Universidad de Matanzas "Camilo Cienfuegos"
Facultad de Ingenierías Química – Mecánica

MONOGRAFÍA

ELEMENTOS DE MINERÍA DE DATOS

Jorge Enrique Albelo Mengana
Ramón Quiza Sardiñas
Marcelino Rivas Santana

Departamento de Ingeniería Mecánica

Noviembre, 2007

Elementos de Minería de Datos.

Jorge Enrique Albelo Mengana,
Ramón Quiza Sardiñas,
Marcelino Rivas Santana.

En: CD Monografías 2007.

Autopista a Varadero, km 3½, Matanzas, CP 44740, Cuba.
http://www.umcc.cu

Tabla de Contenido.

1. Introducción.

El concepto de Minería de Datos (MD) apareció hace más de 10 años. El interés en este campo y su explotación en diferentes especialidades (negocios, finanzas, ingeniería, banca, salud, sistemas de energía, meteorología...), se ha incrementado recientemente debido a la combinación de diferentes factores, los cuales incluyen:

- El surgimiento de gran cantidad de datos (en el orden de los terabytes) debido a la medición y la recopilación de datos automática, registros digitales, archivos centralizados de datos y simulaciones de software y hardware.

- El abaratamiento de los costos de los medios de almacenamiento.

- El surgimiento y rápido crecimiento del manejo de sistemas de bases de datos.

- Los avances en la tecnología computacional tal como las computadoras de alta velocidad y las arquitecturas paralelas.

- Los desarrollos continuos en técnicas de aprendizaje automático.

- La posible presencia de incertidumbre en los datos (ruido, outliers, información perdida, etc.)

El propósito general de la minería de datos es procesar la información de la gran cantidad de datos almacenados o que se puedan generar, y

desarrollar procedimientos para manejar los datos y tomar futuras decisiones [Arévalo y Pérez, 2002].

Generalmente, una de las primeras tareas en el proceso de la minería de datos consiste en resumir la información almacenada en la base de datos, con el fin de comprender bien su contenido. Esto se realiza por medio de análisis estadísticos o técnicas de búsqueda y reporte. Las operaciones más complejas consisten en la identificación de modelos para predecir información acerca de objetos futuros.

Un paradigma de especial importancia en la identificación de modelos son los llamados métodos de aprendizaje. En el aprendizaje supervisado (*supervised learning*), también conocido como "aprendizaje con profesor", para cada entrada (*input*) de los objetos de aprendizaje, la salida (*output*) deseada es conocida e utilizada en el aprendizaje. En los métodos de aprendizaje sin supervisión (*unsupervised learning*) o "aprendizaje por observación", la salida correspondiente a cada entrada no es suministrada o conocida del todo, y el método "aprende" por sí solo de los valores de los atributos de entrada.

Hasta ahora, los mayores éxitos en Minería de Datos se pueden atribuir directa o indirectamente a avances en bases de datos (un campo en el que los ordenadores superan a los humanos). No obstante, muchos problemas de representación del conocimiento y de reducción de la complejidad de la búsqueda necesaria (usando conocimiento *a priori*) están aún por resolver. Ahí reside el interés que ha despertado el tema entre investigadores de todo el mundo.

2. Definición de Minería de Datos.

Hoy en día, la cantidad de información que ha sido almacenada en las bases de datos excede nuestra habilidad para reducir y analizar los datos sin el uso de técnicas de análisis automatizadas. Muchas bases de datos comerciales transaccionales y científicas crecen a una proporción fenomenal.

La Minería de Datos se ha definido como "el proceso no trivial de identificación en los datos de patrones válidos, nuevos, potencialmente útiles, y finalmente comprensibles" [Molina y García, 2004]. O sea, que es la extracción de información en grandes bases de datos.

Su nombre deriva de la analogía existente entre buscar dicha información valiosa y minar una montaña para encontrar un yacimiento de metales preciosos, ya que ambos procesos requieren examinar una inmensa cantidad de material o investigar inteligentemente hasta concretar la búsqueda.

La Minería de Datos está formada por un conjunto de tecnologías que ayudan a los usuarios a enfocar sus objetivos sobre la información más importante de sus fuentes de datos. Las herramientas de Minería de Datos pueden responder preguntas que generalmente demandan demasiado tiempo, encontrando información que ni un profesional experto podría hallar porque, frecuentemente, se encuentras fuera de sus expectativas.

La Minería de Datos es, en realidad, un proceso iterativo de descubrimiento de patrones y tendencias dentro de los datos, a través de

métodos automáticos, tanto manuales como, más generalmente, semiautomáticos, y que no serían necesariamente revelados por otros métodos tradicionales de análisis. Las herramientas de Minería de Datos exploran las Bases de Datos en busca de patrones ocultos, permitiendo a partir de éstos predecir las futuras tendencias y comportamientos de información nueva.

Como es de esperar, los patrones descubiertos deben ser significativos en el hecho que conduzcan a alguna ventaja y ésta, generalmente, sea económica.

Otras definiciones de Minería de Datos, reportadas en la literatura son:

- Es la extracción no trivial de información implícita, desconocida previamente, y potencialmente útil desde los datos [Piatesky-Shapiro y Frawley, 1991].

- Es el proceso de extracción y refinamiento de conocimiento útil desde grandes bases de datos [Simoudis y Livezey-Kerber, 1996].

- Es el proceso de extracción de información previamente desconocida, válida y procesable desde grandes bases de datos para luego ser utilizada en la toma de decisiones [Cabena *et al.*, 1997].

- Es la exploración y análisis, a través de medios automáticos y semiautomáticos, de grandes cantidades de datos con el fin de descubrir patrones y reglas significativos [Berry y Linoff, 1997].

- Es el proceso de planteamiento de distintas consultas y extracción de información útil, patrones y tendencias previamente desconocidas desde grandes cantidades de datos posiblemente almacenados en bases de datos [Thuraisingham,1999].

- Es el proceso de descubrir modelos en los datos [Witten y Frank 2000].

Como se puede ver, todas las definiciones se mueven, con mayor o menor amplitud, alrededor de una idea común.

La terminología Minería de Datos es usada comúnmente por los estadísticos, analistas de datos, y por la comunidad de administradores de sistemas informáticos como todo el proceso del descubrimiento, mientras que el término Adquisición del Conocimiento en Bases de Datos (*Knowledge Discovery in Databases*, KDD) es más utilizado por los especialistas en Inteligencia Artificial.

3. Necesidad de la Minería de Datos

Existen extensos volúmenes de datos almacenados en fuentes de información, los cuales se acumulan bajo la creencia que alguien, en algún momento los utilizará. Sin embargo, crece progresivamente la diferencia entre generación de datos y entendimiento de éstos: como el volumen de datos aumenta, el número de personas que entienden estos datos desafortunadamente disminuye.

Debido a que la información oculta en los datos es útil y generalmente no se encuentra en forma explícita para tomar ventaja de ésta y en algunos

casos, los datos no se pueden analizar por métodos estadísticos estándar [Pighin, 2001], porque pueden existir valores perdidos o incompletos, ruidos y valores extraños, o bien, los datos pueden estar en forma cualitativa y no cuantitativa. Además, en ciertas situaciones, el acceso a los datos no es sencillo.

4. Estructura del Proceso de Adquisición del Conocimiento en Bases de Datos.

La Adquisición de Conocimiento en Bases de Datos (*Knowledge Discovery in Databases*, KDD) se ha definido como "el proceso no trivial de identificación en los datos de patrones válidos, nuevos, potencialmente útiles, y finalmente comprensibles" [Molina y García, 2004] en la Fig. 1 se podrá observar el proceso completo de extracción de información, que se encarga además de la preparación de los datos y de la interpretación de los resultados obtenidos.

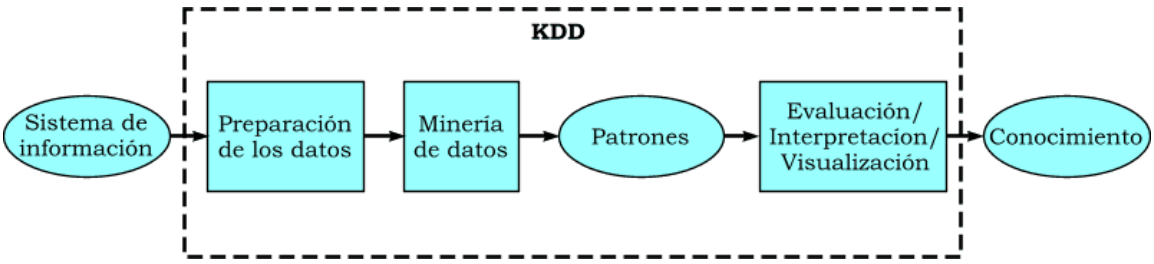

Fig. 1 – Diagrama de bloques del proceso de Adquisición de Conocimiento en Bases de Datos.

5. Objetivo de la Minería de Datos.

El proceso de Minería de Datos tiene como objetivo fundamental descubrir patrones y tendencias en complejas fuentes de datos [Pighin, 2001]. Una vez, que se identifica un patrón particular, el proceso de descubrimiento finaliza y este patrón se convierte en un patrón conocido. Por lo tanto, la Minería de Datos no comprende aquellas aproximaciones analíticas que buscan conjuntos de datos a partir de patrones conocidos, así mismo, las técnicas que requieren implementación de reglas, casos de entrenamiento preestablecidos o aprendizaje automático supervisado son útiles pero no constituyen el proceso de Minería de Datos.

6. Aprendizaje Automático *vs*. Minería de Datos.

El aprendizaje puede definirse de acuerdo al siguiente planteamiento: "un ser aprende, cuando cambia sus comportamientos en la forma que logre la mejor capacidad en el futuro". Este concepto une al aprendizaje más con la capacidad que con el conocimiento, es decir que uno puede evaluar lo aprendido al observar el comportamiento actual y compararlo con el comportamiento anterior.

Por otro lado, el aprendizaje es muy diferente del entrenamiento, porque el aprendizaje implica pensar, implica tener propósitos, implica tener intenciones. El aprendizaje sin objetivos es entrenamiento propiamente dicho.

Por último, el aprendizaje puede resumirse como: "adquisición del conocimiento y habilidad de usarlo ventajosamente" [Pighin, 2001].

La Minería de Datos abarca, más bien, un sentido práctico y no teórico combinando la intervención humana con técnicas de aprendizaje automático. Es una herramienta que ayuda a interpretar los datos, permite encontrar patrones y hacer predicciones a partir de éstos.

7. Consideraciones previas para aplicar MD.

Para explicar de forma efectiva la Minería de Datos deben tenerse en cuenta las siguientes consideraciones:

- Economía, o sea, que el costo de implementación sea menor que las ganancias provocadas por las mejoras obtenidas, produciéndose así un retorno de la inversión.

- Rapidez, ya que generalmente, se espera obtener resultados dentro de un período de tiempo razonable. Si éste se hace muy extenso debería retornarse al inicio y realizar los cambios que se consideren necesarios.

- Accesibilidad de los datos. No es imprescindible un acceso *on-line* a las fuentes de datos, ya que la Minería de Datos no se realiza en tiempo real, pero sí se hace necesario el acceso a toda la información para la ejecución del análisis.

- Factibilidad de la implementación del sistema: Generalmente, nunca se utiliza una simple aplicación, sino más bien una combinación de técnicas y metodologías.

- Efectividad de la toma de decisiones. Como regla general, nunca una herramienta sólo proporcionará la solución buscada, sino simplemente ayuda a encontrarla, como tal lo indica la palabra herramienta. Encontrar la solución al problema propuesto y la total responsabilidad de la toma de decisiones se deposita sobre el o los profesionales que realizan el análisis.

8. Campos de Aplicación de la Minería de Datos.

En puridad, cualquier problema para el que existan datos históricos almacenados es susceptible a ser tratado mediante técnicas de Minería de Datos [Aluja. 2001]. Dentro de sus campos de aplicación más comunes, se incluyen las actividades gubernamentales, los bancos y demás entidades financieras, las industrias manufactureras de diverso tipo, las telecomunicaciones, los servicios, y el comercio.

En la figura 2 se puede observar la distribución de la Minería de Datos por áreas de aplicación. Como se puede ver la manufactura, seguida de la informática, tienen la delantera en este sentido [Pighin, 2001].

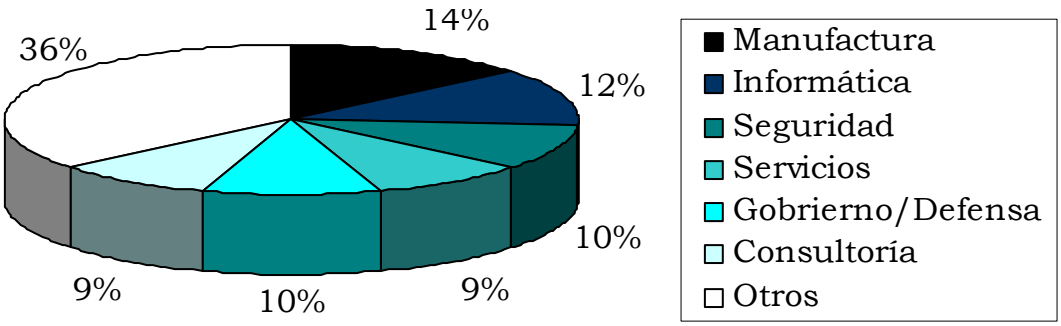

Fig. 2 – Áreas de aplicación de la Minería de Datos.

9. Software disponible para el análisis de datos.

Existen diversos programas para minería de datos. Entre otros, se detacan:

- *KnowledgeSeeker* (Angoss Software International). Herramienta interactiva de clasificación. Basada en los algoritmos de árboles de decisión CHAID y XAID. Se ejecuta sobre plataformas Windows y UNIX .

- *DataCruncher* (DataMind): Herramienta de minería de datos para clasificación y clustering. Basada en Tecnología de agentes de redes (ANT Agent Network Technology). La aplicación servidor se ejecuta sobre UNIX y Windows NT; la aplicación cliente en todas las plataformas Windows.

- *Intelligent Miner* (IBM): Soporta múltiples operaciones de data minino en un entrono cliente servidor. Utiliza redes de neuronas, árboles de inducción y varias técnicas estadísticas. Trabaja sobre clientes Windows, OS/2 y X-Windows, y servidores AIX (incluyendoSP2), OS/400 y OS/390.

- *Clamentine* (Integral Solutions): Herramienta con un entrono de trabajo que soporta todo el proceso de minería de datos. Ofrece árboles de decisión, redes de neuronas, generación de reglas de asociación y características de visualización. Se ejecuta sobre VMS, UNIX o Windows NT.

- *Alice* (Isoft S.A.): Herramienta de escritorio para data minino interactivo. Se basa en tecnología de árboles de decisión. Se ejecuta sobre plataformas Windows.

- *Decisión Series* (NeoVista Software): Herramientas para múltiples operaciones de minería de datos para el desarrollo de modelos basados en servidores. Proporciones algoritmos de redes de neuronas, árboles y reglas de inducción, clustering y análisis de asociaciones. Trabaja sobre sistemas UNIX mono o multi-procesadores de HP y Sun. Accede sólo a ficheros planos, aunque posiblemente las últimas versiones ya trabajaran contra bases de datos relacionales.

- *Pilot Discovery Server* (Pilot Software): Puntos Clave: Herramienta para clasificación y predicción. Basada en la tecnología de árboles de decisión CART. Trabaja sobre UNIX y Windows NT.

- *SAS Solution for Data Mining* (SAS Institute): Un gran número de herramientas de selección, exploración y análisis de datos para entornos cliente-servidor. Las opciones de minería de datos incluyen: aplicaciones de redes de neuronas, de árboles de decisión y herramientas de estadística. Aplicaciones portables para un gran número de entornos PC, UNIX y mainframes.

- *MineSet* (Silicon Graphics): Paquete de herramientas para minería de datos y visualización. Proporciona algoritmos para la generación de reglas para clasificación y asociaciones. Trabaja sobre plataformas SGI bajo IRIS.

- *SPSS* (SPSS): Herramientas de escritorio para clasificación y predicción, clustering, y un gran rango de operaciones estadísticas. Proporciona una herramienta de redes neuronales además de productos de análisis estadístico. SPSS para Windows y Neural Connection son productos que trabajan en modo monopuesto en plataformas Windows.

- *Syllogic Data Mining Tool* (Syllogic): Herramienta con entorno de trabajo multi-estratégico con interface visual. Soporta análisis de árboles de decisión, clasificación k-vecino más próximo, y análisis de clustering y asociaciones por k-means. Trabaja sobre Windows NT y en estaciones UNIX con uno o varios procesadores.

- *Darwin* (Thinking Machines): Herramientas de desarrollo de minería de datos de tipo cliente-servidor para la construcción de modelos de clasificación y predicción. La construcción de modelos utiliza algoritmos de redes de neuronas, árboles de inducción y k-vecino más próximo. Trabaja sobre plataformas Sun de Solaris, AIX de IBM y SP2, con clientes Motif. También existen versiones cliente que trabajan sobre Windows.

- *WEKA* (University of Waikato): Herramientas de escritorio para clasificación y predicción, clustering, y un gran rango de operaciones estadísticas. Proporciona una herramienta de redes de neuronas además de productos de análisis estadístico. WEKA se distribuye como software de libre distribución desarrollado en Java para plataformas Windows y Linux.

10. La entrada al proceso de Minería de Datos

Los datos como tales no pueden ser manejados directamente por Minería de Datos, sino es necesario modelarlos llevándolos a un formato tal que sí pueda ser empleado. El desarrollo de dicho modelo es decisivo, ya que determina los tipos de resultados que se pueden obtener. Este modelado de datos, asume una estructura orientada a objetos, donde la información está representada por objetos, sus atributos descriptivos y las relaciones entre las clases de objetos.

10.1. Clases de Objetos.

Las clases de objetos se consideran como categorías conceptuales. Por ejemplo: personas, lugares, direcciones, y más. La decisión de designar a un campo de variables como una clase de objeto es completamente arbitraria, pero críticamente importante.

Para escoger cuales variables serán clases de objetos, generalmente se prefiere entidades tangibles, pero pueden ser, en cambio, abstractas e incluir estados, fechas, valores, actividades y más.

10.2. Objetos.

Los objetos dentro de una clase son las distintas entidades que ésta contiene. Por ejemplo, en una clase *Persona*, pueden existir tres objetos llamados *Pedro Pérez González*, *María Sánchez Díaz y Carlos Herrera Reyes*. O sea, que la clase es el tipo de datos, mientras que los objetos son los diferentes datos de este tipo que existen. Son dos conceptos que no deben confundirse.

10.3. Atributos.

Los atributos definen los comportamientos de una clase de objeto ya determinada como tal. También ayudan a diferenciar los objetos individuales dentro de una clase. El valor de un atributo es una medida de la cantidad que dicho atributo posee en un objeto. Para caracterizar un objeto cualquiera, cada atributo deber tener un único valor.

En la siguiente figura, se muestra la relación entre una clase (Persona) y sus atributos (Nombre, Edad y Sexo).

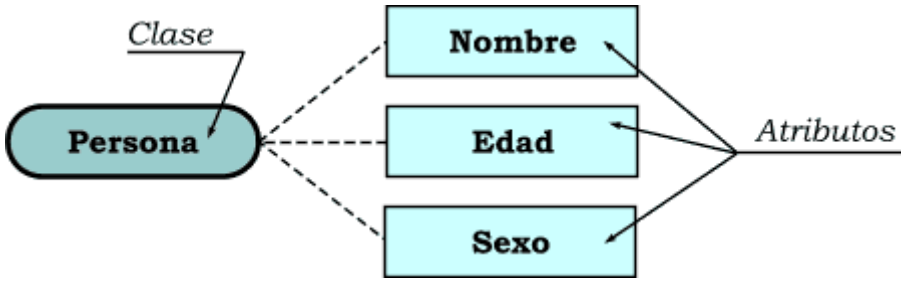

Fig. 3 – Relación clase-atributos.

Para la clase anterior, pueden existir los siguientes objetos, cada uno de los cuales tiene sus respectivos valores para cada uno de los atributos.

Tabla 1 – Ejemplo de objetos con los valores de sus atributos.

Clase	Atributos		
	Nombre	Edad	Sexo
Objetos	Pedro Pérez González	35	Masculino
	María Sánchez Díaz	18	Femenino
	Carlos Herrera Reyes	24	Masculino

Si en el set de datos hay más de un valor, no se puede determinar cual es el correcto. Esto generalmente ocurre cuando dichos valores se toman en distintos intervalos de tiempo. Para estos atributos que cambian con el tiempo, debe plantearse un modelo diferente y esta situación se conoce como "análisis basado en estados" (*state-based analysis*).

Existen casos donde ciertas variables pueden cumplir distintas funciones dentro del modelo, por ejemplo, un campo de datos se puede elegir como una clase de objeto o un atributo de una clase de objeto o ambos según el perfil del análisis.

10.4. Tipos de Atributos.

Los atributos pueden ser:

- Nominales: No tienen un significado numérico concreto. Se subclasifican en:

  - Categóricos: Pueden tomar un número finito de valores, que no tienen ninguna relación de ordenamiento entre ellos. Ejemplo: Marca de la cerveza preferida: Cristal, Bucanero, Beck's…

  - Enumerados: Los valores que pueden tomar, aunque no son numéricos, guardan cierto orden entre sí, por lo que pueden ponerse en correspondencia con los números naturales. Ejemplo: Calidad del servicio: Excelente (5), Buena (4), Regular (3), Mala (2), Pésima (1).

- Booleanos: Son aquellos que solo pueden tener dos valores (Verdadero ó Falso, Sí ó No, 0 ó 1).

♦ Ordinales: Son valores numéricos ya sean discretos o continuos.

Hay que tener presente que esta es una clasificación convencional, por lo que puede diferir de un autor a otro o entre los diferentes programas.

10.5. Relaciones.

Las relaciones enlazan dos clases de objetos según alguna característica en común. Se pueden suponer como clases de objeto, pudiéndole asignárseles atributos tales como fecha, actividades, estados, etc. El valor de cada atributo es único para cada conexión particular, cumpliéndose la regla de un valor por atributo. Se pueden establecer muchas relaciones entre par de objetos, pero son completamente independientes entre sí.

Cuando se construye el modelo, se debe incluir la información que es relevante para el análisis, es decir, decidir qué parte de los datos incluir y cuales omitir, así, las clases de objetos, atributos y relaciones elegidos dependen de los objetivos del problema. Sin embargo, hay que tener presente que durante el proceso de modelado, se puede retroceder y reestructurar el modelo para incluir componentes diferentes del conjunto de datos original.

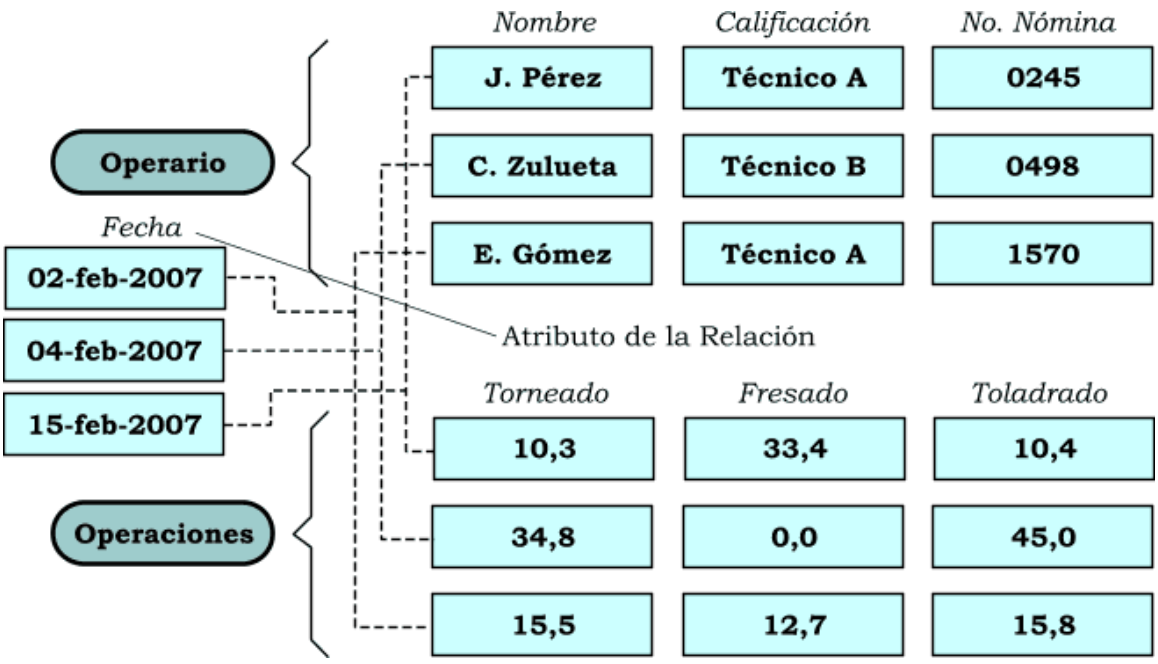

Fig. 4 – Relaciones.

11. El Acceso a la Información.

11.1 Acceso a los Datos.

En esta etapa del proceso de MD, el objetivo es la extracción de los datos de las fuentes originales, y para ello, la condición fundamental es la disponibilidad de los mismos. No siempre el acceso es libre, en ciertas ocasiones, está limitado por la existencia de protección de propiedad literaria, o políticas de seguridad o bien se necesitan procedimientos especiales o protocolos para el acceso.

11.2. Extracción y Tests.

Para llevar a cabo la extracción se utiliza como método la generación de una serie de pruebas que se aplican a las fuentes de datos, permitiendo el acceso buscado. Al confeccionar las pruebas, debe considerarse que la acumulación de gran cantidad de valores tiene un alto costo computacional, para ello existen técnicas de filtrado y segmentación que permiten refinar la extracción. Durante el proceso de extracción debe asegurarse siempre la integridad y consistencia de los datos.

El análisis comienza de forma proactiva, donde se realiza la extracción de un subconjunto de datos que es una muestra representativa del conjunto de datos inicial. Este proceso de extracción es un "corte proactivo" de la información. Aunque si bien, este modo de análisis no se caracteriza por definir los objetivos previamente, es necesario especificar alguna estructura para poder determinar que datos se incluyen en el subconjunto.

Una vez identificado algún objeto de interés en el "corte proactivo", el análisis pasa al modo reactivo, a fin de disponer de toda la información adicional con respecto a dicho objeto y utilizando un lazo de retroalimentación como forma de refinar la investigación sobre los datos.

11.3. Transferencia de Datos

Cuando se habla de extracción, debe tenerse en cuenta el lugar de origen y de destino de los datos, es decir, la transferencia de éstos desde la plataforma principal al ambiente de MD. Esta transferencia no necesariamente se realiza en tiempo real.

Los datos se pueden convertir en archivos de determinado formato, es decir, cualquier tipo de información usada por MD se establece dentro de depósitos temporales donde se puede procesar por software de MD. Estos depósitos temporarios pueden ser: bases de datos, archivos de texto plano u hojas de cálculo.

11.4. Integración de los Datos

Los datos que se usan en un análisis no necesariamente provienen de las mismas fuentes, generalmente, las fuentes y tipos de información pueden ser ilimitados y por lo tanto, luego de acceder a los datos es necesario integrarlos. La idea de integración se conoce como "Almacén de Datos" (Data Warehousing) y la tendencia hacia éste es un reconocimiento de que la información fragmentada puede tener un gran valor cuando se la reúne e integra, por eso, comúnmente se dice que el Almacén de Datos es un precursor de MD.

Es preferible siempre extraer los datos de bases de datos establecidas, pero durante el análisis pueden aparecer formatos no estándares, tales como:

- Texto libre:
  - Encuentran su mayor utilidad con conjuntos de datos pequeños.
  - Estos textos se pueden condensar y resumir.
  - Es un formato difícil de integrar con otros tipos de datos.
- Tablas:
  - Proveen mejores mecanismos para presentar y analizar la información.
  - Permiten ordenar los tipos de datos similares en determinadas zonas para su rápida detección.

- Poseen la capacidad de combinar información de varias fuentes.
- Pero, no transfieren grandes cantidades de datos.

♦ Otros formatos (gráficas, imágenes, fotos, videos, sonidos).

11.5. Normalización de Datos.

Para realizar la integración de los datos de varias fuentes, es necesario que dichos datos estén normalizados: Todas las unidades de medida deben llevarse a una misma escala. Debe asegurarse una terminología consistente. Los tipos de datos similares convienen representarse juntos. Es beneficioso aplicar técnicas de reducción de datos, eliminando información duplicada.

11.6. Limpieza de los Datos.

En general, cualquier base de datos contiene datos inconsistentes, incompletos o erróneos, los cuales pueden ocurrir por varias razones, tales como, caracteres transformados y/o mal deletreados en la entrada de datos, datos perdidos, formatos incompatibles, datos ingresados incorrectamente en pantallas de entrada, etc.

11.7. Valores Perdidos.

Se asume comúnmente, que un valor perdido es un valor no conocido. Pero pueden existir muchas causas para que esto ocurra, por lo tanto, la responsabilidad de decidir si el valor perdido es significativo o no, si es posible, cae sobre alguien bien familiarizado con la información.

11.8. Valores Erróneos.

Valores incorrectos pueden ocurrir cuando éstos cambian de forma insignificante, ya que cualquier perturbación pequeña en el deletreado de un valor da como resultado múltiples representaciones de la misma entrada.

Pueden existir errores no en el deletreado, sino cuando se crean diferentes valores para un mismo objeto.

Se hace importante la limpieza de los datos de manera de asegurar la consistencia y falta de ambigüedad de los mismos. Pueden aplicarse rutinas de limpieza con propósitos diferentes: eliminar los caracteres aislados (iniciales del segundo nombre), eliminar sufijos y prefijos (Sr., Ing., Dr.), etc.

Luego pueden aplicarse métodos más sofisticados como algoritmos o rutinas automáticas donde el objetivo es siempre reducir los conjunto de datos eliminando duplicaciones.

11.9. Información en Formato de Texto.

Algunas fuentes de datos comprenden solo texto con un arreglo consistente de la información, pero en la mayor parte de los casos, dicho formato consistente no existe. En tales situaciones, pueden emplearse distintas aproximaciones, por ejemplo, algunas toman el contenido de un documento y lo separan en un conjunto de estructuras con indicadores de referencia. Al solicitar una palabra o clave determinada, se pueden encontrar de forma automática todas las ocurrencias de éstos términos

dentro de los documentos, produciendo un conjunto de palabras que se clasificarán según diversos factores, tal como, frecuencia de ocurrencia. A veces, los resultados de dicha búsqueda son muy extensos y se pueden resumir sobre la base de un valor o vector que se crea en función de los contenidos verdaderos del documento. Así, cualquier otro documento que posea valores similares, se supone que incluye contenidos semejantes.

11.10. Análisis de Características Estructurales.

El análisis en este contexto se orienta a la agrupación donde las representaciones visuales pueden transmitir, por sí solas, grandes cantidades de información al ser simplemente examinadas.

Por la ubicación de los objetos en la representación, pueden detectarse fácilmente patrones importantes como también características inusuales, las cuales transfieren información acerca de datos que exceden las condiciones límites mediante la representación panorámica de los datos. Pueden localizarse fácilmente aquellos valores alejados de la representación principal. Si tales valores superan condiciones límites o posibles se consideran como erróneos o malos y pueden eliminarse.

11.11. Datos Perdidos

Utilizando la alternativa de agrupación de datos, la representación permite detectar inmediatamente aquellos datos perdidos, lo cual sería tedioso de concretar utilizando un formato tabular.

11.12. Patrones Anómalos.

En algunas situaciones, los conjuntos de datos se representan como eventos que suceden en un orden particular, y si dicho orden se altera, se produce una anomalía que puede señalar patrones importantes.

12. Técnicas de Minería de Datos.

De acuerdo con la bibliografía, se propone una clasificación de las técnicas útiles en minería de datos encontradas en las diferentes fuentes consultadas (ver Tabla 2). La clasificación se basa en la tarea para la cual son útiles cada una de las técnicas y algoritmos encontrados.

Tabla 2. Clasificación de las técnicas de MD.

Tarea	Técnicas	
	De inteligencia artificial	Estadísticas
Clasificación	- Redes neuronales. - Árboles de decisión. - Inducción de reglas. - Programación lógica inductiva.	- Análisis discriminante
Agrupamiento	- Redes neuronales. - Inducción de reglas. - Modelos gráficos probabilísticos. - Hipergráficos.	- Análisis por agrupamiento.

Tabla 2. Clasificación de las técnicas de MD (cont.).

Tarea	Técnicas	
	De inteligencia artificial	Estadísticas
Dependencia	- Análisis de varianza. - Análisis de regresión.	- Modelos gráficos probabilísticos. - Programación lógica inductiva. - Inducción de ecuación. - Inducción de reglas.
Series	- Suavizado de curvas.	- Redes neuronales. - Inducción de reglas.

13. Tareas de Minería de Datos dentro del KDD.

Varios autores coinciden con este orden de tareas de la MD salvo algunas particularidades de procesos específicos [Martínez y Alcántar, 2003; Ferri, 2004].

13.1. Clasificación.

Se trata de obtener un modelo que permita asignar un caso de clase desconocida a una clase concreta (seleccionada de un conjunto predefinido de clases). Dividiendo un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible uno del otro, donde la distancia está medida con respecto a

la(s) variable(s) específica(s) que se están tratando de predecir. [Martínez y Alcántar, 2003; Ferri, 2004].

13.2. 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 similitud. Dividiendo un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible uno del otro, donde la distancia está medida con respecto a todas las variables disponibles [Ferri, 2004]. Permite la identificación de tipologías o grupos donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros grupos. Así se puede segmentar el colectivo de clientes, el conjunto de valores e índices financieros, el espectro de observaciones astronómicas, el conjunto de zonas forestales, el conjunto de empleados y de sucursales u oficinas, etc. La segmentación está teniendo mucho interés desde hace ya tiempo dadas las importantes ventajas que aporta al permitir el tratamiento de grandes colectivos de forma pseudo particularizada, en el más idóneo punto de equilibrio entre el tratamiento individualizado y aquel totalmente masificado. Las herramientas de segmentación se basan en técnicas de carácter estadístico, de empleo de algoritmos matemáticos, de generación de reglas y de redes neuronales para el tratamiento de registros. Para otro tipo de elementos a agrupar o segmentar, como texto y documentos, se usan técnicas de reconocimiento de conceptos. Esta técnica suele servir de punto de partida para después hacer un análisis de clasificación sobre los clusters.

La principal característica de esta técnica es la utilización de una medida de similitud que, en general, está basada en los atributos que describen a los objetos, y se define usualmente por la proximidad en un espacio multidimensional. Para datos numéricos, suele ser preciso preparar los datos antes de realizar MD sobre ellos, de manera que en primer lugar se someten a un proceso de estandarización. Una de las técnicas empleadas para conseguir la normalización de los datos es utilizar la medida *z* (*z-score*) que elimina las unidades de los datos. Esta medida, *z*, es la que se muestra en la ecuación (1), donde μ_f es la media de la variable *f* y σ_f la desviación típica de la misma.

$$z_{f(i)} = \frac{x_{f(i)} - \mu_f}{\sigma_f} . \qquad (1)$$

Entre las medidas de similitud se destaca la distancia euclidiana:

$$d(x_{f(i)}, x_{f(j)}) = \sqrt{\sum_{k=1}^{N} (x_{f(i)}^k - x_{f(j)}^k)^2} . \qquad (2)$$

Hay varios algoritmos estadísticos de agrupamiento. A continuación se exponen los más conocidos:

Agrupamiento numérico (K-medias).

Es uno de los algoritmos más utilizados para hacer agrupamiento. Se caracteriza por su sencillez. En primer lugar se debe especificar por adelantado cuantos grupos (*clusters*) se van a crear, éste es el parámetro *k*, para lo cual se seleccionan *k* elementos aleatoriamente, que representaran el centro o media de cada grupo. A

continuación cada una de las instancias, ejemplos, es asignada al centro del grupo más cercano de acuerdo con la distancia euclideana que le separa de él. Para cada uno de los grupos así construidos se calcula el centroide de todas sus instancias. Estos centroides son tomados como los nuevos centros de sus respectivos grupos.

Finalmente se repite el proceso completo con los nuevos centros de los grupos. La iteración continúa hasta que se repite la asignación de los mismos ejemplos a los mismos grupos, ya que los puntos centrales de los clusters se han estabilizado y permanecerán invariables después de cada iteración.

Para obtener los centroides, se calcula la media o la moda según se trate de atributos numéricos o simbólicos.

Agrupamiento conceptual (Cobweb).

El algoritmo de k-medias se encuentra con un problema cuando los atributos no son numéricos, ya que en ese caso la distancia entre ejemplares no está tan clara. Para resolver este problema Michalski presenta la noción de agrupamiento conceptual, que utiliza para justificar la necesidad de un agrupamiento cualitativo frente al agrupamiento cuantitativo, basado en la vecindad entre los elementos de la población.

En este tipo de agrupamiento, una partición de los datos es buena si cada clase tiene una buena interpretación conceptual (modelo cognitivo de jerarquías). Una de las principales motivaciones de la categorización de un conjunto de ejemplos, que básicamente supone la formación de

conceptos, es la predicción de características de las categorías que heredarán sus subcategorías.

Esta conjetura es la base de Cobweb [Molina y García, 2004]. A semejanza de los humanos, Cobweb forma los conceptos por agrupación de ejemplos con atributos similares. Representa los grupos como una distribución de probabilidad sobre el espacio de los valores de los atributos, generando un árbol de clasificación jerárquica en el que los nodos intermedios definen subconceptos. El objetivo de Cobweb es hallar un conjunto de clases o grupos (subconjuntos de ejemplos) que maximice la utilidad de la categoría (partición del conjunto de ejemplos cuyos miembros son clases).

Agrupamiento probabibilístico.

Los algoritmos de agrupamiento estudiados hasta el momento presentan ciertos defectos entre los que destacan la dependencia que tiene el resultado del orden de los ejemplos y la tendencia de estos algoritmos al sobreajuste (*overfitting*). Una aproximación estadística al problema del agrupamiento resuelve estos problemas. Desde este punto de vista, lo que se busca es el conjunto de grupos más probables dados los datos.

Ahora los ejemplos tienen ciertas probabilidades de pertenecer a un grupo. La base de este tipo de agrupamiento se encuentra en un modelo estadístico llamado mezcla de distribuciones (finite mixtures). Cada distribución representa la probabilidad de que un objeto tenga un conjunto particular de pares atributo-valor, si se supiera que es miembro de ese grupo. Se tienen k distribuciones de probabilidad que representan los k grupos. La mezcla más sencilla se tiene cuando los atributos son

numéricos con distribuciones gaussianas. Cada distribución (normal) se caracteriza por dos parámetros: la media (μ)y la varianza (σ^2). Además, cada distribución tendrá cierta probabilidad de aparición *p*, que vendrá determinada por la proporción de ejemplos que pertenecen a dicho grupo respecto del número total de ejemplos. En ese caso, si hay *k* grupos, habrá que calcular un total de 3*k*-1 parámetros: las *k* medias, *k* varianzas y *k*-1 probabilidades de la distribución dado que la suma de probabilidades debe ser 1, con lo que conocidas *k*-1 se puede determinar la *k*-ésima.

Si se conociera el grupo al que pertenece, en un principio, cada uno de los ejemplos de entrenamiento sería muy sencillo obtener los 3*k*-1 parámetros necesarios para definir totalmente las distribuciones de dichos grupos, ya que simplemente se aplicarían las ecuaciones de la media y de la varianza para cada uno de los grupos. Además, para calcular la probabilidad de cada una de las distribuciones únicamente se dividiría el número de ejemplos de entrenamiento que pertenecen al grupo en cuestión entre el número total de ejemplos de entrenamiento. Una vez obtenidos estos parámetros, si se deseara calcular la probabilidad de pertenencia de un determinado ejemplo de test a cada grupo, simplemente se aplicaría el teorema de Bayes.

13.3. Asociación o Dependencia.

Este tipo de técnicas se emplea para establecer las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes; pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros. Son utilizadas cuando el objetivo es realizar análisis exploratorios, buscando relaciones

dentro del conjunto de datos. Las asociaciones identificadas pueden usarse para predecir comportamientos, y permiten descubrir correlaciones y co-ocurrencias de eventos. Debido a sus características, estas técnicas tienen una gran aplicación práctica en muchos campos como, por ejemplo, el comercial ya que son especialmente interesantes a la hora de comprender los hábitos de compra de los clientes y constituyen un pilar básico en la concepción de las ofertas y ventas cruzada, así como del "merchandising" [Molina y García 04].

En otros entornos como el sanitario, estas herramientas se emplean para identificar factores de riesgo en la aparición o complicación de enfermedades. Para su utilización es necesario disponer de información de cada uno de los sucesos llevados a cabo por un mismo individuo o cliente en un determinado período temporal. Por lo general esta forma de extracción de conocimiento se fundamenta en técnicas estadísticas, como los análisis de correlación y de variación.

Bibliografía.

1. Aluja, T., 2001 “La Minería de Datos entre la Estadística y la Inteligencia Artificial”. Cataluña (España): Universitat Politécnica de Catalunya QUESTIIO, Vol. 25 (3) pp. 479-498.

2. Arévalo, J. L. y Pérez, R., 2002, “Estado del arte en la utilización de técnicas avanzadas para la búsqueda de información trivial a partir de de datos en sistema de abastecimiento de agua potable”. Valencia (España) Universidad Politécnica de Valencia.

3. Berry, M. y Linoff, G, 1997, "Data Mining Techniques for Marketing, Sales, and Customer Support" New York (USA):John Wiley & Song.

4. Bucheit, R.B.; Garrett, J.H; Lee, S.R, y Brahme, R., 2000, “A Knowledge Discovery framework for city civil infrastructure: a case study of the intelligent workplace”, Engineering with Computers, 16, pp. 264-274.

5. Cabena, P; Hadjinian, P.; Stadler, R. ;Verhees, J. y Zanasi, A.,1997, “Discovering Data Mining From concept to implementation”. New York (USA): Prentice Hall.

6. Carbone, P.; 1997, “Data Mining or "Knowledge Discovery in Databases" An Overview”, Mitre Corporation.

7. Fayyad, U. M., Piatetski-Shapiro, G. y Padhraic, S., 1996. “From data mining to knowledge discovery: An overview”. En: Fayyad, U.

M. et al. (editores.). Advances in knowledge discovery and data mining", Menlo Park, CA (U.S.A.). MIT Press. pp: 1-36.

8. Felgaer, P., Britos, P.; Sicre, J.; Servetto, A.; García-Martínez, R. y Perichinsky, G.,2003 "Optimización de redes bayesianas en técnicas de aprendizaje por inducción", Buenos Aires(Argentina): Universidad de Buenos Aires.

9. Ferri, C., 2004, "Técnicas del aprendizaje automático para la asistencia en la toma de decisiones" Valencia (España): Universidad Politécnica de Valencia.

10. Grossman, R.; Kasif, S.; Moore, R.; Rocke, D. Y Ullman, J.; 1998, "Data mining research: opportunities and challenges".

11. Martínez, Salvador Vázquez; Alcántara, Antonio F. Martínez,"Una arquitectura para el análisis automatizado de bases de datos" Comunicaciones en Socioeconomía, Estadística e Informática. 2003. Vol. 7 Núm. 2. pp. 89-106

12. Molina López, José Manuel. García Herrero, Jesús. "Técnicas de Análisis de Datos. Aplicaciones Prácticas Utilizando Microsoft Excel y Weka". Universidad Carlos III de Madrid. 2004.

13. Ohrn, A. 1999. "Discernibility and Rough Sets in Medicine: Tools and Applications", Trondheim, (Norway): Universidad Noruega de Ciencia y Tecnología.

14. Olaru, C.; Wehenkel, L. 1999. "Data Mining". IEEE Computer Applications in Power, 12, (3), pp. 19-25.

15. Piatesky-Shapiro, G. y Frawley W.J. 1991, "Knowledge Discovery in Databases". Cambridge (U.S.A.): MIT Press.

16. Pighin, S., 2001, "Informática Aplicada a la Ingeniería de Procesos I (Orientación I)", Rosario (Argentina): Universidad Tecnológica Nacional.

17. Simoudis, E. Livezey, B. y Kerber, R. 1996 "Integrating Inductive and Deductive Reasoning for Data Mining". En: Fayyad, U.M. et al. (editores), "Advances in knowledge Discovery and Data Mining", Menlo Park, CA (U.S.A.): MIT Press, pp: 353-373.

18. Thuraisingham, B., 1999, "Data Mining: Technologies, Techniques, Tools and Trends." CRC Press.

19. Witten, H. y Frank, E., 2000, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations". San Francisco, CA: Morgan Kaufmann.