en el modelo de calidad de agua para la contaminación de cauces receptoras se lleva a cabo en [López, 2001] y [Espert, 1999] utilizando AGs. [Fielding, 1999] presenta varias aplicaciones de la DM, y especialmente técnicas de AG, en la modelación de adecuación de hábitats. [Jeffers, 1999] utiliza algoritmos genéticos para descubrir reglas que describen preferencias en el hábitat de diferentes especies en ríos británicos. El diseño óptimo de las redes de distribución de agua incluyendo su formalización y utilizando AGs se considera en [Wu, 2002] y [Matías, 2004]. Las tendencias actuales se dirigen hacia las combinaciones de RNAs, Lógica Borrosa y AGs, que están mostrando gran fertilidad en muchos problemas. Como ejemplo, en [Oh, 2003] se introduce una categoría general de redes neuro-borrosas que utiliza mecanismos de optimización genética, y se usa como modelo de las emisiones de NO<sub>x</sub> de una planta eléctrica.

### d) Árboles de Decisión y Reglas de Inducción

Son técnicas y métodos capaces de aprender modelos de datos en forma simbólica. Son estructuras jerárquicas que proveen información en forma de condiciones que son más comprensibles para los humanos (que las redes neuronales, etc. a veces llamadas cajas-negras) siéndolo por lo tanto también para los sistemas semi-automáticos que procesan reglas. Así, estas técnicas son, de entre todas

las técnicas de modelación de DM, las más comprensibles y simples de utilizar. Un árbol de decisión es una estructura jerárquica, donde cada nodo interno contiene un test sobre un atributo, cada rama corresponde a un resultado del test, y cada nodo-hoja da una predicción para el valor de la variable de la clase. Los árboles de decisión han sido utilizados durante siglos y son especialmente adecuados para representar procedimientos en muy diversos campos. La principal característica de un árbol de decisión es que una cierta condición conduce a estados mutuamente excluyentes. Para analizar una cierta situación hay que seguir el árbol de manera lógica hasta finalmente alcanzar una sola hoja, acción o decisión a tomar. Por otro lado, las reglas de inducción son una generalización de los árboles de decisión en los que la exhaustividad no es obligatoria para las condiciones. Así que resulta factible el poder elegir más de una regla. Los árboles de decisión pueden ser transformados en reglas de inducción, pero no recíprocamente. Los árboles de decisión son más adecuados para la clasificación. Ya que la clasificación trabaja con conjuntos de clases disjuntas un árbol de decisión conducirá a un ejemplo hacia una única hoja, asignando de este modo una única clase al objeto. Esta propiedad es la idea sencilla que subyace en los primeros algoritmos de aprendizaje para árboles de decisión. Siguiendo una estrategia de divide y vencerás gestionan completamente la clasificación del espacio de acuerdo con una cierta partición. El criterio adoptado para designar cada partición da pie a diferentes algoritmos para construir árboles de decisión. Los más importantes son CART [Breiman, 1984], ID3 [Quinlan, 1983], [Quinlan 1986], C4.5 [Quinlan, 1993], ASSISTANT [Cestnik, 1987]. Las reglas de inducción no siguen necesariamente el criterio de exhaustividad pero en ocasiones dirigen la clasificación de la información de manera más adecuada. Así, en lugar de usar una partición del espacio, las reglas son generadas una detrás de otra mientras se cubren más y más ejemplos. Al final resulta, entonces, posible obtener reglas que se solapan. El denominado algoritmo de cubrimiento es utilizado para generar las reglas. Este algoritmo trabaja como sigue. Primero se construye una regla que clasifica correctamente algunos ejemplos. Entonces, los ejemplos que cumplen la regla son eliminados del conjunto de entrenamiento y el proceso se repite hasta que no quedan más ejemplos. Hay varias implementaciones del algoritmo de cubrimiento. Las más importantes son AQ [Michalski, 1983] [Michalski, 1986], CN2 [Clark, 1989] [Clark, 1991], FOIL [Quinlan 1990] [Quinlan & Cameron-Jones 1993] y FFOIL [Quinlan 1996].

## Software para KDD

La DM está bajo constante evolución. Se presentan nuevas técnicas, las técnicas existentes se mejoran, el mercado del software cambia y aparecen cada día nuevas áreas de aplicación... Con este panorama, es difícil tanto ofrecer una visión absoluta sobre las herramientas existentes como dar criterios para seleccionar una herramienta adecuada para una aplicación dada. Hay varios lugares de Internet donde encontrar información específica y de esta manera ayudar a encontrar la solución correcta al problema de análisis de datos en cuestión. Nos limitamos a citar dos de los más populares.

- <http://www.kdnuggets.com/> (KD significa Descubrimiento del Conocimiento) es una fuente importante de información en Minería de Datos, Minería de la Web, Descubrimiento del Conocimiento, y Temas de Soporte de la Decisión, incluyendo Noticias, Software, Soluciones, Compañías, Trabajos, Cursos, Reuniones, Publicaciones, y más. De especial interés es el enlace dedicado al *software*, un directorio de propósito general de software de DM, y los sub-enlaces *classification* incluyendo modelos de árboles de decisión y redes neuronales y otros planteamientos, y *suites* ofreciendo una amplia visión de los paquetes de software de DM y KD tanto comerciales como gratuitos. No se puede dejar de mencionar *KDnuggets News*, una publicación bi-semanal que ha sido ampliamente

reconocida como el boletín informativo líder en Minería de Datos y Descubrimiento del Conocimiento.

- EvoWeb, <http://evonet.lri.fr/index.php>, website de EvoNet – la Red Europea de Excelencia en Computación Evolutiva – es otro interesante sitio de Internet que promueve la innovación, transmite tecnología y formación y provee información completa principalmente en el campo de la computación evolutiva, aunque también se consideran otras técnicas de DM.

## Conclusiones y futuros desarrollos

Muchos artículos de investigación que usan técnicas de KDD para abordar problemas de muchas áreas y, en particular en el campo de la gestión de los recursos del agua (IWM), han sido escritos principalmente durante los últimos 10 ó 15 años. Estas técnicas han probado ser útiles para los modelos tradicionales o derivados de la teoría pero también para la identificación y aprendizaje de las relaciones y patrones ocultos sobre los datos medidos obtenidos de procesos demasiado complejos (con incertidumbre y no-linealidad altas) para poder ser descritos por modelos basados en la teoría.

La literatura y la experiencia de los autores permiten sintetizar las ventajas de usar las técnicas del KDD para problemas relacionados con el agua en las siguientes ideas:

- Las técnicas del KDD pueden ser utilizadas bien para reemplazar modelos basados en conocimiento o para complementarlos y producir los denominados modelos híbridos.
- Un área de dominio particular se beneficiará de la modelación basada en los datos si:
  - Hay una considerable cantidad de datos disponibles;
  - No hay cambios de consideración en el sistema modelado durante el periodo cubierto por la modelación;
  - Es difícil construir modelos de simulación basados en la física, o que, en casos particulares, no son lo suficientemente adecuados;
  - Es necesario validar los resultados de los modelos de simulación con otros tipos de modelos.
- No es necesario tener un profundo conocimiento del problema bajo consideración, pero la experiencia del modelador es usualmente muy importante, requiriendo una preparación específica: análisis de las relaciones lógicas entre las diferentes variables, elección de esas variables, preparación de la colección de datos, pre-proceso...
- Los modelos del KDD son usualmente más rápidos que los basados en la teoría que utilizan modelos integro-diferenciales tradicionales para obtener soluciones numéricas.

El futuro se encuentra en usar modelos híbridos combinando modelos de diferentes tipos y siguiendo distintos paradigmas de modelación, incluyendo la combinación con modelos basados en la física. Se puede anticipar que la inteligencia computacional (aprendizaje automático) no sólo será usada para construir modelos a partir de los datos, sino también para construir estructuras de modelos

óptimos y adaptados de tales modelos híbridos. Además, la Computación Granular (GC), concerniente al proceso de información sobre entidades complejas, a través de gránulos de información, es un paradigma conceptual emergente que está penetrando numerosos campos de aplicación. A través del uso de las ideas subyacentes en la GC un sistema puede ser observado y analizado en diferentes escalas de tiempo, espacio y complejidad. La granulación sirve [Bargiela 2003] ‘como un eficiente vehículo para subdividir un problema’, ‘ayuda a entender mejor lo esencial en vez de quedar enterrados en detalles innecesarios’, ‘sirve como un mecanismo de abstracción que reduce la carga conceptual’, y ‘es humano-céntrico en el sentido de que el usuario, el diseñador, el desarrollador están en el centro de todas esos esfuerzos’. No es difícil reconocer que esos factores ocurren de manera ostensible en problemas relacionados con IWM. Ser capaces de reconocer los niveles de abstracción adecuados y tener las herramientas disponibles para tratar el problema en cuestión será de suprema importancia para construir cualquier DSS (Sistema de Soporte a la Decisión) útil para la gestión integrada del agua. Como escriben Bargiela y Pedrycz, ‘la computación granular está orientada al conocimiento... Indudablemente, un proceso dirigido hacia el conocimiento se revela como una piedra angular de la minería de datos, de los modelos basados en reglas, de las bases de datos inteligentes, del control supervisado y jerárquico, sólo por nombrar un pocos ejemplos representativos’.

## Agradecimientos

El contenido de este trabajo ha sido parcialmente presentado en las reuniones de la CCMS-NATO de Värskä (Estonia) y Lisboa (Portugal) y ha sido parcialmente desarrollado bajo el abrigo del Plan Nacional de I + D + I (2004-2007) del Ministerio de Ciencia y Tecnología de España (ref.: DPI2004-04330).

## Referencias

- [1] Abbot, M.B., Babovic, V.M., y Cunge, J.A. (2001). Towards the hydraulics of the hydroinformatics era. *Journal of Hydraulic Research*, 39(4), 339-349.
- [2] Anctil, F., Tape, D. G. (2004). An exploration of artificial neural network rainfall-runoff forecasting combined with wavelet decomposition. *J. Environ. Eng. Sci.* 3: S121-S128.
- [3] Baratti, R., Cannas, B., Fanni, A., Pintus, M., Sechi, G. M., Toreno, N. (2003). River flow forecast for reservoir management through neural networks. *Neurocomputing*, 55, 421-437.
- [4] Bardossy, A., Duckstein, L. (1995). *Fuzzy rule-based modeling with application to geophysical, economic, biological and engineering systems*, CRC, London.
- [5] Bargiela, A., Pedrycz, W. (2003). *Granular Computing. An Introduction*. Kluwer Academic Publishers, Boston, Dordrecht, London.
- [6] Baxter, C. W., Smith, D. W., Stanley, S. J. (2004). A comparison of artificial neural networks and multiple regression methods for the analysis of pilot-scale data. *J. Environ. Eng. Sci.* 3: S45-S58.
- [7] Bazartseren, B., Hildebrandt, G., Holz, K.-P. (2003). Short-term water level prediction using neural networks and neuro-fuzzy approach. *Neurocomputing*, 55, 439-450.
- [8] Bhattacharya, B., Lobbrecht, A. H., Solomatine, D. P. (2003). Neural Networks and Reinforcement Learning in Control of Water Systems. *J. Water Res. Plan. and Mgmt.*, ASCE, 129:6, 458-465.



- [9] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*, Wadsworth and Brooks/Cole, Monterey.
- [10] Cichocki, A., Unbehauen, R. (1993). *Neural Networks for Optimization and Signal Processing*. John Wiley & Sons Ltd. & B. G. Teubner, Stuttgart.
- [11] Clark, P., Boswell, R. (1991). Rule induction with CN2: Some recent improvements, in *Proceedings of the Fifth European Working Session on Learning*, pp. 151-163, Springer, Berlin.
- [12] Clark, P., Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, 3(4): 261-283.
- [13] Doyle, P., Lane, J. I., Theeuwes, J. J., Zayatz, L. M. (Eds.), (2001), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier.
- [14] Dzeroski, S. (2001). Data Mining in a nutshell, in Dzeroski, S. and Lvrac, N., (Eds.), *Relational Data Mining*, pp. 3-27. Springer, Berlin.
- [15] El-Din, A. G., Smith, D. W., El-Din, M. G. (2004). Application of artificial neural networks in wastewater treatment. *J. Environ. Eng. Sci.* 3: S81-S95.
- [16] Espert, V., López, P. A. Izquierdo, J. (1999). Fundamentals of a water quality model solution for dissolved oxygen in one-dimensional receiving system. *Numerical Modelling of Hydrodynamic Systems. Proc. of Intl. Workshop*, pp. 444-445.
- [17] Faye, R. M., Sawadogo, S., Lishou, C., Mora-Camino, F. (2003). Long-term fuzzy management of water resource systems. *Applied Mathematics and Computation*, 137, 459-475.
- [18] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery: An overview, in Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds.) *Advances and knowledge Discovery and Data Mining*, pp. 1-34. MIT Press, Cambridge, MA.
- [19] Fielding, A. H. (Ed.) (1999). *Machine Learning Methods for Ecological Applications*. Dordrecht, Netherlands: Kluwer Academic Press.
- [20] Izquierdo, J., Pérez, R., Iglesias, P. L. (2004a). Mathematical Models and Methods in the Water Industry, *Mathematical and Computer Modeling*, 39, 1353-1374.
- [21] Izquierdo, J., (2004b). Detection and identification of anomalies in water distribution systems using neural networks. NATO/CCMS 2nd workshop of the pilot study Integrated Water Management, Génova.
- [22] Jacob, C. (2003). Stochastic Search Methods, in *Intelligent Data Analysis. An Introduction*, Berthold, M, Hand, D. J. (Eds.), pp. 350-401.
- [23] Jeffers, J. N. R. (1999). Genetic Algorithms I, in *Machine Learning Methods for Ecological Applications*, Fielding, A. H. (Ed.), pp- 107-121.
- [24] Juang, Ch. F., (2003). Temporal problems solved by dynamic fuzzy network based on genetic algorithm with variable-length chromosomes. *Fuzzy Sets and Systems*, 142, 199-219.
- [25] Kaufmann A., Gupta, M. M. (1991). *Introduction to fuzzy arithmetics: Theory and Applications*. Van Nostrand Reinhold, New York.
- [26] Keim, D, Ward, M. (2003). Visualization, in Berthold, M, Hand D. J. (Eds.), *Intelligent Data Analysis. An Introduction*, pp. 403-427, Springer, Berlin.
- [27] Kohonen, T. (2001). *Self-Organizing Maps*. Springer-Verlag, Berlin, Heidelberg.
- [28] Kuncheva, L. I., Wrench, J., Jain, L. C., Al-Zaidan, A. S. (2000). A fuzzy model of heavy metal loadings in Liverpool bay. *Env. Modeling and Software*, 15, 161-167.
- [29] Lavkulich, L. M. (2004). Some remarks on the topic 'Environmental Indicators/Human Health', NATO/CCMS pilot study Integrated Water Management, Genova.
- [30] López, P. A. (2001). Metodología para la calibración de modelos matemáticos de dispersión de contaminantes incluyendo regímenes no permanentes. Doctoral Dissertation. Polytechnic University of Valencia (Spain).
- [31] Matías, A. (2004). Diseño de redes de distribución de agua contemplando la fiabilidad, mediante algoritmos genéticos. Doctoral Dissertation. Polytechnic University of Valencia (Spain).

- [32] Michalewicz, Z. (1992). *Genetic algorithms + data structures = evolutionary programs*. Springer-Verlag, New York, Inc., New York, N.Y.
- [33] Michalski, R.S., Larson, J. (1983). Incremental generation of  $vl_1$  hypotheses: the underlying methodology and the description of program aq11, *ISG 83-5*, Dept. of Computer Science, Univ. of Illinois at Urbana-Champaign, Urbana.
- [34] Michalski, R.S., Mozetic, I., Hong, J., Lavrac, N. (1986). The AQ15 inductive learning system: an overview and experiments, in *Proceedings of IMAL 1986*, Orsay, France, Université de Paris-Sud.
- [35] Millot, J., Rodríguez, M. J., Sérodes, J. B. (2002). Contribution of Neural Networks for Modelling Trihalomethanes Occurrence in Drinking Water. *J. Water Res. Plan. and Mgmt.* 128:5, 370-376.
- [36] Moore, R. (1979). *Methods and Applications of Interval Analysis*. Siam Philadelphia.
- [37] Nauck, D., Klawonn, F., Kruse, R. (1997). *Foundations of Neuro-Fuzzy Systems*. John Wiley, New York.
- [38] Oh, S. K., Pedrycz, W. (2004). Self-organizing polynomial neural networks based on polynomial and fuzzy polynomial neurons: analysis and design. *Fuzzy Sets and Systems*, 142, 163-198.
- [39] Oh, S. K., Pedrycz, W., Park, H. S. (2003). Multi-FNN identification based on HCM clustering and evolutionary fuzzy granulation. *Simulation Modelling Practice and Theory* 11, 627-642.
- [40] Panella, M., Rizzi, A., Martinelli, G. (2003). Refining accuracy of environmental data prediction by MoG neural networks. *Neurocomputing*, 55, 521-549.
- [41] Quinlan, J.R. (1983). Learning Efficient Classification Procedures and Their Application to Chess End Games, in Michalski, R.; Carbonell, J.; Mitchell, T. (Eds.) *Machine Learning: An Artificial Intelligence Approach*. Morgan Kaufmann, San Mateo, CA.
- [42] Quinlan, J.R. (1986). Induction of Decision Trees. *Machine Learning*, 1, 81-106.
- [43] Quinlan, J.R. (1990). Learning logical definitions from Relations. *Machine Learning*, 5:239--266.
- [44] Quinlan, J.R. (1993). *C4.5. Programs for Machine Learning*, San Francisco, Morgan Kaufmann.
- [45] Quinlan, J.R. (1996). Learning first-order definitions from relations. *Machine Learning*, 5(3), 239-266.
- [46] Rowland, J., Andrews, W. S., reber, K. A. M. (2004). A neural network approach to selecting indicators for a sustainable ecosystem. *J. Environ. Eng. Sci.* 3: S129-S136.
- [47] Ruck, B. M., Walley, W. J., Hawkes, H. A. (1993). Biological classification of river water quality using neural networks, in *Applications of Artificial Intelligence in Engineering*, Rzevski, G, Pastor, J., Adey, R. A. (Eds.). *Applications and Techiques* 2, 361-372. Sothampton, UK: Computational Mechanics Publications.
- [48] Torra, V. (2003). Trends in Information Fusion in Data Mining, in Torra, V. (Ed.), *Information Fusion in Data Mining*, pp. 1-6, Springer-Verlag, Heidelberg.
- [49] Torra, V., Domingo-Ferrer, J. (2003). Record linkage methods for multidatabase data mining, in Torra, V. (Ed.), *Information Fusion in Data Mining*, pp. 101-132, Springer-Verlag, Heidelberg.
- [50] Tronchi, S., Giona, M., Baratti, R. (2003). Reconstruction of chaotic time series by neural models: a case study. *Neurocomputing* 55, 581-591.
- [51] Tsumoto, S. (2003). Discovery of Temporal Knowledge in Medical Time-Series Databases using Moving Average, Multiscale Matching and Rule Information, in Torra, V. (Ed.), *Information Fusion in Data Mining*, pp. 79-100, Springer-Verlag, Heidelberg.
- [52] U.S. EPA Terms (2000). *Terms of the Environment*. Document order number EPA175B97001, National Service Center for Environmental Publications. Also available at <http://www.epa.org/OCEPaterms/>.
- [53] Walley, W. J., Dzeroski, S. (1996). Biological monitoring: a comparison between Bayesian, neural and machine learning methods of water quality classification, in *Proceedings of the International symposium on Environmental Software Systems*, pp. 229-240. London: Chapman and Hall.
- [54] Walley, W. J., Martin, R. W., O'Connor, M. A. (2000). Self organizing maps for the classification and diagnosis of river quality from biological and environmental data, in *Environmental Software Systems: Environmental Information and Decision Support*, Denzer, R., Swayne, A., Purvis, M., Schimak, G., pp. 27-41. Dordrecht, Netherlands: Kluwer.

- [55] Wu, Z. Y., Simpson, A. R. (2002). A self-adaptative boundary search genetic algorithm and its application to water distribution systems. *J. Hydr. Research*, Vol. 40, No. 2, 191-199.
- [56] Yager, R. R. (2003). Data Mining Using Granular Linguistic Summaries, in Torra, V. (Ed.), *Information Fusion in Data Mining*, pp. 211-229, Springer-Verlag, Heidelberg.
- [57] Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8, 338-353.
- [58] Zadeh, L. A. (1995). Probability Theory and fuzzy logic are complementary rather than competitive. *Technometrics*, 37(3), 271-276.