

Minería de Datos

5. Otros Aspectos

José Hernández Orallo

jorallo@dsic.upv.es

Máster y Cursos de Postgrado del DSIC
Universitat Politècnica de València

Objetivos Tema 5

- Conocer la metodología CRISP-DM
- Conocer las tendencias del área de minería de datos, en particular los lenguajes de consulta inductivos.
- Tener noción de cuestiones legales que pueden afectar al proceso de minería de datos.
- Presentar referencias de recursos en DM.

3

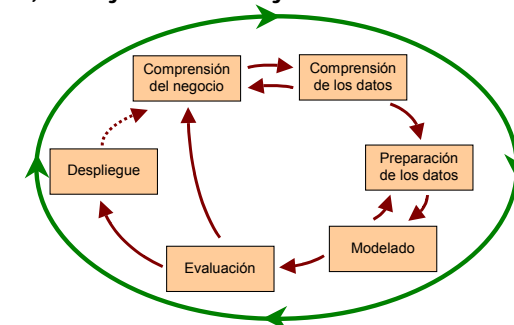
Temario

1. Introducción a la Minería de Datos (DM)
 - 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 Técnicas de Minería de Datos
 - 2.3. Sistemas Comerciales y Herramientas de Minería de Datos
 - 2.4. Preparación y Visualización de Datos
3. Técnicas de Minería de Datos
 - 3.1. El Problema de la Extracción Automática de Conocimiento.
 - 3.2. Evaluación de Hipótesis
 - 3.3. Técnicas no supervisadas y descriptivas.
 - 3.4. Técnicas supervisadas y predictivas.
4. Web Mining
 - 4.1. Los Problemas de la Información No Estructurada.
 - 4.2. Extracción de Con. a partir de Documentos HTML y texto.
 - 4.3. Extracción de Información semi-estructurada (XML).
5. Otros Aspectos

2

Metodología CRISP-DM

- **CRISP-DM** (www.crisp-dm.org) (*C*Ross-*I*ndustry *S*tandard *P*rocess for *D*ata *M*ining)
 - es un consorcio de empresas (inicialmente bajo una subvención inicial de la Comisión Europea), incluyendo SPSS, NCR y DaimlerChrysler.



4

Metodología CRISP-DM

- **Comprensión del negocio:**

- entender los objetivos y requerimientos del proyecto desde una perspectiva de negocio.

Subfases:

- **establecimiento de los objetivos de negocio** (contexto inicial, objetivos y criterios de éxito),
- **evaluación de la situación** (inventario de recursos, requerimientos, suposiciones y restricciones, riesgos y contingencias, terminología y costes y beneficios),
- **establecimiento de los objetivos de minería de datos** (objetivos de minería de datos y criterios de éxito) y
- **generación del plan del proyecto** (plan del proyecto y evaluación inicial de herramientas y técnicas).

5

Metodología CRISP-DM

- **Comprensión de los datos:**

- **recopilar y familiarizarse con los datos**, identificar los problemas de calidad de datos y ver las primeras potencialidades o subconjuntos de datos que puede ser interesante analizar (según los objetivos de negocio en la fase anterior). Subfases:

- **recopilación inicial de datos** (informe de recopilación),
- **descripción de datos** (informe de descripción),
- **exploración de datos** (informe de exploración) y
- **verificación de calidad de datos** (información de calidad).

6

Metodología CRISP-DM

- **Preparación de los datos:**

- el objetivo de esta fase es obtener la “vista minable”. Aquí se incluye la integración, selección, limpieza y transformación. Subfases:

- **selección de datos** (razones de inclusión / exclusión),
- **limpieza de datos** (informe de limpieza de datos),
- **construcción de datos** (atributos derivados, registros generados),
- **integración de datos** (datos mezclados) y
- **formateo de datos** (datos reformateados).

7

Metodología CRISP-DM

- **Modelado:**

- es la aplicación de técnicas de modelado o de minería de datos propiamente dichas a las vistas minables anteriores. Subfases:

- **selección de la técnica de modelado** (técnica de modelado, suposiciones de modelado),
- **diseño de la evaluación** (diseño del test),
- **construcción del modelo** (parámetros elegidos, modelos, descripción de los modelos) y
- **evaluación del modelo** (medidas del modelo, revisión de los parámetros elegidos).

8

Metodología CRISP-DM

• Evaluación:

- es necesario evaluar (desde el punto de vista de la finalidad) los modelos de la fase anterior. Es decir, si el modelo nos sirve para responder a algunos de los requerimientos del negocio.

Subfases:

- **evaluación de resultados** (evaluación de los resultados de minería de datos, modelos aprobados),
- **revisar el proceso** (revisión del proceso) y
- **establecimiento de los siguientes pasos** (lista de posibles acciones, decisión).

9

Metodología CRISP-DM

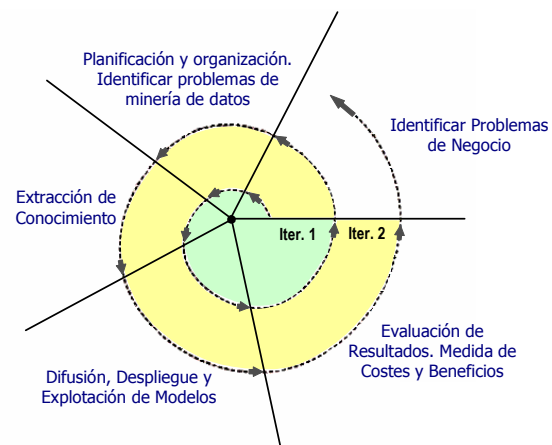
• Despliegue:

- se trata de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisión de la organización, difundir informes sobre el conocimiento extraído, etc. Subfases:
 - **planificación del despliegue** (plan del despliegue),
 - **planificación de la monitorización y del mantenimiento** (plan de la monitorización y del despliegue),
 - **generación del informe final** (informe final, presentación final) y
 - **revisión del proyecto** (documentación de la experiencia).

10

Metodología CRISP-DM

• Implantación progresiva en una organización:



11

Lenguajes de Consulta Inductivos

Nuevos Lenguajes de Consulta:

- El descubrimiento en bases de datos se ve como un proceso de consulta a una base de datos. La situación se parece al desarrollo de lenguajes de consulta en los sesenta y setenta.
- Una consulta inductiva o de búsqueda de patrones debe permitir al usuario restringir la búsqueda inductiva en los siguientes aspectos (Han et al. 1999):
 - La parte de la base de datos a ser minada (también llamada la vista minable o vista relevante).
 - El tipo de patrón/reglas a ser minado (también llamado restricciones del conocimiento).
 - Cuantificadores estadísticos: representatividad (support) $\exists\%$, precisión (confidence/accuracy) $\forall\%$.
 - Otras propiedades que el patrón debería cumplir (número y forma de las reglas, interés, novedad, etc.).

12

Lenguajes de Consulta Inductivos

Propuesta M-SQL (Imielinski et al. 1996)

Basada en modelos de consulta...

Ejemplo:

```
SELECT FROM MINE(T): R
WHERE R.Consequent = { (Age = *) }
      R.Support > 1000
      R.Confidence > 0.65;
```

R es una variable de regla y se puede utilizar:

- R.Consequent
- R.Body (antecedente)
- R.Support
- R.Confidence.

13

Lenguajes de Consulta Inductivos

Propuesta DMQ (Data-Mining Query) language (Ng et al. 1998):

- Utiliza la sintaxis del SQL para la vista minable
- También basado en modelos de consulta.

EJEMPLO:

Esquema:

```
SALES(customer_name, item_name, transaction_id)
LIVES(customer_name, district, city)
ITEM(item_name, category, price)
TRANSACTION(transaction_id, day, month, day)
```

Consulta Inductiva (lenguaje natural):

“buscar las ventas de qué artículos baratos (con una suma de precios menor que \$100) que puede motivar las ventas de qué artículos caros (con el precio mínimo de \$500) de la misma categoría de los clientes de Vancouver en 1998”.

14

Lenguajes de Consulta Inductivos

Propuesta DMQ. EJEMPLO:

Ejemplo de Consulta Inductiva:

```
mine associations as
  lives(C, “Vancouver”) and
  sales+(C, ?[I], {S}) => sales+(C, ?[J], {T})
from sales
where S.year = 1998 and T.year = 1998 and I.category = J.category
group by C, I.category
having sum(I.price) < 100 and min(J.price) >= 500
with min_support = 0.01 and min_confidence = 0.5
```

Ejemplo de Respuesta:

```
lives(C, “Vancouver”) and
sales(C, “Census_CD”, _) and sales(C, “Ms/Office97”, _)
=> sales(C, “Ms/SQLServer”, _) [0.015, 0.68]
```

+: operador regular (1 o más tuplas)
?[I] : utilizar clave ajena. I es la tupla instanciada.

Es un patrón relacional.

Support & Confidence.

15

Lenguajes de Consulta Inductivos

Propuesta “OLE DB for Data Mining” de Microsoft.

- Extensión del protocolo de acceso a BB.DD. OLE DB.
- Implementa una extensión del SQL que trabaja con DMM(Data Mining Model) y permite:

1. Crear el modelo
2. Entrenar el modelo
3. Realizar predicciones

16

Lenguajes de Consulta Inductivos

Propuesta “OLE DB for Data Mining”:

Ejemplo: CREACIÓN DEL MODELO (DMM):

```
CREATE MINING MODEL CredikRisk
```

```
(  
  [Customer ID] LONG KEY,  
  [Profession] TEXT DISCRETE,  
  [Income] TEXT DISCRETE,  
  [Age] LONG CONTINUOUS,  
  [Risk Level] TEXT DISCRETE PREDICT,  
)  
USING [Microsoft Decision Tree]
```

Esto crea un DMM vacío.

17

Lenguajes de Consulta Inductivos

Propuesta “OLE DB for Data Mining”:

Ejemplo: ENTRENAR EL MODELO:

Se usa una sentencia INSERT INTO. A diferencia de insertar datos como en una tabla normal lo que hace es analizar los casos que le introduzcamos y construir el contenido del DMM.

```
INSERT INTO [CreditRisk]  
( [CustomerID],[Profession],[Income],[Age],[RiskLevel] )  
OPENROWSET ('[Provider]='MSOLESQL','user','pwd',  
  'SELECT [CustomerID],[Profession],  
  [Income],[Age],[Risk]  
  FROM [Customers]')
```

18

Lenguajes de Consulta Inductivos

Propuesta “OLE DB for Data Mining”:

Ejemplo: USAR EL MODELO:

- *El modelo se aplica a nuevos datos. La manera de hacerlo es similar a la concatenación de dos tablas relacionales, considerando el modelo como una tabla y los datos a predecir como otra tabla. El resultado es una nueva tabla con los datos que queramos (todos o sólo las predicciones).*

```
SELECT [CustomerID],[Income],[Age], CreditRisk.RiskLevel,  
  PredictProbability(CreditRisk.RiskLevel)  
FROM CreditRisk PREDICTION JOIN Customers  
  ON CreditRisk.Profession=Customers.Profession  
  AND CreditRisk.Income=Customers.Income  
  AND CreditRisk.Age=Customers.Age
```

19

Lenguajes de consulta inductivos para Web Usage Mining

- También existen lenguajes de consulta para seleccionar patrones relativos a uso de páginas web:

- P.ej. En el sistema WUM (Web Utilization Miner) (Berendt & Spiliopoulou 2000), basado también en un grafo de secuencias de visitas, se puede utilizar el lenguaje MINT para hacer consultas del estilo:

```
SELECT t  
FROM NODE AS a b,  
TEMPLATE a * b AS t  
WHERE a.support > 7  
AND (b.support / a.support) >= 0.4  
AND b.url != "G.html"
```

- *Seleccionaría pares de páginas visitadas consecutivamente en la que la primera se ha visitado al menos 7 veces y de éstas, al menos el 40% han llegado a la segunda. Además la segunda no puede ser "G.html".*

Algunas Cuestiones Legales

- Hay dos cuestiones importantes respecto a un uso indiscriminado de KDD:
 - El primero es si los clientes o otros usuarios externos en general se pueden ver incomodados o amenazados por la compañía al atacar su privacidad o someterlos a márketing abusivo.
 - El segundo es si estas políticas pueden ser ilegales.
- Consecuencias:
 - En el primer caso, la compañía o institución obtienen mala prensa y antipatía (lo cual se puede traducir en una pérdida económica).
 - En el segundo caso, la compañía puede ser demandada por miles de clientes, con unos costes de millones de euros.

21

Algunas Cuestiones Legales

Si nos centramos sólo en las cuestiones legales del KDD:

- Uso de datos de fuentes internas a la compañía:
 - se pueden utilizar los datos como se quiera, siempre internamente.
- Uso de datos de fuentes externas a la compañía:
 - Sí se puede si los datos son públicos (la persona ha decidido que lo sean). Ejemplo: páginas web, guía telefónica, visitas en la web.
 - Sí se puede si los datos son agregados (i.e. no contienen individuos). Ejemplo: asociaciones entre productos, llamadas por distrito/hora, segmentaciones, etc ...
 - NO se puede si la persona los ha cedido por transacción u operación habitual y privada con la otra compañía. Ejemplo: datos bancarios, horarios y números de llamadas, cestas de la compra, historiales clínicos, viajes, etc.

22

Algunas Cuestiones Legales

Cuestiones legales del KDD. KDD y Discriminación:

*Una parte importante de los objetivos del KDD es **discriminar** poblaciones (especialmente clientes).*

- No existe una línea clara entre discriminación legal/ilegal...
- ¿Enviar una campaña/oferta de viajes sólo a no-jubilados es legal?
 - ¿No asegurar ciclomotores sólo a varones?
 - ¿Enviar una campaña/oferta de libros científicos sólo a mujeres es legal?
 - ¿Enviar una campaña/oferta de bronceadores sólo a clientes de piel blanca (determinado por análisis de las fotos) es legal?
 - ¿Enviar una campaña/oferta de biblias a cristianos (determinado por los minutos de visión del Christian Channel en un paquete digital de pago)?

23

Algunas Cuestiones Legales

Transparencia en decisiones:

Algunas decisiones (en banca, compañías de seguros, etc.) se toman sin unas reglas públicas e iguales para todos

- La Comisión Europea ha obligado a que dichas reglas sean “transparentes”.
 - Un modelo de minería de datos sobre el que se base la decisión deberá ser comprensible:
 - Redes neuronales → reglas

24

Recursos Web

- **Minería de Datos:**
 - Knowledge Discovery Mine (<http://www.kdnuggets.com>)
 - The Data Mine (<http://www.the-data-mine.com>)
 - Thearling (<http://www.thearling.com/>)
- **Business Intelligence:**
 - BI-SPAIN (<http://www.bi-spain.com>). Documentos libres (hace falta registrarse).
 - *The Datawarehousing Institute* (<http://www.tdwi.org/>). La mayoría de documentos requieren ser miembro.
 - Datawarehouse (<http://www.datawarehouse.com> , <http://www.dmreview.com/>). Documentos libres mayoritariamente
 - OLAPreport (<http://www.olapreport.com>). Informes.

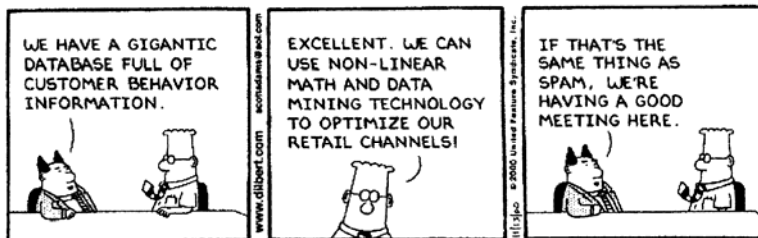
25

Entidades y Consorcios

- **España:**
 - RED ESPAÑOLA DE MINERÍA DE DATOS (<http://www.lsi.us.es/redmidas/>)
- **Internacional:**
 - CRISP - DM, un consorcio industrial (<http://www.crisp-dm.org>)
 - DMG - The Data Mining Group (<http://www.dmg.org/>), un consorcio mixto para crear estándares para intercambiar modelos.

26

...



DILBERT reprinted by permission of United Feature Syndicate, Inc. [2000].

27