
Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial
ISSN: 1137-3601
revista@aepia.org
Asociación Española para la Inteligencia Artificial
España

Riquelme, José C.; Ruiz, Roberto; Gilbert, Karina
Minería de Datos: Conceptos y Tendencias
Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, vol. 10, núm. 29, primavera, 2006, pp. 11-18
Asociación Española para la Inteligencia Artificial
Valencia, España

# Minería de Datos: Conceptos y Tendencias

**José C. Riquelme[(1)], Roberto Ruiz[(2)], Karina Gilbert[(3)]**

[(1)] Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
riquelme@lsi.us.es
[(2)] Área de Lenguajes y Sistemas Informáticos
Universidad Pablo de Olavide, Sevilla
rruisan1@upo.es
[(3)] Departamento de Estadística e Investigación Operativa
Universidad Politécnica de Cataluña, Barcelona
karina.gibert@upc.edu

**Resumen**

Hoy en día, la minería de datos (MD) está consiguiendo cada vez más captar la atención de las empresas. Todavía es infrecuente oír frases como "deberíamos segmentar a nuestros clientes utilizando herramientas de MD", "la MD incrementará la satisfacción del cliente", o "la competencia está utilizando MD para ganar cuota de mercado". Sin embargo, todo apunta a que más temprano que tarde la minería de datos será usada por la sociedad, al menos con el mismo peso que actualmente tiene la Estadística. Así que ¿qué es la minería de datos y qué beneficios aporta? ¿Cómo puede influir esta tecnología en la resolución de los problemas diarios de las empresas y la sociedad en general? ¿Qué tecnologías están detrás de la minería de datos? ¿Cuál es el ciclo de vida de un proyecto típico de minería de datos? En este artículo, se intantarán aclarar estas cuestiones mediante una introducción a la minería de datos: definición, ejemplificar problemas que se pueden resolver con minería de datos, las tareas de la minería de datos, técnicas usadas y finalmente retos y tendencias en minería de datos.

**Palabras clave:** Minería de Datos.

## 1. Introducción

La revolución digital ha hecho posible que la información digitalizada sea fácil de capturar, procesar, almacenar, distribuir, y transmitir [10]. Con el importante progreso en informática y en las tecnologías relacionadas y la expansión de su uso en diferentes aspectos de la vida, se continúa recogiendo y almacenando en bases de datos gran cantidad de info rmación.

Descubrir conocimiento de este enorme volumen de datos es un reto en sí mismo. La minería de datos (MD) es un intento de buscarle sentido a la explosión de información que actualmente puede ser almacenada [10].

Hoy en día, los datos no están restringidos a tuplas representadas únicamente con números o caracteres. El avance de la tecnología para la gestión de bases de datos hace posible integrar diferentes tipos de datos, tales como imagen, video, texto, y otros datos numéricos, en una base de datos sencilla, facilitando el procesamiento multimedia. Como resultado, la mezcla tradicional ad hoc de técnicas estadísticas y herramientas de gestión de datos no son adecuadas por más tiempo para analizar esta vasta colección de datos desiguales.

La tecnología de Internet actual y su creciente demanda necesita el desarrollo de tecnologías de minería de datos más avanzadas para interpretar la

información y el conocimiento de los datos distribuidos por todo el mundo. En este siglo la demanda continuará creciendo, y el acceso a grandes volúmenes de datos multimedia traerá la mayor transformación para el global de la sociedad. Por tanto, el desarrollo de la tecnología de minería de datos avanzada continuará siendo una importante área de estudio, y en consecuencia se espera gastar muchos recursos en esta área de desarrollo en los próximos años. Existen diversos dominios donde se almacenan grandes volúmenes de información en bases de datos centralizadas y distribuidas, como por ejemplo librerías digitales, archivos de imágenes, bioinformática, cuidados médicos, finanzas e inversión, fabricación y producción, negocios y marketing, redes de telecomunicación, etc.

Es conocida la frase "los datos en bruto raramente son beneficiosos directamente". Su verdadero valor se basa en: (a) la habilidad para extraer información útil la toma de decisiones o la exploración, y (b) la comprensión del fenómeno gobernante en la fuente de datos. En muchos dominios, el análisis de datos fue tradicionalmente un proceso manual. Uno o más analistas familiarizados con los datos, con la ayuda de técnicas estadísticas, proporcionaban resúmenes y generaban informes. En efecto, el analista hacía de procesador de preguntas sofisticado. Sin embargo, tal enfoque cambió como consecuencia del crecimiento del volumen de datos. Cada vez es más común encontrarse con bases de los datos con un número de ejemplos del orden de $10^9$ o superior y $10^3$ dimensiones. Cuando la escala de manipulación de datos, exploración e inferencia va más allá de la capacidad humana, se necesita la ayuda de las tecnologías informáticas para automatizar el proceso.

Todo apunta a la necesidad de metodologías de análisis inteligente de datos, las cuales puedan descubrir conocimiento útil de los datos. El término KDD (inciales de *Knowledge Discovery in Databases*), acuñado en 1989 se refiere a todo el proceso de extracción de conocimiento a partir de una base de datos y marca un cambio de paradigma en el que lo importante es el conocimiento útil que seamos capaces de descubrir a partir de los datos. En el primer estado del arte sobre el área, [Fayy96] se dice:

> *"La mayoría de los trabajos previos en KDD, se centraban en [...] la etapa de Minería de Datos. Sin embargo, los otros pasos son de considerable importancia para el éxito de las aplicaciones de KDD en la práctica."*

lo que claramente apunta a la importancia de incluir en la metodología el preproceso de los datos, o la formalización del conocimiento descubierto.

En realidad, los términos MD y KDD son a menudo confundidos como sinónimos. En general se acepta que la MD es un paso particular en el proceso consistiendo en la aplicación de algoritmos específicos para extraer patrones (modelos) de los datos. Otros pasos en el proceso KDD, son la preparación de los datos, la selección y limpieza de los mismos, la incorporación de conocimiento previo, y la propia interpretación de los resultados de minería. Estos pasos aplicados de una manera iterativa e interactiva aseguran que un conocimiento útil se extraiga de los datos.

Las tareas propias de la fase de minería de datos pueden ser descriptivas, (i.e. descubrir patrones interesantes o relaciones describiendo los datos), o predictivas (i.e. clasificar nuevos datos basándose en los anteriormente disponibles). En otras palabras, es un campo interdisciplinar con el objetivo general de predecir las salidas y revelar relaciones en los datos [10]. Para ello se utilizan herramientas automáticas que (a) emplean algoritmos sofisticados para descubrir principalmente patrones ocultos, asociaciones, anomalías, y/o estructuras de la gran cantidad de datos almacenados en los data warehouses u otros repositorios de información, y (b) filtran la información necesaria de las grandes bases de datos.

El concepto de KDD se ha desarrollado, y continúa desarrollándose, desde la intersección de la investigación de áreas tales como bases de datos, aprendizaje automático, reconocimiento de patrones, estadística, teoría de la información, inteligencia artificial, razonamiento con incertidumbre, visualización de datos y computación de altas prestaciones. Los sistemas KDD incorporan teorías, algoritmos, y métodos de todos estos campos. Una buena perspectiva general del KDD se puede encontrar en la referencias [8] y [10].

Otro concepto relacionado con el KDD es el de Data warehousing [4,8,10,12], que se refiere a las tendencias actuales en la recolección y limpieza de datos transaccionales para dejarlos disponible para el análisis y la toma de decisiones. La MD debe trabajar mano a mano con los almacenes de datos, sobre todo en los casos de volumnes de datos muy grandes o de inter-relaciones entre los datos complejas, es decir, que no puedan ser expresadas en una tabla plana. El KDD se centra en el proceso

global de descubrir conocimiento de grandes volúmenes de datos, incluyendo el almacenaje y acceso a tales datos, escalado de algoritmos a bases de datos masivas, interpretación y visualización de resultados, y el modelado y soporte de la interacción general hombre máquina. Un almacenaje eficiente de los datos, y por lo tanto su estructura, es muy importante para su representación acceso. Los conocimientos de las tecnologías de comprensión modernas deberían ser utilizados para explorar como estos mecanismos de almacenaje pueden ser mejorados.

Como hemos señalado el concepto de MD también se solapa con los conceptos de aprendizaje automático y de estadística. En general, la estadística es la primera ciencia que historicamente extrae información de los datos básicamente mediante metodologías procedentes de las matemáticas. Cuando se empezó a usar los ordenadores como apoyo para esta tarea surgió el concepto de *Machine Learning*, traducido como Aprendizaje Automático. Posteriormente con el incremento del tamaño y con la estructuración de los datos es cuando se empieza a hablar de MD.

De esta manera la MD hace hincapié en:

- la escalabilidad del número de atributos y de instancias
- algoritmos y arquitecturas (proporcionando la estadística y el aprendizaje automático los fundamentos de los métodos y las formulaciones), y
- la automatización para manejar grandes volúmenes de datos heterogéneos.

En el resto del artículo consideraremos la minería de datos desde diversas perspectivas. En la próxima sección se proporcionan las bases del descubrimiento de conocimiento y la minería de datos. En la sección 3 se enumeran algunas de las aplicaciones más frecuentes en los negocios. En la sección 4 se describe brevemente las técnicas más utilizadas. Las tendencias en la minería de datos se muestran en la sección 5. Y para finalizar se extraerán las principales conclusiones.

## 2. Extracción de conocimiento en bases de datos y minería de datos

El descubrimiento de conocimiento en bases de datos (KDD) se define como el proceso de identificar patrones significativos en los datos que sean válidos, novedosos, potencialmente útiles y comprensibles para un usuario [4,8,10,12]. El proceso global consiste en transformar información de bajo nivel en conocimiento de alto nivel. El proceso KDD es interactivo e iterativo conteniendo los siguientes pasos:

1. Comprender el dominio de aplicación: este paso incluye el conocimiento relevante previo y las metas de la aplicación.
2. Extraer la base de datos objetivo: recogida de los datos, evaluar la calidad de los datos y utilizar análisis exploratorio de los datos para familiarizarse con ellos.
3. Preparar los datos: incluye limpieza, transformación, integración y reducción de datos. Se intenta mejorar la calidad de los datos a la vez que disminuir el tiempo requerido por el algoritmo de aprendizaje aplicado posteriormente.
4. Minería de datos: como se ha señalado anteriormente, este es la fase fundamental del proceso. Está constituido por una o más de las siguientes funciones, clasificación, regresión, clustering, resumen, recuperación de imágenes, extracción de reglas, etc.
5. Interpretación: explicar los patrones descubiertos, así como la posibilidad de visualizarlos.
6. Utilizar el conocimiento descubierto: hacer uso del modelo creado.

El paso fundamental del proceso es el señalado con el número 4. A continuación se comentan brevemente las tareas más comunes de la minería de datos, con un ejemplo de uso.

- Clasificación: clasifica un dato dentro de una de las clases categóricas predefinidas. Responde a preguntas tales como, ¿Cuál es el riesgo de conceder un crédito a este cliente? ¿Dado este nuevo paciente qué estado de la enfermedad indican sus análisis?
- Regresión: el propósito de este modelo es hacer corresponder un dato con un valor real de una variable. Responde a cuestiones como ¿Cuál es la previsión de ventas para el mes que viene? ¿De qué depende?
- Clustering: se refiere a la agrupación de registros, observaciones, o casos en clases de objetos similares. Un cluster es una colección de registros que son similares entre sí, y distintos a los registros de otro cluster. ¿Cuántos tipos de clientes vienen a mi

negocio? ¿Qué perfiles de necesidades se dan en un cierto grupo de pacientes?

- Generación de reglas: aquí se extraen o generan reglas de los datos. Estas reglas hacen referencia al descubrimiento de relaciones de asociación y dependencias funcionales entre los diferentes atributos. ¿Cuánto debe valer este indicador en sangre para que un paciente se considere grave? ¿Si un cliente de un hipermercado compra pañales también compra cerveza?
- Resumen o sumarización: estos modelos proporcionan una descripción compacta de un subconjunto de datos. ¿Cuáles son las principales características de mis clientes?
- Análisis de secuencias: se modelan patrones secuenciales, como análisis de series temporales, secuencias de genes, etc. El objetivo es modelar los estados del proceso, o extraer e informar de la desviación y tendencias en el tiempo. ¿El consumo de energía eléctrica de este mes es similar al del año pasado? Dados los niveles de contaminación atmosférica de la última semana cuál es la previsión para las próximas 24 horas.

Como resumen podríamos señalar que el rápido crecimiento del interés en la minería de datos es debido (i) al avance de la tecnología de Internet y a la gran participación en aplicaciones multimedia en este dominio, (ii) a la facilidad en la captura de datos y el abaratamiento de su almacenaje, (iii) a compartir y distribuir los datos en la red, junto con el aumento de nuevas bases de datos en los repositorios, (iv) al desarrollo de algoritmos de aprendizaje automático robustos y eficientes para procesar estos datos, (v) al avance de las arquitecturas de las computadoras y la caída del coste del poder computacional, permitiendo utilizar métodos computacionalmente intensivos para el análisis de datos, (vi) la falta de adaptación de los métodos de análisis y consulta convencionales a nuevas formas de interacción y finalmente (vii) a la potencia que este tipo de análisis vienen mostrando como herramientas de soporte a la toma de decisiones frente a realidades complejas (viii) fuerte presión de los productos comerciales disponibles.

## 3. Aplicaciones de la minería de datos

Algunas de las tareas importantes de la minería de datos incluyen la identificación de aplicaciones para las técnicas existentes, y desarrollar nuevas técnicas para dominios tradicionales o de nueva aplicación, como el comercio electrónico y la bioinformática.. Existen numerosas áreas donde la minería de datos se puede aplicar, prácticamente en todas las actividades humanas que generen datos:

- Comercio y banca: segementación de clientes, previsión de ventas, análisis de riesgo.
- Medicina y Farmacia: diagnostico de enfermedades y la efectividad de los tratamientos.
- Seguridad y detección de fraude: reconocimiento facial, identificaciones biométricas, accesos a redes no permitidos, etc.
- Recuperación de información no numerica: minería de texto, minería web, búsqueda e identificación de imagen, video, voz y texto de bases de datos multimedia.
- Astronomía: identificación de nuevas estrellas y galaxias.
- Geología, minería, agricultura y pesca: identificación de áreas de uso para distintos cultivos o de pesca o de explotación minera en bases de datos de imágenes de satelites
- Ciencias Ambientales: identificación de modelos de funcionamiento de ecosistemas naturales y/o artificiales (p.e. plantas depuradoras de aguas residuales) para mejorar su observación, gestión y/o control.
- Ciencias Sociales: Estudio de los flujos de la opinión pública. Planificación de ciudades: identificar barrios con conflicto en función de valores sociodemográficos.

En la actualidad se puede afirmar que la MD ha demostrado la validez de una primera generación de algoritmos mediante diferentes aplicaciones al mundo real. Sin embargo estas técnicas todavía están limitadas por bases de datos simples, donde los datos se describen mediante atributos numéricos o simbólicos, no conteniendo atributos de tipo texto o imágenes, y los datos se preparan con una tarea

concreta en mente. Sobrepasar este límite será un reto a conseguir.

Señalemos por último que existen cientos de productos de minería de datos y de compañías de consultoría. KDNuggets (kdnuggets.com) tiene una lista de estas compañías y sus productos en el campo de la minería de datos. Pueden resaltarse por su mayor expansión las siguientes: SAS con SAS Script y SAS Enterprise Miner; SPSS y el paquete de minería Clementine; IBM con Intelligent Miner; Microsoft incluye características de minería de datos en las bases de datos relacionales; otras compañías son Oracle, Angoss y Kxen. En la línea del software libre Weka [13] es un producto con mayor orientación a las técnicas provenientes de la IA, pero de fuerte impacto.

## 4. Técnicas usadas por la minería de datos

La Minería de Datos se podría abstraer como la construcción de un modelo que ajustado a unos datos proporciona un conocimiento.
Por tanto podemos distinguir dos pasos en una tarea de MD, por un lado la elección del modelo y por otro el ajuste final de éste a los datos.
La elección del modelo viene determinada básicamente por dos condicionantes: el tipo de los datos y el objetivo que se quiera obtener. Así por ejemplo no sería apropiado aplicar regresión a unos datos constituidos por texto o modelos basados en distancia a datos simbólicos.
En cuanto a la relación modelo-objetivo, la literatura presenta un catalogo de distintos modelos para los diferentes objetivos. Así, si se tiene un problema de clasificación se utilizarán máquinas de vectores soporte o árboles de decisión, si es un problema de regresión se pueden usar árboles de regresión o redes neuronales, si se desea hacer clustering se puede optar por modelos jerárquicos o interrelacionados, etc.
También es importante en esta elección el nivel de comprensibilidad que se quiera obtener del modelo final, ya que hay modelos fáciles de "explicar" al usuario como por ejemplo las reglas de asociación y otros que entrañan claras dificultades como las redes neuronales o los vectores soporte.
El segundo paso consiste en realizar una "fase de aprendizaje" con los datos disponibles para ajustar el modelo anterior a nuestro problema particular. Así si tenemos una red neuronal habrá que definir su arquitectura y ajustar los valores de los pesos de sus conexiones. Si vamos a obtener una recta de regresión hay que hallar los valores de los coeficientes y si usamos los k-vecinos más cercanos necesitamos fijar una métrica y k, etc.
Esta fase de aprendizaje ajusta el modelo buscando unos valores que intenten maximizar la "bondad" del mismo. Esta cuestión nos vuelve a plantear dos problemas: uno ¿Cómo se define la bondad de un modelo para unos datos? Y dos, ¿Cómo realizar esa búsqueda?
Respecto a la primera, normalemente todo modelo debe venir acompañado por una función de adaptación que sea capaz de medir el ajuste (en inglés se emplea el concepto de fitness function). Esto es fácil en numerosos casos, por ejemplo en problemas de clasificación o regresión, sin embargo puede plantear serios retos en otros como el clustering.
Además relacionado con este concepto se encuentra un fenómeno conocido como sobreajuste, es decir, que se "aprendan" los datos de entrenamiento pero no se generalice bien para cuando vengan nuevos casos. Exisen numerosos estudios en la literatura sobre distintas formas de separar convenientemente datos de entrenamiento de datos de prueba [1,2,5].
En cuanto a la búsqueda de los valores que maximizan la bondad, se dispone de un importante número de posibilidades: desde la clásicas procedentes del análisis matemático cuando la función de bondad se conoce completamente hasta las heurísticas que proporciona la investigación operativa, pasando por técnicas como los Algoritmos Evolutivos (sin duda una de las más presentes en la literatura), búsquedas tabú, búsquedas dispersas, etc.
Debido a que esta búsqueda u optimización está presente en todos los procesos de MD, a menudo se confunden, pudiendo presentarse por ejemplo los algoritmos evolutivos como un modelo de MD, cuando realmente es una técnica que se puede usar para ajustarlo.
Por último, otro factor a tener en cuenta junto con los anteriores es el tratamiento que deseamos dar a la incertidumbre que el propio modelo genera. Por ejemplo, supongamos un modelo basado en reglas que define una así:

*Si* $x \in [1.4, 3.4]$ *entonces* $y \in [-2.1, 6.5]$

¿Qué podríamos afirmar si $x$ vale 3.5 ó 1.3? ¿y si vale 3.6 ó 1.2? Este razonamiento lleva a usar lógicas distintas de la clásica como son la lógica borrosa o difusa (fuzzy) o los menos conocidos *rough sets*. Relacionado con esto aparece un último concepto: *softcomputing*, para referirse al conjunto de técnicas computacionales (lógica borrosa, razonamiento probabilístico, algoritmos evolutivos, …) que posibilitan las herramientas de aprendizaje.

Softcomputing se refiere a la característica de imprecisión o incertidumbre que acompaña por su propia naturaleza al concepto de MD [15].
Todos los conceptos presentados en esta sección (modelo, tipo de datos, lógica, función de bondad y técnica de búsqueda) convenientemente hibridizados han dado lugar a infinidad de metodologías en MD. Así es fácil encontrar referencias a "redes neuronales borrosas para datos numéricos entrenadas mediante algoritmos evolutivos", "clustering mediante *rough sets* aplicando una búsqueda dispersa", "definición de una métrica para búsqueda tabú de reglas que clasifiquen texto", etc.

## 5. Retos y tendencias de la minería de datos

Existen algunos retos que superar antes de que la minería de datos se convierta en una tecnología de masas [9,14]. Señalamos en este epígrafe algunos de los retos actualmente planteados.

**Aspectos metodológicos**: Sería muy útil la existencia de una API Standard, de forma que los desarrolladores puedan integrar sin dificultad los resultados de los diversos algoritmos de minería. Esto podría facilitar también la tarea de automatizar y simplificar todo el proceso, integrando aspectos como muestreo, limpieza de datos, minería, visualización, etc.. En este mismo sentido sería deseable que los productos de minería de datos estuvieran orientados al programador para fomentar su uso y ampliación. Sería asimismo necesario unificar la teoría sobre la materia: así se puede observar que los estados del arte no son generalizables, no existe un estándar para la validación de resultados y, en general, la investigación se realiza demasiado aislada. Asimismo se necesitaría mejorar la formación en esta área entre los titulados universitarios, que sería la mejor manera de expandir su uso, y finalmente, sigue siendo un asunto pendiente la integración del conocimiento del dominio en el algoritmo, y viceversa, es decir, mejorar la interpretabilidad y facilidad de uso del modelo hallado.

**Escalabilidad**: la escalabilidad de la minería de datos hacia grandes volúmenes de datos es y será siempre una de las tendencias futuras, ya que el volumen de información que se ha de tratar crece de manera exponencial, con lo que los avances en esta área quedan siempre superados por las necesidades crecientes. Datos con miles de atributos es ya algo habitual, pero es probable que las técnicas no estén preparadas aún para centenares de miles o incluso millones de características. Dentro de esta línea también se localiza la minería de *data streams* de muy alta velocidad con posibles cambios de estructura, dimensión o modelo de generación dinámico durante la fase de entrenamiento. Esto obliga a terner un modelo de conocimiento en todo momento.

**Simulación, integración en la toma de decisiones y minería de datos**: los modelos extraídos para un ámbito de interés de una organización. Básicamente se trata de utilizar las salidas de unos modelos como entradas de otros y maximizar el beneficio del conjunto de modelos. Además, pueden añadirse al modelo global restricciones de valores máximos o mínimos (saturación), etc. Las técnicas tradicionales de combinación de modelos [6,8] no pueden aplicarse directamente. Las técnicas de simulación en minería de datos (véase el capítulo 18 de [8]), más relacionadas con el problema de una maximización global no han recibido la atención suficiente desde el área de la minería de datos. La obtención de modelos que globalmente se comporten bien y que se mantengan dentro de unas restricciones, requiere no sólo de matrices de costes y de técnicas como el análisis ROC [7], sino de otros tipos de métricas y técnicas para el aprendizaje y la evaluación. La predicción local más idónea para un problema puede implicar la elección de una menos idónea para otro, mientras puede existir una decisión global mejor. Si bien este tipo de decisiones globales han sido estudiadas por la teoría de la decisión [3] y por el área de planificación en inteligencia artificial [11], esta interrelación entre modelos predictores, su aprendizaje y problemas de optimización y planificación no ha sido estudiada a fondo.

**Minería para datos con una estructura compleja**: en numerosas ocasiones los datos procedentes de aplicaciones del mundo real no tienen una representación directa en forma de una única tabla, sino que deben ser representadas mediante estructuras jerárquicas (árboles), interrelacionadas (grafos), conjuntos, etc. Por lo tanto, el reto que se lanza a la comunidad científica que investiga en aprendizaje automático y minería de datos, es el de adaptar o proponer nuevas técnicas que permitan trabajar directamente con este tipo de representaciones. En este campo también entraría la minería de datos distribuida, donde los datos no se encuentran en una única localización sino como es cada vez más habitual en una red de computadores. Un caso particular sería la minería de datos multimedia, para datos que integran voz, imágenes, texto, video, y que, debido a la complejidad de los datos, el volumen y el gran abanico de aplicaciones

posibles constituye un reto en la actualidad.

Otros temas que se están abordando y donde se debe profundizar son: la comprensibilidad de los patrones extraídos; potenciar las aplicaciones en campos nuevos como privacidad, anti-terrorismo, crisis energética, medioambiente, bioinformática; asegurar la privacidad e integridad de los datos que son sometidos a minería; datos no balanceados entre las distintas clases; datos sensitivos al coste, no sólo en el error al asignar una clase sino en la obtención de los atributos; datos en secuencia y series temporales cada vez más utilizadas; etc.

Podemos concluir señalando que la minería de datos se considera todavía un nicho y un mercado emergente. Una de las razones es que la mayoría de los paquetes de minería de datos están dirigidos a expertos, y esta cuestión no facilita su uso por los usuarios. Se piensa que en los próximos años habrá más desarrolladores de aplicaciones comerciales de gestión que sean capaces de integrar en éstas módulos de minería de datos. Con ello se conseguirá extender y generalizar su uso a usuarios de los más diversos campos de la actividad humana.

## 6. Conclusiones

La minería de datos es un área de estudio científico con grandes expectativas para la comunidad investigadora, principalmente por las expectativas de transferencia a la sociedad que plantea. Desde hace más de 50 años se han publicado infinidad de artículos en conferencias y revistas destacadas sobre la materia. Sin embargo, queda por delante un campo fértil y prometedor con muchos retos en investigación. Este artículo ha proporcionado una introducción al descubrimiento de conocimiento y la minería de datos. Se han descrito las principales posibilidades que la minería de datos proporciona, así como una relación de las principales metodologías usadas. Además se han resaltado diferentes dominios de aplicación y los principales retos y tendencias en investigación.

## Agradecimientos

Los autores agradecen a los profesores Francisco Herrera de la U. de Granada y José Hernández-Orallo de la U.P. de Valencia las sugerencias aportadas para la redacción de este artículo. 'Inteligencia Artificial' es una publicación periódica distribuida por la Asociación Española para la Inteligencia Artificial (AEPIA).

## Referencias

[1] E. Alpaydin. Combined 5x2 cv f test for comparing supervised classification learning algorithms. Neural Computation, 11: 1885-1892, 1999.

[2] C. Ambroise and G. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA 99, pages 6562-6566, 2002.

[3] Y. Ben-Haim. Information-Gap Decision Theory. Academic Press, 2001.

[4] M.J.A. Berry and G.S. Linoff. Data mining techniques for marketing, sales, and Customer Relationship Management. Wiley Publishing, 2004.

[5] T. Dietterich. Approximate statistical test for comparing supervised classification learning algorithms. Neural Computation, 10 (7): 1895-1924, 1998.

[6] V. Estruch, C. Ferri, J. Hernández-Orallo and M. J. Ramírez-Quintana. Bagging Decision Multi-trees. Multiple Classifier Systems, pages 41-51, 2004.

[7] P. Flach, H. Blockeel, C. Ferri, J. Hernández-Orallo and J. Struyf. Decision Support for Data Mining: Introduction to ROC analysis and its applications. Book chapter in Data Mining and Decision Support, Kluwer, 2003.

[8] J. Hernández-Orallo, M. J. Ramírez-Quintana and C. Ferri. Introducción a la Minería de Datos. Prentice Hall / Addison-Wesley, 2004.

[9] H. Kargupta, A. Joshi, K. Sivakumar and Y.Yesha. Data mining: next generation challenges and future directions. MIT/AAAI Press, 2004.

[10] S. Mitra and T. Acharya. Data mining: multimedia, soft computing and bioinformatics. John Wiley & Sons, 2003.

[11] S. J. Russell and P. Norvig. Artificial Intelligence: A Modern Approach. Prentice Hall, 2002.

[12] Z. Tang and J. MacLennan. Data Mining with SQL Server 2005. Wiley Publishing, 2005.

[13] Witten, IH and Frank, E: "Data Mining: Practical Machine Learning Tools and Techniques", 2nd Edition. Morgan Kaufmann, 2005

[14] Q. Yang and X. Wu. Challenging Problems in Data Mining Research ICDM 2005 http://www.cs.ust.hk/~qyang.

[15] L.A. Zadeh. What is Soft Computing?. Soft Computing, 1(1), 1, 1997.