

Introducción a la Prospección de Datos Masivos ("Data Mining")

José Hernández Orallo
jorallo@dsic.upv.es

Transparencias y otra documentación en:
<http://www.dsic.upv.es/~jorallo/master/>

Máster de Ingeniería del Software. DSIC

Objetivos

- Conocer las características especiales de la extracción automática de conocimiento de bases de datos.
- Entender el proceso de extracción de conocimiento, sus fases y sus aplicaciones.
- Conocer las técnicas más apropiadas y su adaptación a estos problemas, especialmente clasificación y agrupamiento.
- Comparar y evaluar productos comerciales.
- Saber utilizar un paquete de minería de datos para resolver problemas sencillos de extracción de conocimiento.

2

Temario

- | | |
|---------------------------------|---|
| 1. Introducción | 1.1. Motivación
1.2. Problemas tipo y aplicaciones
1.3. Relación de DM con otras disciplinas |
| 2. El proceso de KDD | 2.1. Las Fases del KDD
2.2. Tipología de Patrones de Minería de Datos
2.3. Ejemplo |
| 3. Técnicas de Minería de Datos | 3.1. Taxonomía de Técnicas.
3.2. Evaluación de Hipótesis
3.3. Técnicas no supervisadas y descriptivas.
3.4. Técnicas supervisadas y predictivas. |
| 4. Desarrollo e Implantación | 4.1. Sistemas Comerciales
4.2. Tendencias
4.3. Para saber más |

3

1. Introducción

1.1. Motivación

1.2. Problemas tipo y aplicaciones

1.3. Relación de DM con otras disciplinas

4

Motivación

- El **aumento del volumen y variedad de información** que se encuentra informatizada en bases de datos digitales ha crecido espectacularmente en la última década.
- Gran parte de esta **información es histórica**, es decir, representa transacciones o situaciones que se han producido.
- Aparte de su función de "memoria de la organización", la información histórica es **útil para predecir la información futura**.

5

Motivación

- La mayoría de **decisiones** de empresas, organizaciones e instituciones se basan también en información de experiencias pasadas extraídas de fuentes muy diversas.
- las **decisiones colectivas** suelen tener consecuencias mucho más graves, especialmente económicas, y, recientemente, se deben basar en **volúmenes de datos que desbordan la capacidad humana**.

El área de la extracción (semi-)automática de conocimiento de bases de datos ha adquirido recientemente una importancia científica y económica inusual

6

Motivación

- Tamaño de datos poco habitual para algoritmos clásicos:
 - número de registros (ejemplos) muy largo (10^8 - 10^{12} bytes).
 - datos altamente dimensionales (n° de columnas/atributos): 10^2 - 10^4 .
- El usuario final no es un experto en aprendizaje automático ni en estadística.
- El usuario no puede perder más tiempo analizando los datos:
 - industria: ventajas competitivas, decisiones más efectivas.
 - ciencia: datos nunca analizados, bancos no cruzados, etc.
 - personal: "information overload"...

Los sistemas clásicos de estadística son difíciles de usar y no escalan al número de datos típicos en bases de datos.

7

Relación de DM con otras disciplinas

Aparece...

- "Descubrimiento de Conocimiento a partir de Bases de Datos" (KDD, del inglés *Knowledge Discovery from Databases*).
"proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos". (Fayyad et al. 1996)
- No se trata sólo de métodos estadísticos: la estadística se utiliza principalmente para validar o parametrizar un *modelo sugerido y preexistente*, no para generarlo.
- Diferencia sutil "Análisis Inteligente de Datos" (IDA, del inglés *Intelligent Data Analysis*) que correspondía con el uso de técnicas de inteligencia artificial en el análisis de los datos.

8

Relación de DM con otras disciplinas

KDD nace como interfaz entre y se nutre de diferentes disciplinas:

- estadística.
- sistemas de información / bases de datos.
- aprendizaje automático / IA.
- visualización de datos.
- computación paralela / distribuida.
- interfaces de lenguaje natural a bases de datos.

9

Áreas de Aplicación

Áreas de Aplicación:

Más importante industrialmente

- Toma de Decisiones (banca-finanzas-seguros, marketing, políticas sanitarias/demográficas, ...)
- Investigación Científica (medicina, astronomía, meteorología, psicología, ...). *Aquí la eficiencia no es tan importante.*
- Soporte al Diseño y Gestión de Bases de Datos.
 - *Reverse Engineering* (dados una base de datos, desnormalizarla para que luego el sistema la normalice).
 - Mejora de Calidad de Datos.
 - Mejora de Consultas (si se descubren dependencias funcionales nuevas u otras condiciones evitables).

10

Áreas de Aplicación. Problemas Tipo.

KDD para toma de decisiones (Dilly 96)

- | | |
|--------------------------|---|
| Comercio/Marketing: | - Identificar patrones de compra de los clientes.
- Buscar asociaciones entre clientes y características demográficas.
- Predecir respuesta a campañas de <i>mailing</i> .
- Análisis de cestas de la compra. |
| Banca: | - Detectar patrones de uso fraudulento de tarjetas de crédito.
- Identificar clientes leales.
- Predecir clientes con probabilidad de cambiar su afiliación.
- Determinar gasto en tarjeta de crédito por grupos.
- Encontrar correlaciones entre indicadores financieros. |
| Seguros y Salud Privada: | - Identificar reglas de mercado de valores a partir de históricos.
- Análisis de procedimientos médicos solicitados conjuntamente.
- Predecir qué clientes compran nuevas pólizas.
- Identificar patrones de comportamiento para clientes con riesgo.
- Identificar comportamiento fraudulento. |
| Transportes: | - Determinar la planificación de la distribución entre tiendas.
- Analizar patrones de carga. |

11

Áreas de Aplicación. Problemas Tipo.

KDD para toma de decisión

Medicina:

- Identificación de terapias médicas satisfactorias para diferentes enfermedades.
- Asociación de síntomas y clasificación diferencial de patologías.
- Estudio de factores (genéticos, precedentes, hábitos, alimenticios, etc.) de riesgo/salud en distintas patologías.
- Segmentación de pacientes para una atención más inteligente según su grupo.
- Predicciones temporales de los centros asistenciales para el mejor uso de recursos, consultas, salas y habitaciones.
- Estudios epidemiológicos, análisis de rendimientos de campañas de información, prevención, sustitución de fármacos, etc.

12

2. El proceso de KDD

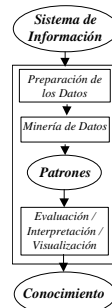
2.1. Las Fases del KDD

2.2. Tipología de Patrones de Minería de Datos

2.3. Ejemplo

13

El Proceso del KDD. FASES



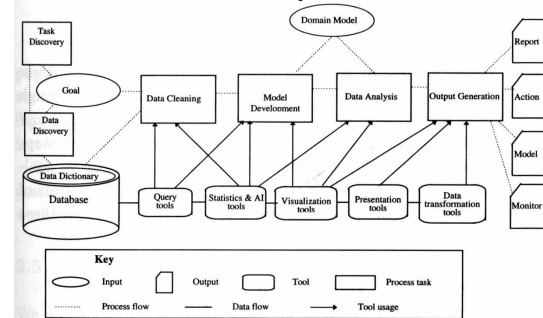
KDD

1. Determinar las fuentes de información que pueden ser útiles y dónde conseguirlas.
2. Diseñar el esquema de un almacén de datos (Data Warehouse) que consiga unificar de manera operativa toda la información recogida.
3. Implantación del almacén de datos que permita la "navegación" y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados.
4. 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).
5. Seleccionar y aplicar el método de minería de datos apropiado.
6. Interpretación, transformación y representación de los patrones extraídos.
7. Difusión y uso del nuevo conocimiento.

14

Fases y Técnicas del KDD

Las distintas técnicas de distintas disciplinas se utilizan en distintas fases:



Fases del KDD: Recogida de Datos

Las primeras fases del KDD determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original.

Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra:

- en bases de datos y otras **fuentes muy diversas**,
- tanto internas como **externas**.
- muchas de estas fuentes son las que se utilizan para el trabajo **transaccional**.

El análisis posterior será mucho más sencillo si la fuente es **unificada, accesible** (interna) y desconectada del trabajo **transaccional**.

16

Fases del KDD: Recogida de Datos

El proceso subsiguiente de minería de datos:

- Depende mucho de la fuente:
 - OLAP u OLTP.
 - Datawarehouse o copia con el esquema original.
 - ROLAP o MOLAP.
- Depende también del tipo de usuario:
 - 'picapedreros' (o 'granjeros'): se dedican fundamentalmente a realizar informes periódicos, ver la evolución de determinados parámetros, controlar valores anómalos, etc.
 - 'exploradores': encargados de encontrar nuevos patrones significativos utilizando técnicas de minería de datos.

17

Fases del KDD: Selección, Limpieza y Transformación de Datos

(Pyle 1999) es un libro dedicado al problema de la preparación de datos

Limpieza (data cleansing) y criba (selección) de datos:

Se deben eliminar el mayor número posible de datos erróneos o inconsistentes (limpieza) e irrelevantes (criba).

Métodos estadísticos casi exclusivamente.

- histogramas (detección de datos anómalos).
- selección de datos (muestreo, ya sea verticalmente, eliminando atributos, u horizontalmente, eliminando tuplas).
- redefinición de atributos (agrupación o separación).

18

Fases del KDD: Selección, Limpieza y Transformación de Datos

Transformación del Esquema:

- Esquema Original:
 - Ventajas: Las R.I. se mantienen (no hay que reaprenderlas, no despidan)
 - Inconvenientes: Muchas técnicas no se pueden utilizar.
- Tabla Universal: *Cualquier Esquema Relacional se puede convertir (en una correspondencia 1 a 1) a una tabla universal.*
 - Ventajas: Modelos de aprendizaje más simples (proposicionales).
 - Desventajas: Muchísima Redundancia (tamaños ingentes).
- Desnormalizado Tipo Estrella o Copo de Nieve (*datamarts*):
 - Ventajas: Se pueden buscar reglas sobre información sumariada y si resultan factibles se pueden comprobar con la información detallada. Con operadores propios: *Roll-up, Drill-down, Slicing and Dicing.*
 - Desventajas: Orientadas a extraer un tipo de información (granjeros).¹⁹

Fases del KDD: Selección, Limpieza y Transformación de Datos

Intercambio de Dimensiones: (filas por columnas)

EJEMPLO:

Una tabla de cestas de la compra, donde cada atributo indica si el producto se ha comprado o no.

- Objetivo: Ver si dos productos se compran conjuntamente (regla de asociación).

Es muy costoso: hay que mirar al menos la raíz cuadrada de todas las filas (cestas).

Y puede haber millones en una semana...

Sin embargo...

Productos sólo hay unos 10.000.

20

Fases del KDD: Selección, Limpieza y Transformación de Datos

Intercambio de Dimensiones: EJEMPLO

Si se intercambian filas por columnas tenemos:

	B1	B2	B3	B4	B5	B6	...
Jabón	X		X				
Huevos		X			X		
Patatas Fritas		X			X		
Champú						X	
Jabón + Champú	X		X			X	
Huevos + Pat. Fritas							

Sólo es necesario hacer XOR entre dos filas para saber si hay asociación.

21

Fases del KDD: Selección, Limpieza y Transformación de Datos

Transformación de los Campos:

- Numerización / Etiquetado
 - Ventajas: Se reduce espacio. Ej: apellido \Rightarrow entero. Se pueden utilizar técnicas más simples.
 - Desventajas: Se necesita meta-información para distinguir los datos inicialmente no numéricos (la cantidad no es relevante) de los inicialmente numéricos (la cantidad es relevante: precios, unidades, etc.). A veces se puede "sesgar" el modelo (*biasing*).
- Discretización:
 - Ventajas: Se reduce espacio. Ej. 0..10 \Rightarrow (pequeño, mediano, grande). Se pueden utilizar árboles de decisión y construir reglas discretas.
 - Desventajas: Una mala discretización puede invalidar los resultados.

22

Fases del KDD: Selección, Limpieza y Transformación de Datos

Pick & Mix:

- En minería de datos, generalmente los datos sugieren la creación de nuevos campos (columnas) por pick & mix.
- Ejemplos:
 - $\text{height}^2 / \text{weight}$ (índice de obesidad)
 - $\text{debt} / \text{earnings}$
 - $\text{passengers} * \text{miles}$
 - $\text{credit limit} - \text{balance}$
 - $\text{population} / \text{area}$
 - $\text{minutes of use} / \text{number of telephone calls}$
 - $\text{activation_date} - \text{application_date}$
 - $\text{number of web pages visited} / \text{total amount purchased}$
- Es conveniente añadirlas, pero siempre una sola combinación. No poner x/y si ya se ha puesto $x-y$.

23

Fases del KDD: La Minería de Datos

Características Especiales de los Datos:

Aparte del gran volumen, ¿por qué las técnicas de aprendizaje automático y estadística no son *directamente* aplicables?

- Los datos residen en el disco. No se pueden escanear múltiples veces.
- Algunas técnicas de muestreo no son compatibles con algoritmos no incrementales.
- Muy alta dimensionalidad (muchos campos).
- Evidencia POSITIVA.
- DATOS IMPERFECTOS...

24

Fases del KDD: La Minería de Datos

Patrones a descubrir:

- Una vez recogidos los datos de interés, un explorador puede decidir qué tipos de patrón quiere descubrir.
- El tipo de conocimiento que se desea extraer va a marcar claramente la *técnica* de minería de datos a utilizar.
- Según como sea la búsqueda del conocimiento se puede distinguir entre:
 - *Directed data mining*: se sabe claramente lo que se busca, generalmente predecir unos ciertos datos o clases.
 - *Undirected data mining*: no se sabe lo que se busca, se trabaja con los datos (*¡hasta que confiesen!*).
- En el primer caso, se trata de elegir el *algoritmo* más idóneo entre los disponibles para un determinado tipo de patrón a buscar.

25

Fases del KDD: Visualización

Las técnicas de visualización de datos se utilizan fundamentalmente con dos objetivos:

- aprovechar la gran capacidad humana de extraer patrones a partir de imágenes.
- ayudar al usuario a comprender más rápidamente patrones descubiertos automáticamente por un sistema de KDD.

26

Fases del KDD: Visualización

Estos dos objetivos marcan dos momentos diferentes del uso de la visualización de los datos (no excluyentes) [Fayyad et al. 2002]:

- *visualización previa* (tb. Visual Data Mining [Wong 1999]): se utiliza para entender mejor los datos y sugerir posibles patrones o qué tipo de herramienta de KDD utilizar.
- *visualización posterior* al proceso de minería de datos (tb. Model Visualization) : se utiliza para mostrar los patrones y entenderlos mejor.

27

Tipología de Patrones de Minería de Datos

Tipos de conocimiento:

- **Asociaciones**: Una asociación entre dos atributos ocurre cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente es relativamente alta.
 - Ejemplo, en un supermercado se analiza si los pañales y los potitos de bebé se compran conjuntamente.
- **Dependencias**: Una dependencia funcional (aproximada o absoluta) es un patrón en el que se establece que uno o más atributos determinan el valor de otro. Ojo! Existen muchas dependencias nada interesantes (causalidades inversas).
 - Ejemplo: que un paciente haya sido ingresado en maternidad determina su sexo.

28

Tipología de Patrones de Minería de Datos

Tipos de conocimiento (cont.):

- **Clasificación**: Una clasificación se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas.
 - Ejemplo: se sabe (por un estudio de dependencias) que los atributos edad, grado de miopías y astigmatismo han determinado los pacientes para los que su operación de cirugía ocular ha sido satisfactoria.
 - Podemos intentar determinar las reglas exactas que clasifican un caso como positivo o negativo a partir de esos atributos.
- **Agrupamiento / Segmentación**: El agrupamiento (o clustering) es la detección de grupos de individuos. Se diferencia de la clasificación en el que no se conocen ni las clases ni su número (aprendizaje no supervisado), con lo que el objetivo es determinar grupos o racimos (clusters) diferenciados del resto.

29

Tipología de Patrones de Minería de Datos

Tipos de conocimiento (cont.):

- **Tendencias/Regresión**: El objetivo es predecir los valores de una variable continua a partir de la evolución sobre otra variable continua, generalmente el tiempo, o sobre un conjunto de variables.
 - Ejemplo, se intenta predecir el número de clientes o pacientes, los ingresos, llamadas, ganancias, costes, etc. a partir de los resultados de semanas, meses o años anteriores.
- **Información del Esquema**: (descubrir claves primarias alternativas, R.I.).
- **Reglas Generales**: patrones no se ajustan a los tipos anteriores. Recientemente los sistemas incorporan capacidad para establecer otros patrones más generales.

30

Ejemplo

Tenemos una tabla con datos de empleados:

#	Sueldo	Casado	Coche	Hijos	Alg/Prop	Sindic.	Bajas/Año	Antgdd.	Sexo
1	10000	Si	No	0	Alquiler	No	7	15	H
2	20000	No	Si	1	Alquiler	Si	3	3	M
3	15000	Si	Si	2	Prop	Si	5	10	H
4	30000	Si	Si	1	Alquiler	No	15	7	M
5	10000	Si	Si	0	Prop	Si	1	6	H
6	40000	No	Si	0	Alquiler	Si	3	16	M
7	25000	No	No	0	Alquiler	Si	0	8	H
8	20000	No	Si	0	Prop	Si	2	6	M
9	20000	Si	Si	3	Prop	No	7	5	H
10	30000	Si	Si	2	Prop	No	1	20	H
11	50000	No	No	0	Alquiler	No	2	12	M
12	8000	Si	Si	2	Prop	No	3	1	H
13	20000	No	No	0	Alquiler	No	27	5	M
14	10000	No	Si	0	Alquiler	Si	0	7	H
15	8000	No	Si	0	Alquiler	No	3	2	H

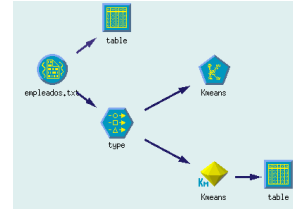
Queremos obtener subgrupos representativos....

31

Ejemplo

- Importamos los datos en un paquete de minería de datos, tipamos los datos, miramos si hay datos anómalos, etc.
- Aplicamos el método k-means para buscar clusters (grupos). Le decimos que nos busque los 3 grupos más significativos.

El proceso de minería de datos resultante es el siguiente:



32

Ejemplo

Podemos ver qué características tiene cada cluster

cluster 1	cluster 2	cluster 3
5 examples	4 examples	6 examples
Sueldo : 226000 Casado : No -> 0.8 Si -> 0.2 Coche : No -> 0.8 Si -> 0.2 Hijos : 0 Alg/Prop : Alquiler -> 1.0 Sindic. : No -> 0.8 Si -> 0.2 Bajas/Año : 8 Antigüedad : 8 Sexo : H -> 0.6	Sueldo : 225000 Casado : No -> 1.0 Coche : Si -> 1.0 Hijos : 0 Alg/Prop : Alquiler -> 0.75 Prop -> 0.25 Sindic. : Si -> 1.0 Bajas/Año : 2 Antigüedad : 8 Sexo : H -> 0.25 M -> 0.75	Sueldo : 188333 Casado : Si -> 1.0 Coche : Si -> 1.0 Hijos : 2 Alg/Prop : Alquiler -> 0.17 Prop -> 0.83 Sindic. : No -> 0.67 Si -> 0.33 Bajas/Año : 5 Antigüedad : 8 Sexo : H -> 0.83 M -> 0.17

¿Cómo interpretamos estos resultados?

33

Ejemplo

También podemos ver qué empleado ha sido asignado a qué cluster:

#	Sueldo	Casado	Coche	Hijos	Alg/Prop	Sindic.	Bajas/Año	Antigüedad	Sexo	Dist. Basano	Dist. Basano
1	10000	Si	No	0	Alquiler	No	7	15	H	cluster-1	1.045
2	20000	No	Si	1	Alquiler	Si	3	3	M	cluster-2	0.512
3	15000	Si	Si	2	Prop	Si	5	10	H	cluster-3	0.728
4	30000	Si	Si	1	Alquiler	No	15	7	M	cluster-3	1.304
5	10000	Si	Si	0	Prop	Si	1	6	H	cluster-3	0.544
6	40000	No	Si	0	Alquiler	Si	3	16	M	cluster-2	0.696
7	25000	No	No	0	Alquiler	Si	0	8	H	cluster-1	0.583
8	20000	No	Si	0	Prop	Si	2	6	M	cluster-2	0.804
9	20000	Si	Si	3	Prop	No	7	5	H	cluster-3	0.63
10	30000	Si	Si	2	Prop	No	1	20	H	cluster-3	0.814
11	50000	No	No	0	Alquiler	No	2	12	M	cluster-1	0.594
12	8000	Si	Si	2	Prop	No	3	1	H	cluster-3	0.629
13	20000	No	No	0	Alquiler	No	27	5	M	cluster-1	1.011
14	10000	No	Si	0	Alquiler	Si	0	7	H	cluster-2	0.454
15	8000	No	Si	0	Alquiler	No	3	2	H	cluster-1	1.07

Y asignar nuevos empleados a los clusters definidos.

34