MINERIA DE DATOS

El arte de sacar conocimiento de grandes volúmenes de datos

Elaborado Por:

- ***Centeno, Hender***
- ***Doffourt, Gineska***
- ***Garcia, Nathaly***
- ***Gómez, Giselle***
- ***González, Eduardo***
- ***Granado, Luis***
- ***Loyo, Sara***
- ***Pérez, Astrid***
- ***Pérez, Daliani***

Profesor:

Iván Turmero

Puerto Ordaz, marzo del 2011.

ÍNDICE

CONCEPTOS ASOCIADOS 6
Datos 7
Información 8
Conocimiento 9
Sabiduría 11
IMPORTANCIA DE LA DATA MINING 12
DESCUBRIMIENTO DE CONOCIMIENTO Y MINERÍA DE DATOS 13
Historia del KDD 14
Metas 15
Relación con otras disciplinas 15
El proceso de KDD 15
DATA MINING 16
Grupos de técnicas principales 19
Visualización. 19
Verificación. 20
Descubrimiento. 21
TAREAS DE LA MINERÍA DE DATOS 22
Descripción de clases 22
Encontrar Asociaciones 23
Clasificación y Predicción 24
Regresion 26
Series de tiempo 26
Agrupacion 27
Similaridad 27
ETAPAS DE LA MINERÍA DE DATOS 28
Carga de trabajo en las fases de un proyecto de datamining 28
APLICACIONES DE LA MINERÍA DE DATOS 30
Ejemplos de Usos 31
En el gobierno 31
En la empresa 31

Prediciendo el tamaño de las audiencias televisivas. 32
En la universidad 32
En investigaciones espaciales 33
En los clubes deportivos 33
Negocios 34
Patrones de fuga 35
Fraudes 35
Recursos humanos 36
Comportamiento en Internet 36
Terrorismo 36
Ciencia e Ingeniería 37
Minería de datos y otras disciplinas análogas 38
De la estadística 38
De la informática 39
HERRAMIENTAS DE SOFTWARE 40
Software DTM - Minería de Datos y Textos (Lebart) 41
CONCLUSIONES 43
FUENTES CONSULTADAS 44

INTRODUCCIÓN

Las empresas suelen generar grandes cantidades de información sobre sus procesos productivos, desempeño operacional, mercados y clientes. Pero el éxito de los negocios depende por lo general de la habilidad para ver nuevas tendencias o cambios en las tendencias. Las aplicaciones de data mining pueden identificar tendencias y comportamientos, no sólo para extraer información, sino también para descubrir las relaciones en bases de datos que pueden identificar comportamientos que no muy evidentes.

Cada día generamos una gran cantidad de información, algunas veces conscientes de que lo hacemos y otras veces inconscientes de ello porque lo desconocemos. Nos damos cuenta de que generamos información cuando registramos nuestra entrada en el trabajo, cuando entramos en un servidor para ver nuestro correo, cuando pagamos con una tarjeta de crédito o cuando reservamos un billete de avión. Otras veces no nos damos cuenta de que generamos información, como cuando conducimos por una vía donde están contabilizando el número de automóviles que pasan por minuto, cuando se sigue nuestra navegación por Internet o cuando nos sacan una fotografía del rostro al haber pasado cerca de una oficina gubernamental.

¿Con qué finalidad queremos generar información? Son muchos los motivos que nos llevan a generar información, ya que nos pueden ayudar a controlar, optimizar, administrar, examinar, investigar, planificar, predecir, someter, negociar o tomar decisiones de cualquier ámbito según el dominio en que nos desarrollemos. La información por sí misma está considerada un bien patrimonial. De esta forma, si una empresa tiene una pérdida total o parcial de información provoca bastantes perjuicios. Es evidente que la información debe ser protegida, pero también explotada.

¿Qué nos ha permitido poder generar tanta información? En los últimos años, debido al desarrollo tecnológico a niveles exponenciales tanto en el área de cómputo como en la de transmisión de datos, ha sido posible que se gestionen de una mejor manera el manejo y almacenamiento de la información. Sin duda existen cuatro factores importantes que nos han llevado a este suceso:

1. El abaratamiento de los sistemas de almacenamiento tanto temporal como permanente.
2. El incremento de las velocidades de cómputo en los procesadores.
3. Las mejoras en la confiabilidad y aumento de la velocidad en la transmisión de datos.

4. El desarrollo de sistemas administradores de bases de datos más poderosos.

Actualmente todas estas ventajas nos han llevado a abusar del almacenamiento de la información en las bases de datos. Podemos decir que algunas empresas almacenan un cierto tipo de datos al que hemos denominado *dato-escritura*, ya que sólo se guarda (o *escribe*) en el disco duro, pero nunca se hace uso de él. Generalmente, todas las empresas usan un dato llamado *dato-escritura-lectura*, que utilizan para hacer consultas dirigidas. Un nuevo tipo de dato al cual hemos denominado *dato-escritura-lectura-análisis* es el que proporciona en conjunto un verdadero conocimiento y nos apoya en las tomas de decisiones. Es necesario contar con tecnologías que nos ayuden a explotar el potencial de este tipo de datos.

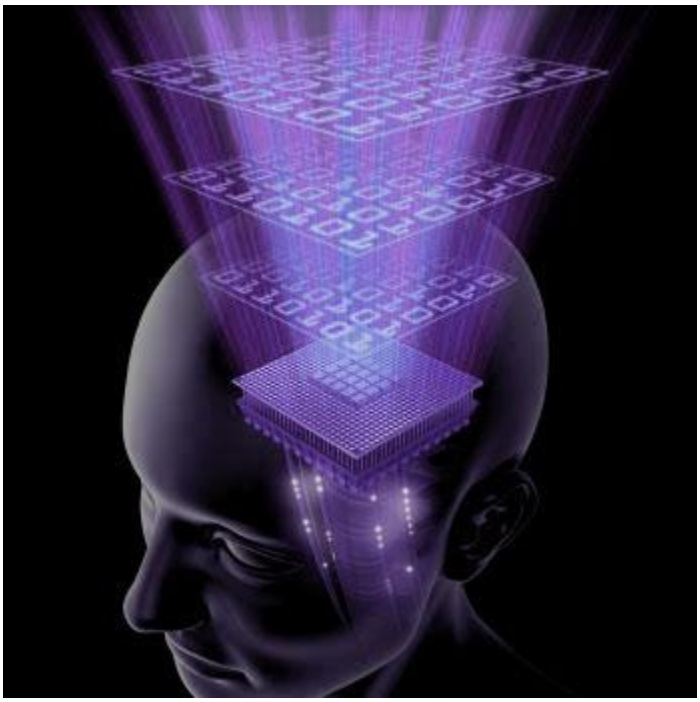
La cantidad de información que nos llega cada día es tan inmensa que nos resulta difícil asimilarla. Basta con ir al buscador *Google* y solicitar la palabra *información* para ver que existen 171.769.416 sitios donde nos pueden decir algo al respecto. Suponiendo que nos tomemos un minuto para ver el contenido de cada página, tardaríamos entonces 326 años en visitarlas todas. Esto es imposible, y, por lo tanto, existe una clara necesidad de disponer de tecnologías que nos ayuden en nuestros procesos de búsqueda y, aún más, de tecnologías que nos ayuden a comprender su contenido.

El data mining surge como una tecnología que intenta ayudar a comprender el contenido de una base de datos. De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación del confronto entre la información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento [Molina, 2002].

La técnica usada para realizar estas operaciones en data mining se denomina modelado y es, simplemente, el acto de construir un modelo en una situación donde usted conoce la respuesta y luego la aplica en otra situación de la cual desconoce la respuesta. Principal diferencia: los algoritmos están adaptados para poder trabajar sobre grandes bases de datos El data mining es un conjunto de actividades utilizadas para encontrar en los datos contextos nuevos, ocultos o inesperados. Utilizando información contenida en un data warehouse (o "depósito de datos"), el data mining puede responder a preguntas que un decisor no hubiera formulado de no contar con estas herramientas. Usando una combinación de técnicas que incluyen el análisis estadístico, la lógica neuronal, la lógica difusa, el análisis multidimensional, la visualización de datos y los agentes inteligentes, puede descubrir patrones útiles para

desarrollar modelos predictivos de conductas o de consecuencias, en una amplia variedad de dominios del conocimiento.

Nuestra capacidad para almacenar datos ha crecido en los últimos años a velocidades exponenciales. En contrapartida, nuestra capacidad para procesarlos y utilizarlos no ha ido a la par. Por este motivo, la data mining se presenta como una tecnología de apoyo para explorar, analizar, comprender y aplicar el conocimiento obtenido usando grandes volúmenes de datos. Descubrir nuevos caminos que nos ayuden en la identificación de interesantes estructuras en los datos es una de las tareas fundamentales en el data mining [Molina, 2002].



CONCEPTOS ASOCIADOS

- Datos: son hechos, medidas u observaciones, que pueden presentarse (o no) en un contexto dado. Datos sin contexto son 60, 62, 66, 72. Los mismos datos, ahora con contexto, podrían representar el peso en kilogramos de Laura, Ana, Juan y Pedro, respectivamente. La validez y la efectividad de los datos vienen determinadas principalmente por su exactitud.

- Información: Son los datos organizados de cierta manera, de forma tal que sean de utilidad y relevancia para quien tiene que resolver un problema de decisión. El criterio clave para evaluar la información es su utilidad.

- Conocimiento: Es una combinación de instintos, ideas, reglas, procesos e información que un decisor aplica para guiar sus acciones y decisiones. El conocimiento es una interpretación realizada por la mente, que será válida cuando pueda explicar las interacciones de un problema con su contexto.

De lo anterior, se infieren dos puntos. El primero es que la información es personal; el segundo, que el conocimiento no es estático: es más, debe cambiar cuando cambia el entorno de decisión. En la figura 1 se ilustra la jerarquía que existe en una base de datos entre dato, información y conocimiento.

Figura 1. Pirámide Informacional

El data mining trabaja en el nivel superior buscando patrones, comportamientos, agrupaciones, secuencias, tendencias o asociaciones que puedan generar algún modelo que nos permita comprender mejor el dominio para ayudar en una posible toma de decisión. La pirámide informacional explica el proceso de transformación asociado a la generación del conocimiento. En esta se indica que el nivel más bajo de los hechos conocidos son los datos. Los datos no tienen un significado por sí mismos, ya que deben ser ordenados, agrupados, analizados e interpretados para entender potencialmente lo que por sí sólo nos quieren indicar. [Camargo, 2006] Cuando los datos son procesados de esta manera, se convierten en información. La información tiene una esencia y un propósito. Cuando la información es utilizada y puesta en el contexto o marco de referencia de una persona junto con su percepción personal se transforma en conocimiento. El conocimiento es la combinación de información, contexto y experiencia. El conocimiento resumido, una vez validado y orientado hacia un objetivo genera inteligencia (sabiduría), la cual pretende ser una representación de la realidad. Estos factores están gobernados por dos criterios: Cantidad y Calidad.

Re caracterizando entonces, los términos previamente conceptualizados:

Datos

Son un conjunto discreto de hechos objetivos acerca de eventos. Cuando un cliente va a un cajero automático y hace un retiro, la operación puede ser descrita parcialmente por los datos: cuándo retiró, cuánto retiró, de que cuenta retiró, etc.; pero no nos informa de las razones por las que retiró, ni puede predecir cuándo volverá. Las organizaciones precisan almacenar y manipular datos y son fuertemente dependientes de ellos. El registro de los datos y la gestión efectiva de los mismos es fundamental para el éxito. Pero no siempre mayor cantidad de datos es mejor, primero porque demasiados datos hace más difícil extraer el sentido de los mismos y segundo, porque describen parcialmente lo que sucede y no dan juicios ni interpretaciones, ni permiten la toma de decisiones. Los datos por supuesto son importantes porque son la materia prima de la información. [Camargo, 2006] Por sí solos son irrelevantes como apoyo a la toma de decisiones. También se pueden ver como un conjunto discreto de valores, que no dicen nada sobre el por qué de las cosas y no son orientativos para la acción. Los datos pueden ser una colección de hechos almacenados en algún lugar físico como un papel, un dispositivo electrónico (CD, DVD, disco duro...), o la mente de una persona. En este sentido las tecnologías de la información han aportado mucho a recopilación de datos. Como cabe suponer, los datos pueden provenir de fuentes externas o internas, pudiendo ser de carácter objetivo o subjetivo, o de tipo cualitativo o cuantitativo, etc.

Información

La información puede definirse como un mensaje significativo que se transmite de la fuente a los usuarios, es la expresión material del conocimiento con fines de uso. Un conocimiento que no se utiliza Minería de Datos no se convierte en información, una información que no se asimile nunca se convierte en conocimiento. Es la forma social de existencia del conocimiento consolidada en una fuente determinada. La información son "datos dotados de relevancia y propósito". Para muchos investigadores la información es mensaje, normalmente en forma de documento o de combinación visible o audible. Cualquier mensaje requiere un emisor y un receptor. La información persigue cambiar la forma en que el receptor percibe algo, influir sobre su juicio y su comportamiento. Así pues, el mensaje debe informar (originalmente dar forma a), es decir modelar a quien recibe el mismo. Es el receptor el que da categoría de información al mensaje. [Camargo, 2006] La información se puede definir como un conjunto de datos procesados y que tienen un significado (relevancia, propósito y contexto), y que por lo tanto son de utilidad para quién debe tomar decisiones, al disminuir su incertidumbre. Está destinada a resolver determinados problemas. Debe estar disponible públicamente y servir para al desarrollo individual y corporativo. Se encuentra presente en todos los niveles de actividad y en todas las ramas de la economía, la política y la sociedad. La información de hoy puede que mañana no lo sea, y lo que para unos es información para otros no tiene por qué serlo. La información para que sea utilizable debe tener tres características básicas: completa, confiable y oportuna. Una información completa debe contar con los elementos necesarios para que pueda ser analizada y procesada; confiable, debe provenir de una fuente veraz y creíble; oportuna, debe llegar a tiempo para su empleo. [Camargo, 2006]

Lo que realmente se debería preguntar acerca de la información que se recibe es: ¿Nos aporta nuevas perspectivas?, ¿Ayuda a que la situación tenga más sentido?, ¿Contribuye a la toma de decisiones o a la solución del problema?. A diferencia de los datos, la información tiene sentido y en si misma tiene forma, está organizada con algún propósito. Los datos se convierten en información cuando se les añade sentido a través de (5 C):

- Contextualizados: Se sabe en qué contexto y para qué propósito se generaron.
- Categorizados: Se conocen sus componentes claves, las unidades de medida que ayudan a interpretarlos
- Calculados: Han sido analizados matemática o estadísticamente.
- Corregidos: Se han eliminado errores e inconsistencias de los datos.
- Condensados: Han sido resumidos, son más concisos (agregación).

Por tanto, la información es la comunicación de conocimientos o inteligencia, y es capaz de cambiar la forma en que el receptor percibe algo, impactando sobre sus juicios de valor y sus comportamientos.

Información = Datos + Contexto (añadir valor) + Utilidad (disminuir la incertidumbre)

La información también puede evaluarse según diferentes parámetros:

- Accesibilidad (información de accesibilidad x)
- Comprensividad (información de comprensividad igual a x)
- Precisión (información precisa, o de precisión igual a x)
- Relevancia (información relevante "vs" información superflua)
- Puntualidad (información de alta/baja puntualidad)
- Claridad (información de alta/baja ambigüedad)
- Flexibilidad (información altamente compartible o de x nivel de compartibilidad)
- Verificabilidad (información de alta o baja verificabilidad)
- Cuantificabilidad (información cuantificable "vs" información no cuantificable).

Conocimiento

El conocimiento es "una verdad justificada". El conocimiento surge cuando una persona considera, interpreta y utiliza la información de manera combinada con su propia experiencia y capacidad. Por esta razón podemos decir que el conocimiento está condicionado por la interpretación que las personas efectúan de la información disponible, condicionada por el contexto en que se desenvuelven y la experiencia que tienen. [Camargo, 2006]

"El conocimiento es una mezcla fluida de experiencias, valores, información contextual y apreciaciones expertas (know-how) que proporcionan un marco para su evaluación e incorporación de nuevas experiencias e información, y es útil para la acción. Se origina y aplica en las mentes de los conocedores. En las organizaciones está, a menudo, embebido no sólo en los documentos y bases de datos, sino también en las rutinas organizacionales, en los procesos, prácticas y normas".

El conocimiento debería tener cuatro particulares:

1. Es tácito: porque los conceptos cambian o se adaptan a la luz de las experiencias de los individuos.

2. Es orientado a la acción: porque posee la cualidad dinámica de generar nuevos conocimientos y superar los antiguos.

3. Está sustentado por reglas: porque la creación de patrones en el cerebro, con el paso del tiempo, permiten actuar con rapidez y eficacia, de forma automática, en situaciones inconcebibles.

4. Está en constante cambio: porque el conocimiento puede ser distribuido, criticado y aumentado.

El conocimiento deriva de la información como esta deriva de los datos, aunque son los humanos los que hacen todo el trabajo para que esta transformación tenga lugar, mediante las siguientes acciones (4 C):

Comparación con otros elementos: ¿Cómo se ajusta la información en la situación dada, comparada con otras situaciones ya conocidas? Predicción de Consecuencias: ¿Qué implicaciones tiene la información para la toma de decisiones y la acción? Búsqueda de Conexiones: ¿Cómo se relaciona este fragmento de conocimiento con otros fragmentos? Conversación con otros portadores de conocimiento: ¿Qué piensan otras personas acerca de esta información?

Características del Conocimiento

El conocimiento es personal, en el sentido de que se origina y reside en las personas, que lo asimilan como resultado de su propia experiencia (es decir, de su propio "hacer", ya sea físico o intelectual) y lo incorporan a su acervo personal estando "convencidas" de su significado e implicaciones, articulándolo como un todo organizado que da estructura y significado a sus distintas "piezas". Su utilización, que puede repetirse sin que el conocimiento "se consuma" como ocurre con otros bienes físicos, permite "entender" los fenómenos que las personas perciben (cada una "a su manera", de acuerdo precisamente con lo que su conocimiento implica en un momento determinado), y también "evaluarlos", en el sentido de juzgar la bondad o conveniencia de los mismos para cada una en cada momento. Sirve de guía para la acción de las personas, en el sentido de decidir qué hacer en cada momento porque esa acción tiene en general por objetivo mejorar las consecuencias, para cada individuo, de los fenómenos percibidos (incluso cambiándolos si es posible). Estas características convierten al conocimiento, en un cimiento sólido para el desarrollo de ventajas competitivas. En efecto, en la medida en que es el resultado de la acumulación de experiencias de personas, su imitación es complicada a menos que existan representaciones precisas que permitan su transmisión a otras personas efectiva y eficientemente.

Sabiduría

El saber es definido como la capacidad de comprender los principios, como contraposición al conocimiento, que comprende patrones, y la información, que comprende relaciones y cuya acumulación puede dar lugar, en términos más prácticos, al capital intelectual. El saber, como estadio superior al conocimiento, tiene que ver con los principios, la introspección, la moral, los arquetipos, tratando de dar respuesta al por qué de las cosas, en tanto que el conocimiento se asocia al cómo, incluyendo estrategias, prácticas, métodos y enfoques y, más abajo, la información que se asocia a las descripciones, definiciones y perspectivas: qué, quién, cuándo, dónde. A los datos, exentos de significado por sí mismos, ni siquiera se le asignan atributos diferenciados. [Camargo, 2006]

Por último, vale la pena señalar la distinción que Gene Meieran (Intel) hace entre saber y conocimiento:

"El saber nos permite tomar decisiones sobre el futuro, mientras que los conocimientos se refieren a las decisiones sobre el presente". Además agrega: "El saber se alcanza, buscando las relaciones entre las cosas e intentando comprenderlas y basando todo juicio sobre la pericia y la experiencia". Niel Fleming presenta un diagrama que asocia el nivel de independencia del contexto y el nivel de entendimiento en torno a los elementos de la cadena informacional: los datos, la información, el conocimiento, la sabiduría y la verdad.

Figura 2. Cadena Informacional

Dónde:

Dato: Es un punto en el espacio y en el tiempo el cual no cuenta con referencias espaciales y temporales.

Información: Una colección de datos no es información. Las piezas de datos representan información de acuerdo a la medida de asociación existente entre ellos, lo cual permite generar discernimiento en torno a ellas. Representa el cuál, el quién, el cuándo y el dónde.

Conocimiento: Una colección de información no es conocimiento. Mientras que la información entrega las asociaciones necesarias para entender los datos, el conocimiento provee el fundamento de cómo cambian (en el caso que lo hagan). Esto claramente puede ser visto como patrones de comportamiento contextuados, es decir una relación de relaciones. Representa el cómo.

Sabiduría: La sabiduría abarca los principios fundacionales responsables de los patrones que representan el conocimiento. Representa el porqué.

Verdad: La totalidad de los factores de sabiduría y sus relaciones. Representa el ser.

IMPORTANCIA DE LA DATA MINING

En el ámbito comercial, resulta interesante encontrar patrones ocultos de consumo de los clientes para poder explorar nuevos horizontes. Saber por ejemplo, que un vehículo deportivo corre un riesgo de accidente casi igual al de un vehículo normal cuando su dueño tiene un segundo vehículo en casa ayuda a crear nuevas estrategias comerciales para ese grupo de clientes. Asimismo, predecir el comportamiento de un futuro cliente, basándose en los datos históricos de clientes que presentaron el mismo perfil, ayuda a poder retenerlo durante el mayor tiempo posible.

Vivimos permanentemente bombardeados por informes sobre la importancia del conocimiento y la información para obtener ventajas competitivas en un mundo globalizado. Las actividades de vigilancia e inteligencia, en el contexto de la "sociedad de la Información", se considera como "la transformación de la información en conocimiento, y del conocimiento en acción". Es por esto la importancia del "análisis de la información" en el seno de todas las actividades en donde se trata de transformar los datos brutos con el fin de extraer los conocimientos que puedan ser explotados y útiles en un determinado campo de acción.

DESCUBRIMIENTO DE CONOCIMIENTO Y MINERÍA DE DATOS

El término descubrimiento de conocimiento en bases de datos (knowledge discovery in databases, KDD para abreviar) se refiere al amplio proceso de búsqueda de conocimiento en bases de datos, y para enfatizar la aplicación a "alto nivel" de métodos específicos de minería de datos. En general, el descubrimiento es un tipo de inducción de conocimiento, no supervisado, que implica dos procesos:

- búsqueda de regularidades interesantes entre los datos de partida,
- formulación de leyes que las describan.

Proceso KDD
1. Pre-procesamiento de Datos: Limpieza, integración y transformación.
2. Minería de Datos: Uso de métodos inteligentes para extraer conocimiento (búsqueda de oro) .
3. Evaluación de patrones encontrados y presentación
Los principales pasos dentro del proceso interactivo e iterativo del KDD pueden verse en la figura 3

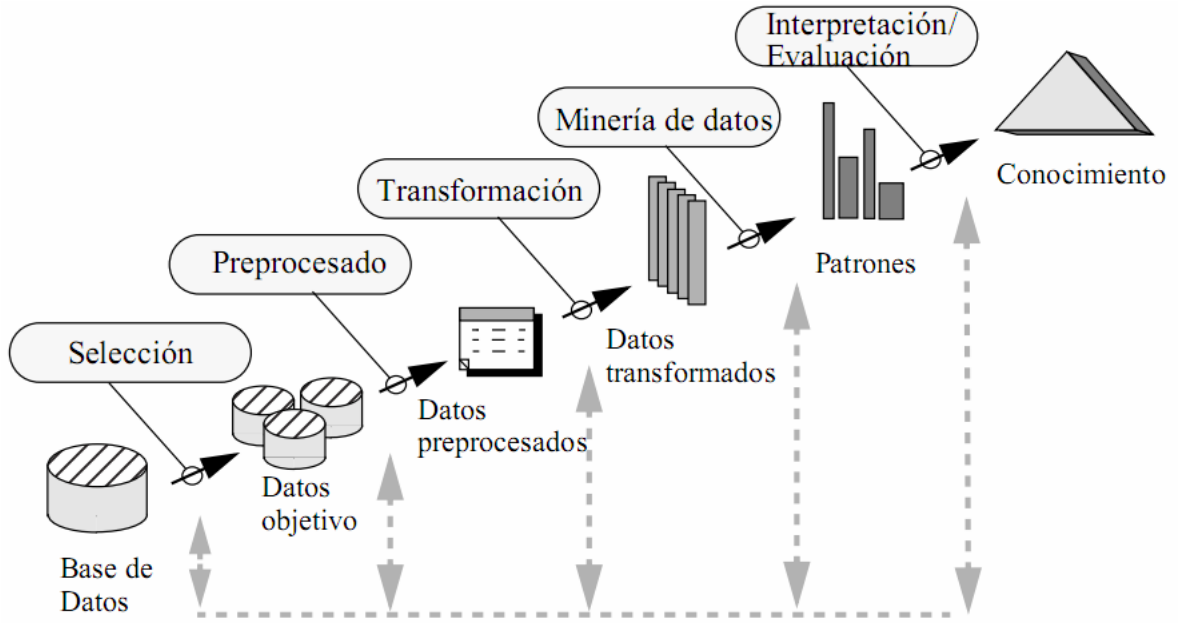


Figura 3.

Historia del KDD

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 (DataMining), 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.

En la figura 4 se ilustra la jerarquía que existe en una base de datos entre datos, información y conocimiento. Se observa igualmente el volumen que presenta en cada nivel y el valor que los responsables de las decisiones le dan en esa jerarquía. El área interna dentro del triángulo representa los objetivos que se han propuesto. La separación del triángulo representa la estrecha unión entre dato e información, no así elntre la información y el conocimiento.

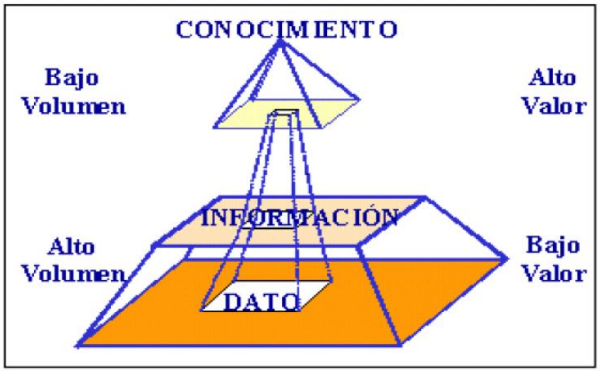


Figura 4: Jerarquía del Conocimiento. **Relación entre dato, información y conocimiento (Molina, 1998).**

Metas

Las metas del KDD son:

• Procesar automáticamente grandes cantidades de datos crudos.
• Identificar los patrones más significativos y relevantes.
• Presentarlos como conocimiento apropiado para satisfacer las metas del usuario.

Relación con otras disciplinas

• KDD nace como interfaz y se nutre de diferentes disciplinas:
• Sistemas de información / bases de datos: tecnologías de bases de datos y bodegas de datos, maneras eficientes de almacenar, accesar y manipular datos.
• Estadística, aprendizaje automático / IA (redes neuronales, lógica difusa, algoritmos genéticos, razonamiento probabilístico): desarrollo de técnicas para extraer conocimiento a partir de datos.
• Reconocimiento de patrones: desarrollo de herramientas de clasificación.
• Visualización de datos: interfaz entre humanos y datos, y entre humanos y patrones.
• Computación paralela / distribuida: cómputo de alto desempeño, mejora de desempeño de algoritmos debido asu complejidad y a la cantidad de datos.
• Interfaces de lenguaje natural a bases de datos.

El proceso de KDD

El proceso de KDD consiste en usar métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con preprocesamientos y post-procesamientos. En la figura 3. se ilustra el proceso de KDD.

Se estima que la extracción de patrones (minería) de los datos ocupa solo el 15% - 20% del esfuerzo total del proceso de KDD. El proceso de descubrimiento de conocimiento en bases de datos involucra varios pasos:

• Determinar las fuentes de información: que pueden ser útiles y dónde conseguirlas.
• Diseñar el esquema de un almacén de datos (Data Warehouse): que consiga unificar de manera operativa toda la información recogida.
• Implantación del almacén de datos: que permita la navegación y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados. Esta es la etapa que puede llegar a consumir el mayor tiempo.
• Selección, limpieza y transformación de los datos que se van a analizar: la selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos).La limpieza y prepocesamiento de datos se logra diseñando una estrategia adecuada para manejar

ruido, valores incompletos, secuencias de tiempo, casos extremos (si es necesario), etc.

• Seleccionar y aplicar el método de minería de datos apropiado: esto incluye la selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc. La selección de él o de los algoritmos a utilizar. La transformación de los datos al formato requerido por el algoritmo específico de minería de datos. Y llevar a cabo el proceso de minería de datos, se buscan patrones que puedan expresarse como un modelo o simplemente que expresen dependencias de los datos, el modelo encontrado depende de su función (clasificación) y de su forma de representarlo (árboles de decisión, reglas, etc.), se tiene que especificar un criterio de preferencia para seleccionar un modelo dentro de un conjunto posible de modelos, se tiene que especificar la estrategia de búsqueda a utilizar (normalmente está predeterminada en el algoritmo de minería)

• Evaluación, interpretación, transformación y representación de los patrones extraídos:

Interpretar los resultados y posiblemente regresar a los pasos anteriores. Esto puede involucrar repetir el proceso, quizás con otros datos, otros algoritmos, otras metas y otras estrategias. Este es un paso crucial en donde se requiere tener conocimiento del dominio. La interpretación puede beneficiarse de procesos de visualización, y sirve también para borrar patrones redundantes irrelevantes.

• Difusión y uso del nuevo conocimiento. Incorporar el conocimiento descubierto al sistema (normalmente para mejorarlo) lo cual puede incluir resolver conflictos potenciales con el conocimiento existente. El conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente para almacenarlo y reportarlo a las personas interesadas. En este sentido, KDD implica un proceso interactivo e iterativo involucrando la aplicación de varios algoritmos de minería de datos.

DATA MINING

Aunque desde un punto de vista académico el término data mining es una etapa dentro de un proceso mayor llamado extracción de conocimiento en bases de datos (Knowledge Discovery in Databases o KDD) en el entorno comercial, ambos términos se usan de manera indistinta. Lo que en verdad hace el data mining es reunir las ventajas de varias áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo, principalmente usando como materia prima las bases de datos. Una definición tradicional es la siguiente: "la integración de un conjunto de áreas que tienen como propósito la identificación de un

conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión" [Molina, 2002].

La idea de data mining no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como data fishing, data mining o data archaeology con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de data mining y KDD. A finales de los años ochenta sólo existían un par de empresas dedicadas a esta tecnología; en 2002 existen más de 100 empresas en el mundo que ofrecen alrededor de 300 soluciones. Las listas de discusión sobre este tema las forman investigadores de más de ochenta países. Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios.

El data mining es una tecnología compuesta por etapas que integra varias áreas y que no se debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones software en cada etapa que pueden ser estadísticas, de visualización de datos o de inteligencia artificial, principalmente. Actualmente existen aplicaciones o herramientas comerciales de data mining muy poderosas que contienen un sinfín de utilerías que facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta. La data mining es la etapa de descubrimiento en el proceso de KDD: Paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos preprocesados (Fayyad et al., 1996) Aunque se suelen usar indistintamente los términos KDD y Minería de Datos.

Los proyectos de Data Mining tienen por objetivo extraer información útil a partir de grandes cantidades de datos y se aplican a todos los sectores y en todos los campos. Así existen proyectos de este tipo en sectores tan dispares como el comercio electrónico, la banca, las empresas industriales o la exploración petrolífera. La extracción de esta información útil es un proceso complejo, que requiere la aplicación de una metodología estructurada para la utilización ordenada y eficiente de las técnicas y herramientas disponibles [Rodríguez, 2005].

La minería de datos (DM, Data Mining) consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras,

la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos.

Bajo el nombre de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos. Está fuertemente ligado con la supervisión de procesos industriales ya que resulta muy útil para aprovechar los datos almacenados en las bases de datos.

Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico. Mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación.

Las **categorías básicas** de las técnicas de minería de datos actualmente en uso se pueden clasificar en: clasificación, asociación, secuencia y cluster.

- ✓ Clasificación: incluye los procesos de minería de datos que buscan reglas para definir si un ítem o un evento pertenecen a un *subset* particular o a una clase de datos. Esta técnica, probablemente la más utilizada, incluye dos subprocesos: la construcción de un modelo y la predicción.

  En términos generales, los métodos de clasificación desarrollan un modelo compuesto por reglas IF-THEN y se aplican perfectamente, por ejemplo, para encontrar patrones de compra en las bases de datos de los clientes y construir mapas que vinculan los atributos de los clientes con los productos comprados. Con un conjunto apropiado de atributos predictivos, el modelo puede identificar los clientes con mayor propensión a realizar una determinada compra durante el próximo mes. Un caso típico de clasificación es el de dividir una base de datos de compañías en grupos homogéneos respecto a variables como "posibilidades de crédito" con valores tales como"bueno" y "malo".

- ✓ Asociación: incluye técnicas conocidas como linkage analysis, utilizadas para buscar patrones que tienen una probabilidad alta de repetición, como ocurre al analizar una canasta en la búsqueda de productos afines.

  Se desarrolla un algoritmo asociativo que incluye las reglas que van a correlacionar un conjunto de eventos con otro. Por ejemplo, un supermercado podría necesitar información sobre los hábitos de compra de sus clientes, para

reubicar los productos que se suelen comprar juntos, para localizar los productos nuevos en el mejor lugar, o para ofrecer promociones.

- ✓ Secuencia: los métodos de analisis de series de tiempo son usados para relacionar los eventos con el tiempo.

  A título ilustrativo: como resultado de este tipo de modelo se puede aprender que las personas que alquilan una película de video tienden a adquirir los productos promocionales durante las siguientes dos semanas; o bien, que la adquisición de un horno de microondas se produce frecuentemente luego de determinadas compras previas.

- ✓ Cluster: Muchas veces resulta difícil o imposible definir los parámetros de una clase de datos. En ese caso, los métodos de clustering pueden usarse para crear particiones, de forma tal que los miembros de cada una de ellas resulten similares entre sí, según alguna métrica o conjunto de métricas.

  El análisis de clusters podría utilizarse, entre otras aplicaciones, al estudiar las compras con tarjetas de crédito, para descubrir—digamos—que los alimentos comprados con una tarjeta dorada de uso empresarial son adquiridos durante los días de semana y tienen un valor promedio de ticket de 152 pesos, mientras que el mismo tipo de compra, pero realizado con una tarjeta platino personal, ocurre predominantemente durante los fines de semana, por un valor menor, pero incluye una botella de vino más del 65 % de las veces.

Como puede verse, en el mundo de la data mining las preguntas que se pueden hacer y responder son infinitas y las metodologías utilizadas para responderlas son diversas y cada vez más variadas. Así como existen muchas técnicas para dar soporte al proceso de decidir con la minería de datos, existen variadas tecnologías para construir los modelos

Grupos de técnicas principales

Visualización.

• Ayudas para el descubrimiento manual de información.
• Se muestran tendencias, agrupamientos de datos, etc.
• Funcionamiento semi-automático.

Técnicas de visualización:

- Visualización en 2D de datos multidimensionales.
- Problema con múltiples atributos:
  - Vibraciones.
  - Temperatura.
  - Horas funcionamiento.
  - Meses desde revisión…
- Se calculan las distancias entre cada 2 instancias de entrenamiento.

$$d(x_1, x_2) = \sqrt{(temp_1 - temp_2)^2 + (vib_1 - vib_2)^2 + (horas_1 - horas_2)^2 + (rev_1 - rev_2)^2}$$

- Se crea un gráfico 2D en el que cada instancia corresponde a un punto y en el que las distancias entre instancias son lo más parecidas posible a las distancias reales.

Verificación.

• Se conoce de antemano un modelo y se desea saber si los datos disponibles se ajustan a él.

• Se establecen medidas de ajuste al modelo.

Técnicas de verificación:

- Se conoce de antemano un modelo y se desea verificar si es un buen modelo para el conjunto de instancias disponible.
- Medidas utilizadas: soporte y precisión.
  - Soporte: dada una regla, porcentaje de instancias que cumplen sus condiciones.
  - Precisión: dada una regla, porcentaje de casos en los que la regla se cumple.

soporte

"si (temp=alta) y (vibr=altas) entonces (fallará)"

precisión

- Interesan soporte y precisión altos.

Descubrimiento.

• Se busca un modelo desconocido de antemano.
• Descubrimiento descriptivo: se busca modelo legible.
• Descubrimiento predictivo: no importa que el modelo no sea legible.

Técnicas de descubrimiento.

- Descriptivo: se busca un modelo legible.
  - Clustering (agrupamiento).
  - Resumen.
  - Generación de reglas en cualquier formato.
    - Justifican la clase (ej. árboles de decisión).
    - Justifican cualquier relación (ej. reglas de asociación).
- Predictivo: no importa si el modelo no es legible.
  - Clasificación.
  - Análisis de series temporales (predicción).
  - Regresión.

Sea como sea la presentación del problema, una de las características presente en cualquier tipo de aprendizaje y en cualquier tipo de técnica de Minería de Datos es su carácter hipotético, es decir, lo aprendido puede, en cualquier momento, ser refutado por evidencia futura.

En muchos casos, los modelos no aspiran a ser modelos perfectos, sino modelos aproximados. En cualquier caso, al estar trabajando con hipótesis, es necesario realizar una evaluación de los patrones obtenidos, con el objetivo de estimar su validez y poder compararlos con otros. Por tanto, la Minería de Datos, más que verificar patrones hipotéticos, usa los datos para encontrar estos patrones. Por lo tanto, es un proceso inductivo.

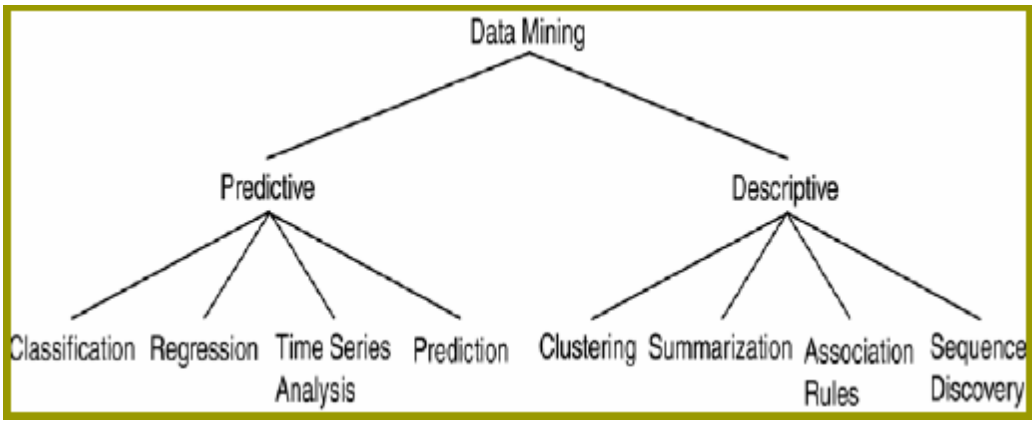
Pueden emplearse diferentes criterios para clasificar los sistemas de minería de datos y, en general, los sistemas de aprendizaje inductivo en ordenadores:

DEPENDIENDO DEL OBJETIVO para el que se realiza el aprendizaje, pueden distinguirse sistemas para: clasificación (clasificar datos en clases predefinidas), regresión (función que convierte datos en valores de una función de predicción), agrupamiento de conceptos (búsqueda de conjuntos en los que agrupar los datos), compactación (búsqueda de descripciones más compactas de los datos), modelado de dependencias (dependencias entre las variables de los datos), detección de desviaciones (búsqueda de desviaciones importantes de los datos respecto de valores anteriores o medios), etc.

DEPENDIENDO DE LA TENDENCIA con que se aborde el problema, se pueden distinguir tres gran- des líneas de investigación o paradigmas: sistemas conexionistas (redes neuronales), sistemas evolucionistas (algoritmos genéticos) y sistemas simbólicos.

DEPENDIENDO DEL LENGUAJE utilizado para representar del conocimiento, se pueden distinguir: representaciones basadas en la lógica de proposiciones, representaciones basadas en lógica de predicados de primer orden, representaciones estructuradas, representaciones a través de ejemplos y representaciones no simbólicas como las redes neuronales.

TAREAS DE LA MINERÍA DE DATOS



Descripción de clases

Sumarizacion/ caracterización de la colección de datos
- Tendencias
- Reportes

Aplicaciones:
- Supermercados (Canasta de mercado)
- Contratos de Mantenimiento (Que debe hacer el almacén para potenciar las ventas de contratos de mantenimiento) ‡98% de la gente que compra llantas y accesorios de autos también obtiene servicios de mantenimiento
- Recomendaciones de páginas Web (URL1 & URL3 -> URL5) ‡60% de usuarios de la Web quien visita la Página A y B compra el ítem T1

Encontrar Asociaciones

Descubrir asociaciones, relaciones / correlaciones entre un conjunto de "ítems"

Id	Ítems
1	{pan, leche}
2	{pan,pañales,cerveza,huevos}
3	{leche, pañales, cerveza, gaseosa}
4	{pan,leche,pañales,cerveza}
5	{pan,leche,pañales,gaseosa}

Interés en analizar los datos para aprender el comportamiento de las compras de sus clientes

- Promociones de mercadeo
- Manejo de inventario
- Relación con el cliente

Asociaciones, relaciones / correlaciones en forma de reglas: X ⇒ Y
(registros en BD que satisfacen X, también satisfacen Y)

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Min. support 50%
Min. confidence 50%

Frequent Itemset	Support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

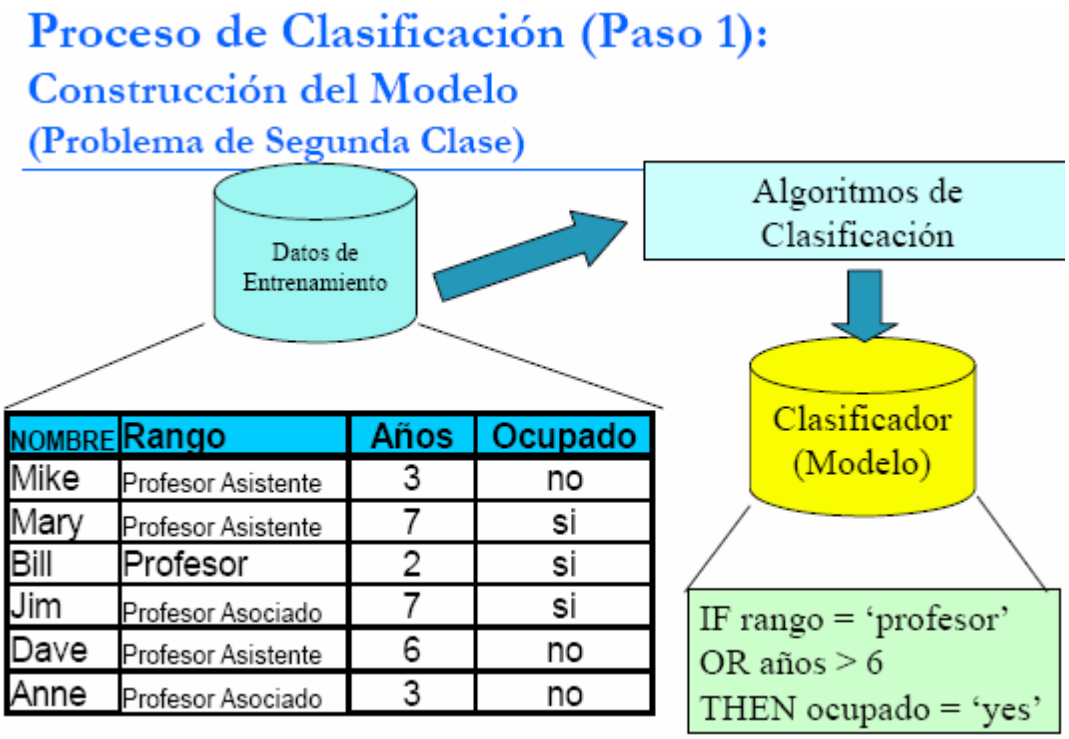
For rule $A \Rightarrow C$: support = support($\{A \cup C\}$) = 50%
confidence = support($\{A \cup C\}$)/support($\{A\}$) = 66.6%

Aplicaciones:
- Aprobación de créditos

- Diagnóstico médico
- Clasificación de documentos de texto (text mining)
- Recomendación de páginas Web automáticamente
- Seguridad

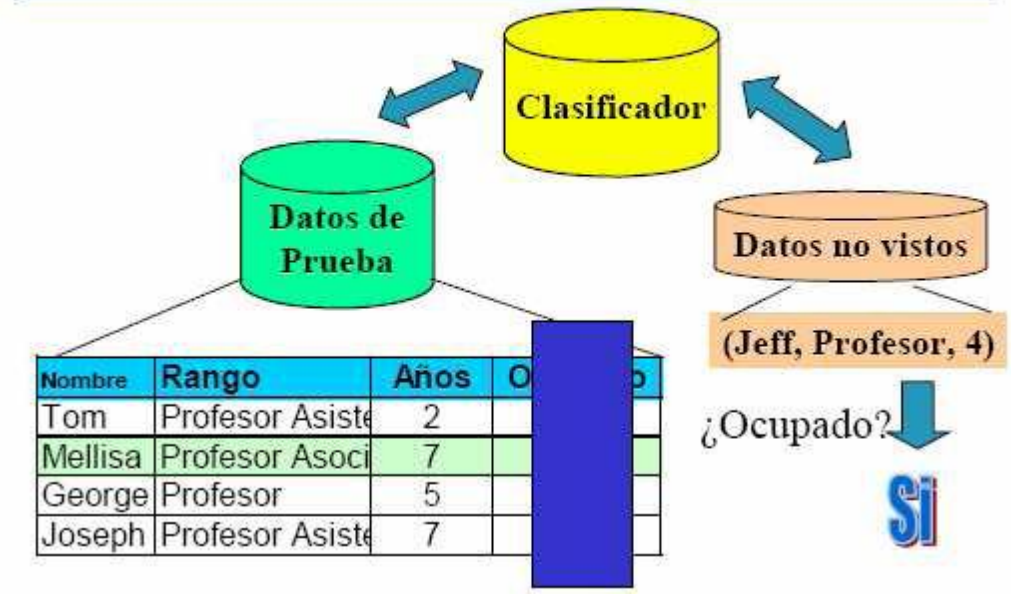
Clasificación y Predicción

Clasificación: Construir un modelo por cada clase de dato etiquetado usado en el entrenamiento del modelo.
Basado en sus características y usado para clasificar futuros datos
Predicción: Predecir valores posibles de datos/atributos basados en similar objetos.



NOMBRE	Rango	Años	Ocupado
Mike	Profesor Asistente	3	no
Mary	Profesor Asistente	7	si
Bill	Profesor	2	si
Jim	Profesor Asociado	7	si
Dave	Profesor Asistente	6	no
Anne	Profesor Asociado	3	no

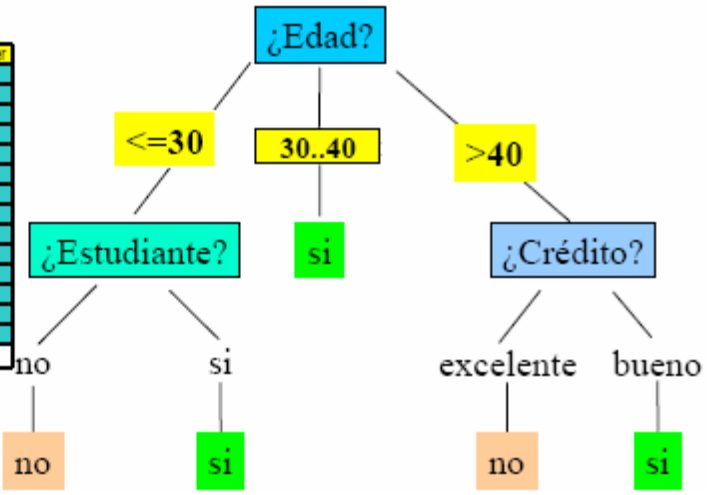
Proceso de Clasificación (Paso 2): Uso del Modelo en la Predicción



Clasificación y Predicción (árboles de decisión)

- **Determinar si clientes compraran determinado producto.**

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
30…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



Evaluación del modelo

- Basado en el número de registros que fueron clasificados correcta e incorrectamente.
- Confusion Matrix

		Clase predicción	
		Clase =1	Clase =2
Clase	Clase =1	f_{11}	f_{10}
	Clase =2	f_{01}	f_{00}

$$Accuracy = \frac{Número \quad correctos}{Total} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$Error \quad rate = \frac{Número \quad incorrectos}{Total} = \frac{f_{10} + f_{10}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Aplicaciones:
- ‡ Procesamiento de Imágenes (segmentar imágenes a color en regiones)
- ‡ Indexamiento de texto e imágenes
- ‡ WWW

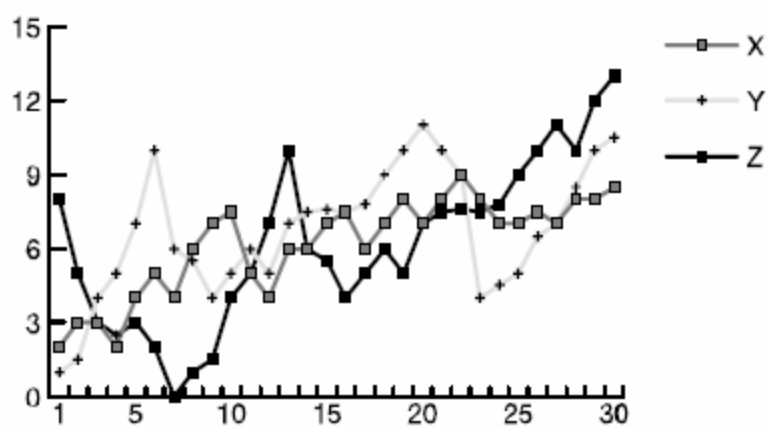
Clasificación de paginas Web (usados por motores de búsqueda -Google)
Agrupar web log para descubrir grupos de patrones de acceso similares (web usage profiles)
- ‡ Seguridad: Descubriendo patrones de acceso a redes (Detección de intrusos)

Regresion

Predicción de una variable real (no categórica)
- Variable real -> regresión
- Variable categórica -> clasificación

Series de tiempo

Predecir valores futuros de acuerdo al tiempo



Agrupacion

Dividir datos sin etiqueta en grupos (clusters) de tal forma que datos que pertenecen al mismo grupo son similares, y datos que pertenecen a diferentes grupos son diferentes

Similaridad

Medidas de similaridad/distancia. Dependen de los tipos de datos
‡Numéricos:
Distancia Euclidean
‡Binarios
Jaccard
Cosine

ETAPAS DE LA MINERÍA DE DATOS

Aunque en datamining cada caso concreto puede ser radicalmente distinto al anterior, el proceso común a todos ellos se suele componer de cuatro etapas principales:

- ✓ Determinación de los objetivos. Trata de la delimitación de los objetivos que el cliente desea bajo la orientación del especialista en data mining.
- ✓ Preprocesamiento de los datos. Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de data mining.
- ✓ Determinación del modelo. Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.
- ✓ Análisis de los resultados. Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

Carga de trabajo en las fases de un proyecto de datamining

Un proceso típico de minería de datos consta de los siguientes pasos generales:

1. Selección del conjunto de datos, tanto en lo que se refiere a las variables objetivo (aquellas que se quiere predecir, calcular o inferir), como a las variables independientes (las que sirven para hacer el cálculo o proceso), como posiblemente al muestreo de los registros disponibles.

2. Análisis de las propiedades de los datos, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).

3. Transformación del conjunto de datos de entrada, se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema, a este paso también se le conoce como preprocesamiento de los datos.

4. Seleccionar y aplicar la técnica de minería de datos, se construye el modelo predictivo, de clasificación o segmentación.

5. Extracción de conocimiento, mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.

6. Interpretación y evaluación de datos, una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Si el modelo final no superara esta evaluación el proceso se podría repetir desde el principio o, si el experto lo considera oportuno, a partir de cualquiera de los pasos anteriores. Esta retroalimentación se podrá repetir cuantas veces se considere necesario hasta obtener un modelo válido.

Una vez validado el modelo, si resulta ser aceptable (proporciona salidas adecuadas y/o con márgenes de error admisibles) éste ya está listo para su explotación. Los modelos obtenidos por técnicas de minería de datos se aplican incorporándolos en los sistemas de análisis de información de las organizaciones, e incluso, en los sistemas transaccionales. En este sentido cabe destacar los esfuerzos del Data Mining Group, que está estandarizando el lenguaje PMML (Predictive Model Markup Language), de manera que los modelos de minería de datos sean interoperables en distintas plataformas, con independencia del sistema con el que han sido construidos. Los principales fabricantes de sistemas de bases de datos y programas de análisis de la información hacen uso de este estándar.

Tradicionalmente, las técnicas de minería de datos se aplicaban sobre información contenida en almacenes de datos. De hecho, muchas grandes empresas e instituciones han creado y alimentan bases de datos especialmente diseñadas para proyectos de minería de datos en las que centralizan información potencialmente útil de todas sus áreas de negocio. No obstante, actualmente está cobrando una importancia cada vez mayor la minería de datos desestructurados como información contenida en ficheros de texto, en Internet, etc.

APLICACIONES DE LA MINERÍA DE DATOS

Cada año, en los diferentes congresos, simposios y talleres que se realizan en el mundo se reúnen investigadores con aplicaciones muy diversas. Sobre todo en los Estados Unidos, el data mining se ha ido incorporando a la vida de empresas, gobiernos, universidades, hospitales y diversas organizaciones que están interesadas en explorar sus bases de datos. Podemos decir que "en data mining cada caso es un caso". [Molina, 2002]

Sin embargo, en términos generales, el proceso se compone de cuatro etapas principales:

1. **Determinación de los objetivos.** Trata de la delimitación de los objetivos que el cliente desea bajo la orientación del especialista en data mining.
2. **Preprocesamiento de los datos**. Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de data mining.
3. **Determinación del modelo.** Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.
4. **Análisis de los resultados**. Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

Las principales metodologías utilizadas por los analistas para la realización de proyectos de Data Mining son: CRISP-DM y SEMMA. Estas metodologías comparten la misma esencia estructurando el proyecto de Data Mining en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de Data Mining en un proceso iterativo e interactivo.

Ejemplos de Usos

A continuación se describen varios ejemplos donde se ha visto involucrado el data mining. Se han seleccionado de diversos dominios y con diversos objetivos para observar su potencial. Respecto a los modelos inteligentes, se ha comprobado que en ellos se utilizan principalmente árboles y reglas de decisión, reglas de asociación, redes neuronales, redes bayesianas, conjuntos aproximados (rough sets), algoritmos de agrupación (clustering), máquinas de soporte vectorial, algoritmos genéticos y lógica difusa.

En el gobierno

Para el FBI analizar las bases de datos comerciales para detectar terroristas.
Departamento de Justicia debe introducirse en la vasta cantidad de datos comerciales referentes a los hábitos y preferencias de compra de los consumidores, con el fin de descubrir potenciales terroristas antes de que ejecuten una acción. Algunos expertos aseguran que, con esta información, el FBI uniría todas las bases de datos y permitirá saber si una persona fuma, qué talla y tipo de ropa usa, su registro de arrestos, su salario, las revistas a las que está suscrito, su altura y peso, sus contribuciones a la Iglesia, grupos políticos u organizaciones no gubernamentales, sus enfermedades crónicas (como diabetes o asma), los libros que lee, los productos de supermercado que compra, si tomó clases de vuelo o si tiene cuentas de banco abiertas, entre otros.

En la empresa

Detección de fraudes en las tarjetas de crédito.
Examinar transacciones, propietarios de tarjetas y datos financieros para detectar y mitigar fraudes. En un principio para detectar fraudes en tarjetas de crédito, luego incorporar las tarjetas comerciales, de combustibles y de débito.

Descubriendo el porqué de la deserción de clientes de una compañía operadora de telefonía móvil. Este estudio fue desarrollado en una operadora española que básicamente situó sus objetivos en dos puntos: el análisis del perfil de los clientes que se dan de baja y la predicción del comportamiento de sus nuevos clientes. Se analizaron los diferentes históricos de clientes que habían abandonado la operadora y de clientes que continuaban con su servicio. También se analizaron las variables personales de cada cliente (estado civil, edad, sexo, nacionalidad, etc.). De igual forma se estudiaron, para cada cliente, la morosidad, la frecuencia y el horario de uso del servicio, los descuentos y el porcentaje de llamadas locales, interprovinciales,

internacionales y gratuitas. Al contrario de lo que se podría pensar, los clientes que abandonaban la operadora generaban ganancias para la empresa; sin embargo, una de las conclusiones más importantes radicó en el hecho de que los clientes que se daban de baja recibían pocas promociones y registraban un mayor número de incidencias respecto a la media. De esta forma se recomendó a la operadora hacer un estudio sobre sus ofertas y analizar profundamente las incidencias recibidas por esos clientes. Al descubrir el perfil que presentaban, la operadora tuvo que diseñar un trato más personalizado para sus clientes actuales con esas características. Para poder predecir el comportamiento de sus nuevos clientes se diseñó un sistema de predicción basado en la cantidad de datos que se podía obtener de los nuevos clientes comparados con el comportamiento de clientes anteriores.

Prediciendo el tamaño de las audiencias televisivas.

La British Broadcasting Corporation (BBC) del Reino Unido emplea un sistema para predecir el tamaño de las audiencias televisivas para un programa propuesto, así como el tiempo óptimo de exhibición [Brachman y otros, 1996]. El sistema utiliza redes neuronales y árboles de decisión aplicados a datos históricos de la cadena para determinar los criterios que participan según el programa que hay que presentar. La versión final se desempeña tan bien como un experto humano con la ventaja de que se adapta más fácilmente a los cambios porque es constantemente reentrenada con datos actuales.

En la universidad

Conociendo si los recien titulados de una universidad llevan a cabo actividades profesionales relacionadas con sus estudios. Se hizo un estudio sobre los recién titulados de la carrera de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Chihuahua II, [1] en Méjico [Rodas, 2001]. Se quería observar si sus recién titulados se insertaban en actividades profesionales relacionadas con sus estudios y, en caso negativo, se buscaba saber el perfil que caracterizó a los exalumnos durante su estancia en la universidad. El objetivo era saber si con los planes de estudio de la universidad y el aprovechamiento del alumno se hacía una buena inserción laboral o si existían otras variables que participaban en el proceso. Dentro de la información considerada estaba el sexo, la edad, la escuela de procedencia, el desempeño académico, la zona económica donde tenía su vivienda y la actividad profesional, entre otras variables. Mediante la aplicación de conjuntos aproximados se descubrió que existían cuatro variables que determinaban la adecuada inserción laboral, que son citadas de acuerdo con su importancia: zona económica donde habitaba el estudiante, colegio de dónde provenía, nota al ingresar y promedio final al

salir de la carrera. A partir de estos resultados, la universidad tendrá que hacer un estudio socioeconómico sobre grupos de alumnos que pertenecían a las clases económicas bajas para dar posibles soluciones, debido a que tres de las cuatro variables no dependían de la universidad.

En investigaciones espaciales

Proyecto SKYCAT.

Durante seis años, el Second Palomar Observatory Sky Survey (POSS-II) coleccionó tres terabytes de imágenes que contenían aproximadamente dos millones de objetos en el cielo. Tres mil fotografías fueron digitalizadas a una resolución de 16 bits por píxel con 23.040 x 23.040 píxeles por imagen. El objetivo era formar un catálogo de todos esos objetos.

El sistema Sky Image Cataloguing and Analysis Tool (SKYCAT) se basa en técnicas de agrupación (clustering) y árboles de decisión para poder clasificar los objetos en estrellas, planetas, sistemas, galaxias, etc. con una alta confiabilidad [Fayyad y otros, 1996]. Los resultados han ayudado a los astrónomos a descubrir dieciséis nuevos quásars con corrimiento hacia el rojo que los incluye entre los objetos más lejanos del universo y, por consiguiente, más antiguos. Estos quásars son difíciles de encontrar y permiten saber más acerca de los orígenes del universo.

En los clubes deportivos

El AC de Milan utiliza un sistema inteligente para prevenir lesiones.

Esta temporada el club comenzará a usar redes neuronales para prevenir lesiones y optimizar el acondicionamiento de cada atleta. Esto ayudará a seleccionar el fichaje de un posible jugador o a alertar al médico del equipo de una posible lesión. El sistema, creado por Computer Associates International, es alimentado por datos de cada jugador, relacionados con su rendimiento, alimentación y respuesta a estímulos externos, que se obtienen y analizan cada quince días. El jugador lleva a cabo determinadas actividades que son monitoreadas por veinticuatro sensores conectados al cuerpo y que transmiten señales de radio que posteriormente son almacenadas en una base de datos. Actualmente el sistema dispone de 5.000 casos registrados que permiten predecir alguna posible lesión. Con ello, el club intenta ahorrar dinero evitando comprar jugadores que presenten una alta probabilidad de lesión, lo que haría incluso renegociar su contrato. Por otra parte, el sistema pretende encontrar las diferencias entre las lesiones de atletas de ambos sexos, así como saber si una determinada lesión se relaciona con el estilo de juego de un país concreto donde se practica el fútbol.

Los equipos de la NBA utilizan aplicaciones inteligentes para apoyar a su cuerpo de entrenadores.

El Advanced Scout es un software que emplea técnicas de data mining y que han desarrollado investigadores de IBM para detectar patrones estadísticos y eventos raros. Tiene una interfaz gráfica muy amigable orientada a un objetivo muy específico: analizar el juego de los equipos de la National Basketball Association (NBA). El software utiliza todos los registros guardados de cada evento en cada juego: pases, encestes, rebotes y doble marcaje (doublé team) a un jugador por el equipo contrario, entre otros. El objetivo es ayudar a los entrenadores a aislar eventos que no detectan cuando observan el juego en vivo o en película. Un resultado interesante fue uno hasta entonces no observado por los entrenadores de los Knicks de Nueva York. El doble marcaje a un jugador puede generalmente dar la oportunidad a otro jugador de encestar más fácilmente. Sin embargo, cuando los Bulls de Chicago jugaban contra los Knicks, se encontró que el porcentaje de encestes después de que al centro de los Knicks, Patrick Ewing, le hicieran doble marcaje era extremadamente bajo, indicando que los Knicks no reaccionaban correctamente a los dobles marcajes.

Para saber el porqué, el cuerpo de entrenadores estudió cuidadosamente todas las películas de juegos contra Chicago. Observaron que los jugadores de Chicago rompían su doble marcaje muy rápido de tal forma que podían tapar al encestador libre de los Knicks antes de prepararse para efectuar su tiro. Con este conocimiento, los entrenadores crearon estrategias alternativas para tratar con el doble marcaje. La temporada pasada, IBM ofreció el Advanced Scout a la NBA, que se convirtió así en un patrocinador corporativo. La NBA dio a sus veintinueve equipos la oportunidad de aplicarlo. Dieciocho equipos lo están haciendo hasta el momento obteniendo descubrimientos interesantes.

Negocios

La minería de datos puede contribuir significativamente en las aplicaciones de administración empresarial basada en la relación con el cliente. En lugar de contactar con el cliente de forma indiscriminada a través de un centro de llamadas o enviando cartas, sólo se contactará con aquellos que se perciba que tienen una mayor probabilidad de responder positivamente a una determinada oferta o promoción.

Por lo general, las empresas que emplean minería de datos ven rápidamente el retorno de la inversión, pero también reconocen que el número de modelos predictivos desarrollados puede crecer muy rápidamente.

En lugar de crear modelos para predecir qué clientes pueden cambiar, la empresa podría construir modelos separados para cada región y/o para cada tipo de cliente. También puede querer determinar qué clientes van a ser rentables durante una ventana de tiempo (una quincena, un mes, ...) y sólo enviar las ofertas a las personas que es probable que sean rentables. Para mantener esta cantidad de modelos, es necesario gestionar las versiones de cada modelo y pasar a una minería de datos lo más automatizada posible.

Hábitos de compra en supermercados

El ejemplo clásico de aplicación de la minería de datos tiene que ver con la detección de hábitos de compra en supermercados. Un estudio muy citado detectó que los viernes había una cantidad inusualmente elevada de clientes que adquirían a la vez pañales y cerveza. Se detectó que se debía a que dicho día solían acudir al supermercado padres jóvenes cuya perspectiva para el fin de semana consistía en quedarse en casa cuidando de su hijo y viendo la televisión con una cerveza en la mano. El supermercado pudo incrementar sus ventas de cerveza colocándolas próximas a los pañales para fomentar las ventas compulsivas.

Patrones de fuga

Un ejemplo más habitual es el de la detección de patrones de fuga. En muchas industrias —como la banca, las telecomunicaciones, etc.— existe un comprensible interés en detectar cuanto antes aquellos clientes que puedan estar pensando en rescindir sus contratos para, posiblemente, pasarse a la competencia. A estos clientes —y en función de su valor— se les podrían hacer ofertas personalizadas, ofrecer promociones especiales, etc., con el objetivo último de retenerlos. La minería de datos ayuda a determinar qué clientes son los más proclives a darse de baja estudiando sus patrones de comportamiento y comparándolos con muestras de clientes que, efectivamente, se dieron de baja en el pasado.

Fraudes

Un caso análogo es el de la detección de transacciones de blanqueo de dinero o de fraude en el uso de tarjetas de crédito o de servicios de telefonía móvil e, incluso, en la relación de los contribuyentes con el fisco. Generalmente, estas operaciones fraudulentas o ilegales suelen seguir patrones característicos que permiten, con cierto grado de probabilidad, distinguirlas de las legítimas y desarrollar así mecanismos para tomar medidas rápidas frente a ellas.

Recursos humanos

La minería de datos también puede ser útil para los departamentos de recursos humanos en la identificación de las características de sus empleados de mayor éxito. La información obtenida puede ayudar a la contratación de personal, centrándose en los esfuerzos de sus empleados y los resultados obtenidos por éstos. Además, la ayuda ofrecida por las aplicaciones para Dirección estratégica en una empresa se traducen en la obtención de ventajas a nivel corporativo, tales como mejorar el margen de beneficios o compartir objetivos; y en la mejora de las decisiones operativas, tales como desarrollo de planes de producción o gestión de mano de obra.

Comportamiento en Internet

También es un área en boga el del análisis del comportamiento de los visitantes —sobre todo, cuando son clientes potenciales— en una página de Internet. O la utilización de la información —obtenida por medios más o menos legítimos— sobre ellos para ofrecerles propaganda adaptada específicamente a su perfil. O para, una vez que adquieren un determinado producto, saber inmediatamente qué otro ofrecerle teniendo en cuenta la información histórica disponible acerca de los clientes que han comprado el primero.

Terrorismo

La minería de datos ha sido citada como el método por el cual la unidad Able Danger del Ejército de los EE. UU. había identificado al líder de los atentados del 11 de septiembre de 2001, Mohammed Atta, y a otros tres secuestradores del "11-S" como posibles miembros de una célula de Al Qaeda que operan en los EE. UU. más de un año antes del ataque. Se ha sugerido que tanto la Agencia Central de Inteligencia y sus homóloga canadiense, Servicio de Inteligencia y Seguridad Canadiense, también han empleado este método.1

Juegos

Desde comienzos de la década de 1960, con la disponibilidad de oráculos para determinados juegos combinacionales, también llamados finales de juego de tablero (por ejemplo, para las tres en raya o en finales de ajedrez) con cualquier configuración de inicio, se ha abierto una nueva área en la minería de datos que consiste en la extracción de estrategias utilizadas por personas para estos oráculos. Los planteamientos actuales sobre reconocimiento de patrones, no parecen poder aplicarse con éxito al funcionamiento de estos oráculos. En su lugar, la producción de patrones perspicaces se basa en una amplia experimentación con bases de datos sobre

esos finales de juego, combinado con un estudio intensivo de los propios finales de juego en problemas bien diseñados y con conocimiento de la técnica (datos previos sobre el final del juego). Ejemplos notables de investigadores que trabajan en este campo son Berlekamp en el juego de puntos-y-cajas (o Timbiriche) y John Nunn en finales de ajedrez.

Ciencia e Ingeniería

En los últimos años la minería de datos se está utilizando ampliamente en diversas áreas relacionadas con la ciencia y la ingeniería. Algunos ejemplos de aplicación en estos campos son:

Genética

En el estudio de la genética humana, el objetivo principal es entender la relación cartográfica entre las partes y la variación individual en las secuencias del ADN humano y la variabilidad en la susceptibilidad a las enfermedades. En términos más llanos, se trata de saber cómo los cambios en la secuencia de ADN de un individuo afectan al riesgo de desarrollar enfermedades comunes (como por ejemplo el cáncer). Esto es muy importante para ayudar a mejorar el diagnóstico, prevención y tratamiento de las enfermedades. La técnica de minería de datos que se utiliza para realizar esta tarea se conoce como "reducción de dimensionalidad multifactorial".2

Ingeniería eléctrica

En el ámbito de la ingeniería eléctrica, las técnicas minería de datos han sido ampliamente utilizadas para monitorizar las condiciones de las instalaciones de alta tensión. La finalidad de esta monitorización es obtener información valiosa sobre el estado del aislamiento de los equipos. Para la vigilancia de las vibraciones o el análisis de los cambios de carga en transformadores se utilizan ciertas técnicas para agrupación de datos (clustering) tales como los Mapas Auto-Organizativos (SOM, Self-organizing map). Estos mapas sirven para detectar condiciones anormales y para estimar la naturaleza de dichas anomalías.3

Análisis de gases

También se han aplicado técnicas de minería de datos para el análisis de gases disueltos (DGA, Dissolved gas analysis) en transformadores eléctricos. El análisis de gases disueltos se conoce desde hace mucho tiempo como herramienta para diagnosticar transformadores. Los Mapas Auto-Organizativos (SOM) se utilizan para analizar datos y determinar tendencias que podrían pasarse por alto utilizando las técnicas clásicas DGA.

Minería de datos y otras disciplinas análogas

Suscita cierta polémica el definir las fronteras existentes entre la minería de datos y disciplinas análogas, como pueden serlo la estadística, la inteligencia artificial, etc. Hay quienes sostienen que la minería de datos no es sino estadística envuelta en una jerga de negocios que la conviertan en un producto vendible. Otros, en cambio, encuentran en ella una serie de problemas y métodos específicos que la hacen distinta de otras disciplinas.

El hecho es, que en la práctica la totalidad de los modelos y algoritmos de uso general en minería de datos —redes neuronales, árboles de regresión y clasificación, modelos logísticos, análisis de componentes principales, etc. — gozan de una tradición relativamente larga en otros campos.

De la estadística

Ciertamente, la minería de datos bebe de la estadística, de la que toma las siguientes técnicas:

- Análisis de varianza, mediante el cual se evalúa la existencia de diferencias significativas entre las medias de una o más variables continúas en poblaciones distintas.

- Regresión: define la relación entre una o más variables y un conjunto de variables predictoras de las primeras.

- Prueba chi-cuadrado: por medio de la cual se realiza el contraste la hipótesis de dependencia entre variables.

- Análisis de agrupamiento o clustering: permite la clasificación de una población de individuos caracterizados por múltiples atributos (binarios, cualitativos o cuantitativos) en un número determinado de grupos, con base en las semejanzas o diferencias de los individuos.

- Análisis discriminante: permite la clasificación de individuos en grupos que previamente se han establecido, permite encontrar la regla de clasificación de los elementos de estos grupos, y por tanto una mejor identificación de cuáles son las variables que definan la pertenencia al grupo.

- Series de tiempo: permite el estudio de la evolución de una variable a través del tiempo para poder realizar predicciones, a partir de ese conocimiento y bajo el supuesto de que no van a producirse cambios estructurales.

De la informática

De la informática toma las siguientes técnicas:

- Algoritmos genéticos: Son métodos numéricos de optimización, en los que aquella variable o variables que se pretenden optimizar junto con las variables de estudio constituyen un segmento de información. Aquellas configuraciones de las variables de análisis que obtengan mejores valores para la variable de respuesta, corresponderán a segmentos con mayor capacidad reproductiva. A través de la reproducción, los mejores segmentos perduran y su proporción crece de generación en generación. Se puede además introducir elementos aleatorios para la modificación de las variables (mutaciones). Al cabo de cierto número de iteraciones, la población estará constituida por buenas soluciones al problema de optimización, pues las malas soluciones han ido descartándose, iteración tras iteración.

- Inteligencia Artificial: Mediante un sistema informático que simula un sistema inteligente, se procede al análisis de los datos disponibles. Entre los sistemas de Inteligencia Artificial se encuadrarían los Sistemas Expertos y las Redes Neuronales.

- Sistemas Expertos: Son sistemas que han sido creados a partir de reglas prácticas extraídas del conocimiento de expertos. Principalmente a base de inferencias o de causa-efecto.

- Sistemas Inteligentes: Son similares a los sistemas expertos, pero con mayor ventaja ante nuevas situaciones desconocidas para el experto.

- Redes neuronales: Genéricamente, son métodos de proceso numérico en paralelo, en el que las variables interactúan mediante transformaciones lineales o no lineales, hasta obtener unas salidas. Estas salidas se contrastan con los que tenían que haber salido, basándose en unos datos de prueba, dando lugar a un proceso de retroalimentación mediante el cual la red se reconfigura, hasta obtener un modelo adecuado.

HERRAMIENTAS DE SOFTWARE

Existen muchas herramientas de software para el desarrollo de modelos de minería de datos tanto libres como comerciales como, por ejemplo:

- Dynamic Data Web
- KXEN
- KNIME
- Orange
- RapidMiner
- SPSS Clementine

El data mining es una de las principales herramientas que se utilizan dentro de los programas de gestión del conocimiento como soporte a la toma de decisiones.

El fin es la extracción de información oculta o análisis de datos mediante técnicas estadísticas de grandes bases de datos.

Las herramientas de data mining o minería de datos pueden responder a preguntas de negocios empresariales a priori no planteadas o que pueden consumir demasiado tiempo para ser resueltas.

Los programas de gestión del conocimiento se complementan con distintas herramientas además del data mining, como puede ser el data warehousing o el groupware. El data mining, como herramienta de búsqueda de información, se utiliza como sistema de apoyo a la toma de decisiones de las altas direcciones de las empresas.

Las técnicas de data mining se centran en analizar el gran volumen de datos, que en una primera selección pueden ser pertinentes, pero que la aplicación de técnicas de selección ceñida a unas determinada demanda, reduce el tamaño de los datos eligiendo las variables más influyentes en el problema.

En definitiva, la minería de datos es una tecnología usada para descubrir información oculta y desconocida, pero potencialmente útil, a partir de las fuentes de información de la propia empresa. Obtiene un conocimiento de un negocio, utilizando técnicas de clustering, redes neuronales, árboles de decisión y reglas de asociación etc.

En resumen, el datamining se presenta como una tecnología emergente, con varias ventajas: por un lado, resulta un buen punto de encuentro entre los investigadores y las personas de negocios; por otro, ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios. Además, no hay duda de que

trabajar con esta tecnología implica cuidar un sinnúmero de detalles debido a que el producto final involucra "toma de decisiones".

En el artículo Data Mining: Torturando a los datos hasta que confiesen, Luis Carlos Molina proporciona una visión muy clarificadora sobre la minería de datos, incluyendo interesantes ejemplos de aplicaciones de la misma. Recomendamos su lectura.

Si no está familiarizado con el concepto de Datamining, puede resultarle útil, además, examinar las siguientes definiciones:

Datamart

Datawarehouse

Cuadro de Mando Integral

Sistemas de Soporte a la Decisión (DSS)

Sistemas de Información Ejecutiva (EIS)

Software DTM - Minería de Datos y Textos (Lebart)

Estadística Exploratoria Multidimensional para datos complejos que incluyen datos numéricos y textuales.

Software bastante completo en los primeros métodos de tratamiento de dato (estadística multivariada). También incluye análisis estadístico de texto. Weka implementa redes neuronales, árboles y bayes.

Aspectos específicos:

- Complementariedad de las técnicas de visualización (Análisis de componentes principales, Análisis de correspondencias simple y múltiple) y la clasificación automática (método mixto que combina clasificación jerárquica [criterio de Ward] y centros móviles [k-means]; mapas autoorganizados de Kohonen [redes neuronales SOM).
- Validación de las técnicas de visualización: Re-muestreo (bootstrap, bootstrap parcial, bootstrap total, bootstrap sobre variables). Tres opciones para el bootstrap total: Tipo 1: simple corrección de señal para los ejes. Tipo 2: como

tipo 1, + corrección de la rotación de ejes. Tipo 3 rotaciones procrustéennes para acercar a los réplicaciones de la muestra inicial.

- Análisis de asociación y métodos vecinos. Mapas de Kohonen (SOM).

***Limitaciones de esta versión*:** 22.500 filas (individuos, líneas), 1.000 variables (numéricas o nominales), 100.000 caracteres para las respuestas de un individuo a cuestiones abiertas (textos).

Fuente: http://ses.telecom-paristech.fr/lebart/

Otros softwares

- Weka: Magnífica suite de minería de datos de libre distribución.
- MLC++: Conjunto de librerías y utilidades de minería de datos.
- Xelopes: Librería con licencia pública GNU para el desarrollo de aplicaciones de minería de datos.
- C4.5: Sistema clásico de aprendizaje de árboles de decisión.
- FOIL: Software que permite el aprendizaje de modelos relacionales.

CONCLUSIONES

Nuestra capacidad para almacenar datos ha crecido en los últimos años a velocidades exponenciales. En contrapartida, nuestra 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. Descubrir nuevos caminos que nos ayuden en la identificación de interesantes estructuras en los datos es una de las tareas fundamentales en el *data mining*.

En el ámbito comercial, resulta interesante encontrar patrones ocultos de consumo de los clientes para poder explorar nuevos horizontes. Saber que un vehículo deportivo corre un riesgo de accidente casi igual al de un vehículo normal cuando su dueño tiene un segundo vehículo en casa ayuda a crear nuevas estrategias comerciales para ese grupo de clientes. Asimismo, predecir el comportamiento de un futuro cliente, basándose en los datos históricos de clientes que presentaron el mismo perfil, ayuda a poder retenerlo durante el mayor tiempo posible.

Las herramientas comerciales de *data mining* que existen actualmente en el mercado son variadas y excelentes. Las hay orientadas al estudio del web o al análisis de documentos o de clientes de supermercado, mientras que otras son de uso más general. Su correcta elección depende de la necesidad de la empresa y de los objetivos a corto y largo plazo que pretenda alcanzar. La decisión de seleccionar una solución de *data mining* no es una tarea simple. Es necesario consultar a expertos en el área con vista a seleccionar la más adecuada para el problema de la empresa.

Como se ha visto a lo largo del este artículo, son muchas las áreas, técnicas, estrategias, tipos de bases de datos y personas que intervienen en un proceso de *data mining*. Los negocios requieren que las soluciones tengan una integración transparente en un ambiente operativo. Esto nos lleva a la necesidad de establecer estándares para hacer un ambiente interoperable, eficiente y efectivo. Esfuerzos en este sentido se están desarrollando actualmente. En Grossman y otros (2002) se exponen algunas iniciativas para estos estándares, incluyendo aspectos en:

Modelos: para representar datos estadísticos y de *data mining*.

Atributos: para representar la limpieza, transformación y agregación de atributos usados como entrada en los modelos.

Interfaces y API: para facilitar la integración con otros lenguajes o aplicaciones de software y API.

Configuración: para representar parámetros internos requeridos para construir y usar los modelos.

Procesos: para producir, desplegar y usar modelos.

Datos remotos y distribuidos: para analizar y explorar datos remotos y distribuidos.

En resumen, el *data mining* se presenta como una tecnología emergente, con varias ventajas: por un lado, resulta un buen punto de encuentro entre los investigadores y las personas de negocios; por otro, ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios. Además, no hay duda de que trabajar con esta tecnología implica cuidar un sinnúmero de detalles debido a que el producto final involucra "toma de decisiones".

FUENTES CONSULTADAS

- http://micesari.blogspot.com/2008/04/dtm-minera-de-datos-y-textos-lebart.html
- http://ses.telecom-paristech.fr/lebart/
- http://ai.frm.utn.edu.ar/micesari/files/01_Matilde_Mineria_datos.pdf
- http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_Vallejos.pdf