Revista de la Facultad de Ingeniería Industrial
Vol. (7) 2: pp. 83-86 (2004) UNMSM
ISSN: 1560-9146 (impreso) / ISSN: 1810-9993 (electrónico)

DATA MINING Y EL DESCUBRIMIENTO DEL CONOCIMIENTO

Recepción: Noviembre de 2004 / Aceptación: Diciembre 2004

[1] Violeta Valcárcel Asencios

RESUMEN

El presente artículo enfatiza el uso del Data Mining para el descubrimiento del conocimiento, a fin de contribuir en la toma de decisiones tácticas y estratégicas en una organización proporcionando un sentido automatizado para la generación de conocimiento. Se incluyen las técnicas, el poder predictivo de los modelos estadísticos y el aporte a las diferentes ramas de la investigación.

Palabras Claves: Minería de datos. Descubrimiento del conocimiento. Modelos predictivos.

DATA MINING AND KNOWLEDGE DISCOVERY
ABSTRACT

The present article emphasizes the use of data mining for the discovery of knowledge, with the purpose of contributing in taking tactical decisions and strategies within an organization providing an automated sense to generate knowledge. Techniques, the predictive power of statistical models and the contribution of the various fields of the research have been included.

Key words: Data mining. Knowledge discovery. Predictable models.

(1) Licenciada en Estadística. Estudiante de la Unidad de Postgrado de la Facultad de Ingeniería Industrial, UNMSM. Cursa estudios de Maestría.
E-mail: postind@unmsm.edu.pe

INTRODUCCIÓN

En los últimos años, ha existido un gran crecimiento en nuestras capacidades de generar y colectar datos, debido básicamente al gran poder de procesamiento de las máquinas como a su bajo costo de almacenamiento.

Sin embargo, dentro de estas enormes masas de datos existe una gran cantidad de información "oculta", de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información. El descubrimiento de esta información "oculta" es posible gracias a la Minería de Datos (Data Mining), que entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad, pero es el descubrimiento del conocimiento (KDD, por sus siglas en inglés) que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados.

Así el valor real de los datos reside en la información que se puede extraer de ellos, información que ayude a tomar decisiones o mejorar nuestra comprensión de los fenómenos que nos rodean. Hoy, más que nunca, los métodos analíticos avanzados son el arma secreta de muchos negocios exitosos. Empleando métodos analíticos avanzados para la explotación de datos, los negocios incrementan sus ganancias, maximizan la eficiencia operativa, reducen costos y mejoran la satisfacción del cliente.

DESCUBRIMIENTO DEL CONOCIMIENTO (KDD)

Según Molina (2001) lo define como *«la extracción no trivial de información potencialmente útil a partir de un gran volumen de datos, en el cual la información está implícita, donde se trata de interpretar grandes cantidades de datos y encontrar relaciones o patrones, para conseguirlo harán falta técnicas de aprendizaje, estadística y bases de datos»*.

Las tareas comunes en KDD son la inducción de reglas, los problemas de clasificación y clustering, el reconocimiento de patrones, el modelado predictivo, la detección de dependencias, etc.

Los datos recogen un conjunto de hechos (una base de datos) y los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese subconjunto), tal como se muestra en la Figura 1. El KDD involucra un proceso iterativo e interactivo de búsqueda de modelos, patrones o parámetros, los cuales descubiertos han de ser válidos, novedosos para el sistema y potencialmente útiles.

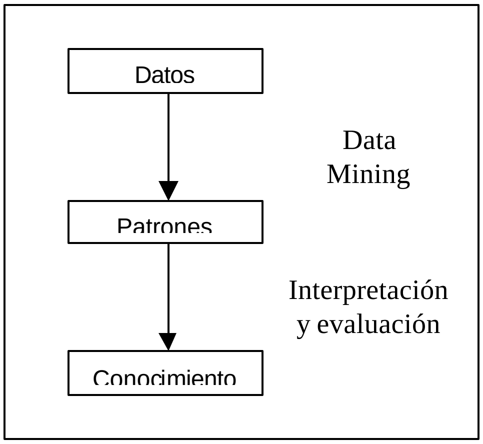

Figura 1. Proceso de descubrimiento del conocimiento (KDD)

El objetivo final de todo esto es incorporar el conocimiento obtenido en algún sistema real, así como tomar decisiones a partir de los resultados alcanzados o, simplemente, registrar la información conseguida y suministrársela a quien esté interesado.

Mientras el descubrimiento de la máquina confía solamente en métodos autónomos para el descubrimiento de la información, KDD típicamente combina métodos automatizados con la interacción humana para asegurar resultados exactos, útiles, y entendibles.

Existen diferentes métodos que son clasificados como las técnicas de KDD, entre ellos los métodos cuantitativos, los probabilísticos y los estadísticos. Se tienen métodos que utilizan las técnicas de visualización y, métodos de clasificación como la clasificación de Bayesian, lógica inductiva, descubrimiento de modelado de datos y análisis de decisión. Otros métodos incluyen la desviación y tendencia al análisis, algoritmos genéticos, redes neuronales y los métodos híbridos que combinan dos o más técnicas.

DATA MINING

Según Molina (2001) menciona que la *Data Mining* se refiere al proceso de extraer conocimiento de bases de datos. Su objetivo es descubrir situaciones anómalas y/o interesantes, tendencias, padrones y secuencias en los datos.

La *Data Mining* es una etapa dentro del proceso completo del descubrimiento del conocimiento, este intenta obtener patrones o modelos a partir de los datos recopilados. Decidir si los modelos obtenidos son útiles o no suele requerir una valoración subjetiva por parte del usuario. Los algoritmos de *data mining* suelen tener tres componentes:

1. El modelo, que contiene parámetros que han de fijarse a partir de los datos de entrada.
2. El criterio de preferencia, que sirve para comparar modelos alternativos.
3. El algoritmo de búsqueda, que viene a ser como cualquier otro programa de inteligencia artificial (IA).

El criterio de preferencia suele ser algún tipo de heurística y los algoritmos de búsqueda empleados suelen ser los mismos que en otros programas de inteligencia artificial. Las principales diferencias entre los algoritmos de *data mining* se hallan en el modelo de representación escogido y la función del mismo, es decir según el objetivo perseguido.

HERRAMIENTAS DE DATA MINING

Las herramientas de *data mining* empleados en el proceso de KDD se pueden clasificar en dos grandes grupos:

- Técnicas de verificación, en las que el sistema se limita a comprobar hipótesis suministradas por el usuario.

- Métodos de descubrimiento, en los que se han de encontrar patrones potencialmente interesantes de forma automática, incluyendo en este grupo todas las técnicas de predicción.

El resultado obtenido con la aplicación de algoritmos de *data mining* pertenecientes al segundo grupo, el de técnicas de descubrimiento, pueden ser de carácter descriptivo o predictivo. Las predicciones sirven para prever el comportamiento futuro de algún tipo de entidad mientras que una descripción puede ayudar a su comprensión.

La aplicación de técnicas de *data mining* en grandes bases de datos persiguen los siguientes resultados:

1. **Clasificación:** Se trata de obtener un modelo que permita asignar un caso de clase desconocida a una clase concreta (seleccionada de un conjunto redefinido de clases), como son los árboles de clasificación (CART), cuyos resultados pueden expresarse mediante reglas ejecutables directamente del SQL o el método de Bayesiano.

2. **Regresión:** Se persigue la obtención de un modelo que permita predecir el valor numérico de alguna variable (modelos de regresión logística).

3. **Agrupamiento (clustering):** Hace corresponder cada caso a una clase, con la peculiaridad de que las clases se obtienen directamente de los datos de entrada utilizando medidas de similaridad. Es decir, agrupan a los datos bajo diferentes métodos y criterios. Las técnicas más usadas son las clásicas (distancia mínima) y las redes neuronales (método de Kohonen o método de Neural-Gas).

4. **Resumen:** Se obtienen representaciones compactas para subconjuntos de los datos de entrada (análisis interactivo de datos, generación automática de informes, visualización de datos).

5. **Modelado de Dependencias:** Se obtienen descripciones de dependencias existentes entre variables. El análisis de relaciones (por ejemplo las reglas de asociación), en el que se determinan relaciones existentes entre elementos de una base de datos, podría considerarse un caso particular de modelado de dependencias.

6. **Análisis de Secuencias:** Se intenta modelar la evolución temporal de alguna variable, con fines descriptivos o predictivos (redes neuronales multicapas).

LOS TIPOS DE MODELOS ESTADÍSTICOS Y LA DATA MINING

Como en todo lo producido por la máquina, las predicciones estadísticas fabricadas por la *data mining* deben ser inspeccionadas por especialistas en el área, de manera a comprender y verificar lo que fue producido.

Asimismo, es importante mencionar que existe un término medio entre la claridad del modelo y su poder de predicción. Mientras más sencilla sea la forma del modelo, más fácil será su comprensión, pero tendrá menor capacidad para tomar en cuenta dependencias sutiles o demasiado variadas (no lineales). La Figura 2 ilustra una representación de dicho término medio.

Los árboles de decisión y las bases de reglas se interpretan muy fácilmente pero no conocen sino los límites "duros" de comparación en niveles de decisión Si-No. Adolecen de una fineza predictiva.

Las evaluaciones por puntuación, lineales o con funciones logísticas son un poco más "sofisticadas" pero como sólo adicionan resultados no pueden dar cuenta de relaciones multivariables.

Las redes neuronales tienen la virtud de adaptarse a valores bastante indefinidos e incluso ausentes, pero son difíciles en el momento de inspeccionar. Sólo las predicciones realizadas pueden ser inspeccionadas y visualizadas. Sin embargo, una buena herramienta de visualización le da la posibilidad al usuario de reconstruir el "razonamiento" de la red neuronal. Según cual sea el precio a pagar, y una vez que se haya establecido la confianza en la herramienta establecida, el usuario notará, la mayoría de las veces, que la

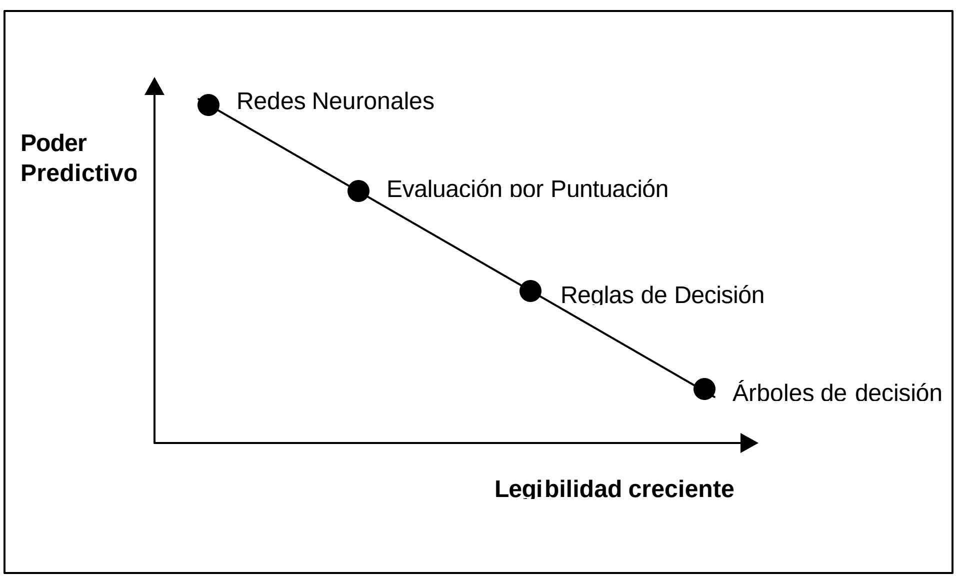

Figura 2. Representación del témino medio en el modelamiento predictivo de data mining

pérdida parcial de comprensión será más que compensada por la calidad de las predicciones.

APLICACIONES DE DATA MINING

En la actualidad, existe una gran cantidad de aplicaciones, en áreas tales como:

- **Astronomía:** clasificación de cuerpos celestes.
- **Aspectos climatológicos:** predicción de tormentas, etc.
- **Medicina:** caracterización y predicción de enfermedades, probabilidad de respuesta satisfactoria a tratamiento médico.
- **Industria y manufactura:** diagnóstico de fallas.
- **Mercadotécnia:** identificar clientes susceptibles de responder a ofertas de productos y servicios por correo, fidelidad de clientes, selección de sitios de tiendas, afinidad de productos, etc.
- **Inversión en casas de bolsa y banca (credit scoring, redes neuronales o regresión logística):** análisis de clientes, aprobación de prestamos, determinación de montos de crédito, etc.
- **Detección de fraudes y comportamientos inusuales:** telefónicos, seguros, en tarjetas de crédito, de evasión fiscal, electricidad, etc.
- **Análisis de canastas de mercado:** para mejorar la organización de tiendas, segmentación de mercado (clustering) determinación de niveles de audiencia de programas televisivos.
- **Normalización automática:** de bases de datos.

CONCLUSIONES

La capacidad para almacenar datos ha crecido en los últimos años a velocidades exponenciales. En contrapartida, la capacidad para procesarlos y utilizarlos no ha ido a la par. Por este motivo, el data mining se presenta como una tecnología de apoyo para explorar, analizar, comprender y aplicar el conocimiento obtenido usando grandes volúmenes de datos. Sin embargo, en su aplicación sólo se obtienen patrones que no sirven de gran cosa mientras no se les encuentre significado y su valor real reside en la información que se puede extraer de ellos: información que ayude a tomar decisiones o mejorar la comprensión de los fenómenos que nos rodean.

Las técnicas estadísticas son fundamentales a la hora de validar hipótesis y analizar datos, por lo cual la estadística desempeña un papel muy importante en KDD. La Estadística proporciona herramientas para cuantificar adecuadamente la incertidumbre resultante de la inferencia de patrones a partir de datos particulares. Las herramientas de KDD pretenden automatizar (hasta donde se pueda) el proceso completo de análisis de datos.

La *data mining* y el descubrimiento del conocimiento (KDD) contribuye a la toma de decisiones tácticas y estratégicas, proporcionando un sentido automatizado para la generación de conocimiento y por ende a la toma acertada de decisiones y su aplicación es amplia en las diferentes ramas de la investigación.

BIBLIOGRAFÍA

1. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (1996). *Advances in Knowledge and Data Mining*. MIT Press. Massachussets, USA.

2. Lyn, Thomas; Edelman, David; Crook, Jonathan. (2002). *Credit Scoring and its Applications*. SIAM. Filadelfia, USA.

3. Molina, Luis Carlos. (2000). *Torturando los Datos Hasta que Confiesen*. Departamento de Lenguajes y Sistemas Informáticos, Universidad Politécnica de Cataluña. Barcelona, España.

4. Urdaneta, Elymir. (2001). *El Data Mining*. Universidad de Caracas. Venezuela.

5. Zavala, Mauricio. (2004). *Modelamiento Predictivo*. En: http://www.gm.et/bluetech/edicion11.3/Datamining.