Capítulo 1

Data Mining & Knowledge Discovery in Databases (KDD)

1. Introducción

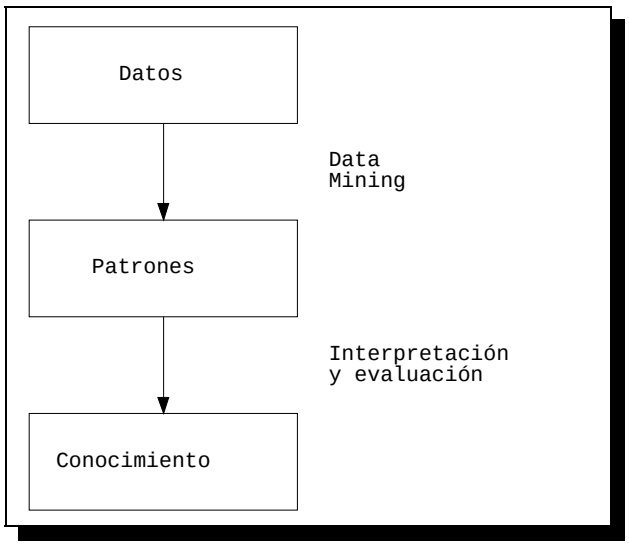
"Data Mining (also called Knowledge Discovery in Databases)
is the efficient discovery of previously unknown patterns in large databases"

Rakesh Agrawal & John C. Shafer: "Parallel Mining of Association Rules"
IEEE Transactions on Knowledge and Data Engineering, December 1996

Data Mining es un término genérico que engloba resultados de investigación, técnicas y herramientas usadas para extraer información útil de grandes bases de datos. Los algoritmos de *Data Mining* se enmarcan en el proceso completo de extracción de información conocido como KDD [*Knowledge Discovery in Databases*], que se encarga además de preparación de los datos y de la interpretación de los resultados obtenidos. No debemos olvidar que de la simple aplicación de técnicas de *Data Mining* sólo se obtienen patrones que no sirven de gran cosa mientras no se les encuentre significado [*data dredging*].

KDD se ha definido como la extracción no trivial de información potencialmente útil a partir de un gran volumen de datos en el cual la información está implícita (aunque no se conoce previamente). Se trata de interpretar grandes cantidades de datos y encontrar relaciones o patrones. Para conseguirlo harán falta técnicas de aprendizaje [*Machine Learning*], estadística y bases de datos.

Las investigaciones en estos temas incluyen análisis estadístico de datos, técnicas de representación del conocimiento, razonamiento basado en casos [CBR: *Case Based Reasoning*], razonamiento aproximado, adquisición de conocimiento, redes neuronales y visualización de datos. Tareas comunes en KDD son la inducción de reglas, los problemas de clasificación y clustering, el reconocimiento de patrones, el modelado predictivo, la detección de dependencias, etc..



KDD

KDD: "The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"

Fayyad, Piatetsky-Shapiro & Smyth: "From data mining to knowledge discovery: An overview"
Advances in Knowledge Discovery and Data Mining (AAAI / MIT Press, 1996)

Los datos recogen un conjunto de hechos (una base de datos) y los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese subconjunto). KDD involucra un **proceso iterativo e interactivo** de búsqueda de modelos, patrones o parámetros. Los patrones descubiertos han de ser válidos, novedosos para el sistema (para el usuario siempre que sea posible) y potencialmente útiles.

Se han de definir medidas cuantitativas para los patrones obtenidos (precisión, utilidad, beneficio obtenido...). Se debe establecer alguna medida de interés [*interestingness*] que considere la validez, utilidad y simplicidad de los patrones obtenidos mediante alguna de las técnica de *Data Mining*. El objetivo final de todo esto es incorporar el conocimiento obtenido en algún sistema real, tomar decisiones a partir de los resultados alcanzados o, simplemente, registrar la información conseguida y suministrársela a quien esté interesado.

En muchos lugares se han preocupado de recopilar gran cantidad de información de todo tipo. Es fácil digitalizar información, ya no es excesivamente caro almacenarla y, en principio, los datos recogidos creemos que pueden llegar a sernos útiles.

Ha llegado un momento en el que disponemos de tanta información que nos vemos incapaces de sacarle provecho. Los datos tal cual se almacenan [*raw data*] no suelen proporcionar beneficios directos. Su valor real reside en la información que podamos extraer de ellos: información que nos ayude a tomar decisiones o a mejorar nuestra comprensión de los fenómenos que nos rodean.

El análisis de la información recopilada (por ejemplo, en un experimento científico) es habitual que sea un proceso completamente manual (basado por lo general en técnicas estadísticas). Sin embargo, cuando la cantidad de datos de los que disponemos aumenta la resolución manual del problema se hace intratable. Aquí es donde entra en juego el conjunto de técnicas de análisis automático al que nos referimos al hablar de *Data Mining* o *KDD*.

Hasta ahora, los mayores éxitos en *Data Mining* 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. Data Mining

Como ya se ha comentado, las técnicas de *Data Mining* (una etapa dentro del proceso completo de KDD) intentan obtener patrones o modelos a partir de los datos recopilados. Decidir si los modelos obtenidos son útiles o no suele requerir una valoración subjetiva por parte del usuario. Los algoritmos de *Data Mining* suelen tener tres componentes:

- El MODELO, que contiene parámetros que han de fijarse a partir de los datos de entrada.
- El CRITERIO DE PREFERENCIA, que sirve para comparar modelos alternativos
- El ALGORITMO DE BÚSQUEDA (como cualquier otro programa de IA).

El criterio de preferencia suele ser algún tipo de heurística y los algoritmos de búsqueda empleados suelen ser los mismos que en otros programas de IA. Las principales diferencias entre los algoritmos de *Data Mining* se hallan en el modelo de representación escogido y la función del mismo (el objetivo perseguido).

Por ejemplo, un modelo de clasificación basado en árboles de decisión suele utilizar un algoritmo greedy (una búsqueda sin vuelta atrás) y una heurística que favorezca la construcción de árboles de decisión con pocos nodos.

Las herramientas de Data Mining empleados en el proceso de KDD se pueden clasificar en dos grandes grupos: técnicas de verificación (en las que el sistema se limita a comprobar hipótesis suministradas por el usuario) y métodos de descubrimiento (en los que se han de encontrar patrones potencialmente interesantes de forma automática, incluyendo en este grupo todas las técnicas de predicción).

El resultado obtenido con la aplicación de algoritmos de *Data Mining* (pertenecientes al segundo grupo, el de técnicas de descubrimiento) puede ser de carácter descriptivo o predictivo. Las predicciones nos sirven para prever el comportamiento futuro de algún tipo de entidad mientras que una descripción nos puede ayudar a su comprensión. De hecho, los modelos predictivos pueden ser descriptivos (hasta donde sean comprensibles por personas) y los modelos descriptivos pueden emplearse para realizar predicciones.

Algunos de los objetivos perseguidos al aplicar técnicas de Data Mining en grandes bases de datos son los siguientes:

- ✓ 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).

- ✓ REGRESIÓN: Se persigue la obtención de un modelo que permita predecir el valor numérico de alguna variable.

- ✓ AGRUPAMIENTO [*clustering*]: Hace corresponder cada caso a una clase, con la peculiaridad de que las clases se obtienen directamente de los datos de entrada utilizando medidas de similaridad.

- ✓ RESUMEN [*summarization*]: Se obtienen representaciones compactas para subconjuntos de los datos de entrada (vg: análisis interactivo de datos, generación automática de informes, visualización de datos...).

- ✓ MODELADO DE DEPENDENCIAS: Se obtienen descripciones de dependencias existentes entre variables. El análisis de relaciones (vg. reglas de asociación), en el que se determinan relaciones existentes entre elementos de una base de datos, podría considerarse un caso particular de modelado de dependencias.

- ✓ ANÁLISIS DE SECUENCIAS [*deviation and trend analysis*]: Se intenta modelar la evolución temporal de alguna variable, con fines descriptivos o predictivos.

3. Materias relacionadas

Los aspectos abarcados por el proceso de KDD incluyen desde el almacenamiento eficiente de los datos hasta la visualización de los resultados. Es esencial que los algoritmos empleados en Data Mining sean eficientes, escalables y robustos a la hora de manipular grandes cantidades de información con ruido.

3.1 Estadística

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

3.2 Ingeniería del conocimiento: Modelos de representación del conocimiento

Algunos de los modelos de representación del conocimiento utilizados en técnicas de *Data Mining* son los árboles de decisión, las reglas de producción o las redes bayesianas. El modelo escogido determina la flexibilidad de la representación y la facilidad con la que una persona pueda interpretar el conocimiento obtenido. Los modelos más complejos pueden adaptarse mejor a los datos aunque suelen ser más difíciles de interpretar, por lo que en la práctica muchas veces se utilizan modelos simplificados.

Las técnicas para el manejo de la incertidumbre se hallan asociadas a los modelos de representación del conocimiento y son esenciales en KDD, ya que los datos suelen incluir errores (ruido) y ser incompletos.

3.3 Bases de datos

"I never waste memory on things that can easily be stored and retrieved from elsewhere"

Albert Einstein, 1879-1955

Habitualmente, los algoritmos empleados en Inteligencia Artificial (en *Machine Learning* para ser más concretos) y en reconocimiento de patrones presuponen que los datos sobre los que se aplican han de cargarse en la memoria principal del ordenador. Cuando tenemos tantos datos que no podemos cargarlos en memoria no nos queda más remedio que recurrir a técnicas empleadas en bases de datos, otro campo fundamental para las investigaciones en *Data Mining*.

4. Referencias

Michael J. A. Berry & Gordon Linoff
"Data Mining Techniques: for Marketing, Sales, and Customer Support"
John Wiley and Sons, 1997

Un libro destinado a ejecutivos y managers, para que éstos se familiaricen con distintas técnicas de Data Mining existentes, su uso y sus limitaciones. Se cubre desde el análisis de las transacciones comerciales [*basket data analysis*] hasta la utilización de árboles de decisión, técnicas de clustering, redes neuronales y algoritmos genéticos.

R.J. Brachman, T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro & E. Simoudis
"Mining Business Databases"
Communications of the ACM, November 1996

En este artículo se comenta el uso de técnicas de Data Mining en distintos ámbitos relacionados con el mundo de los negocios: marketing [vg: *database marketing* o *mailshot response*], inversiones financieras, detección de fraudes, CAM [*computer aided manufacturing*], redes de telecomunicaciones...

Oren Erzioni
"The World-Wide-Web: Quagmire or Gold Mine?"
Communications of the ACM, November 1996

El autor considera que Internet, la WWW, ofrece grandes oportunidades para la extracción automática de conocimiento útil. Según él, "Web Mining" es factible y las técnicas de Data Mining no han de limitarse a bases de datos bien estructuradas.

Usama Fayyad, David Haussler & Paul Stolorz
"Mining Scientific Data"
Communications of the ACM, November 1996

Una de las aplicaciones más interesantes de KDD es el análisis de datos obtenidos en experimentos científicos, permitiendo de ese modo que los científicos se centren en tareas más creativas (como la formación de teorías e hipótesis) y dejando que las máquinas realicen todo el trabajo rutinario. En este artículo se ilustra el potencial del proceso de KDD con algunas aplicaciones reales (como SKICAT, JARtool o CONQUEST).

Usama Fayyad, Gregory Piatetsky-Shapiro & Padhraic Smyth
"The KDD Process for Extracting Useful Knowledge from Volumes of Data"
Communications of the ACM, November 1996

El artículo ofrece una visión general de aquello a lo que nos referimos al hablar de KDD, revisa algunos temas relacionados y concluye con una enumeración de los desafíos a los que han de enfrentarse las investigaciones: gran cantidad de datos, interacción con el usuario, información incompleta, redundancias, técnicas incrementales...

Usama Fayyad & Ramasamy Uthurusamy
"Data Mining and Knowledge Discovery in Databases"
Communications of the ACM, November 1996

Este breve artículo sirve de introducción a una sección especial de la publicación más difundida de la ACM dedicada a Data Mining. Fayyad (de Microsoft) y Uthurusamy (de General Motors) ponen de relieve lo interesante que puede ser la aplicación de técnicas de KDD, un campo prometedor en el que muchos investigadores están trabajando actualmente.

Clark Glymour, David Madigan, Daryl Pregibon & Padhraic Smyth
"Statistical Inference and Data Mining"
Communications of the ACM, November 1996

El artículo trata de lo que la Estadística puede aportar a las técnicas de *Data Mining*: básicamente, la evaluación de las hipótesis generadas y de los resultados obtenidos.

Tomasz Imielinski & Heikki Mannila
"A Database Perspective on Knowledge Discovery"
Communications of the ACM, November 1996

En esta ocasión se ven los métodos empleados en Data Mining desde otra perspectiva, la de las bases de datos. En este artículo se ponen de manifiesto las limitaciones de SQL a la hora de construir aplicaciones de Data Mining y se expone la necesidad de idear lenguajes de consulta más potentes: "El modelo relacional representa el lenguaje ensamblador de los sistemas modernos (y futuros) de bases de datos" [C.J. Date].

W. H. Inmon
"The Data Warehouse and Data Mining"
Communications of the ACM, November 1996

En este breve artículo se hace hincapié en que la calidad de los datos recopilados (y de la forma en que se han almacenado) es esencial para obtener buenos resultados al aplicar técnicas de *Data Mining*.

James S. Ribeiro, Kenneth A. Kaufman & Larry Kerschberg
"Knowledge Discovery from Multiple Databases"
First International Conference on Knowledge Discovery (KDD-95), August 1995

En este artículo se propone en realizar un análisis individualizado de las relaciones existentes entre distintas tablas de una base de datos (mediante el uso de claves externas) frente a la alternativa clásica de la construcción de una relación universal a la hora de extraer conocimiento mediante algoritmos del tipo de ID3 (construcción de árboles de decisión) o AQ (generación progresiva de reglas). Las técnicas descritas se han incorporado a un prototipo denominado INLEN [*INference and LEarNing*].

Peggy Wright
"Knowledge Discovery in Databases"
CROSSROADS The ACM Student Magazine, Winter 1998

Un artículo de carácter informativo bastante bueno en el que se citan distintas técnicas usadas en Data Mining y KDD cuya complejidad reside esencialmente en los algoritmos de aprendizaje empleados. Aunque parezca increíble, no se mencionan explícitamente las reglas de asociación.