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

Temario

4. Desarrollo e Implantación de DM

Sistemas

Elder Research, www.dataminglab.com

Sistemas

Tipos de Sistemas:

- *Standalone*: Los datos se deben exportar/convertir al formato interno del sistema de data mining: Knowledge Seeker IV (Angoss International Limited, Groupe Bull).
- *On-top*: funcionan sobre un sistema propietario (microstrategy sobre Oracle).
- *Embedded* (propietarios): Oracle Discoverer, Oracle Darwin, IBM...
- Extensible (Tecnología *Plug-ins*): proporcionan unas herramientas mínimas de interfaz con los datos, estadísticas y visualización, y los algoritmos de aprendizaje se pueden ir añadiendo con plug-ins. (ej. KEPLER).

Sistemas

Producto	Compañía	Técnicas	Plataformas	Interfaz
Knowledge Seeker	Angoss http://www.angoss.com/	Decision Trees, Statistics	Win NT	ODBC
CART	Salford Systems www.salford-systems.com	Decision Trees	UNIX/NT	
Clementine	SPSS/Integral Solutions Limited (ISL) www.spss.com	Decision Trees, ANN, Statistics, Rule Induction, Association Rules, K Means, Linear Regression.	UNIX/NT	ODBC
Data Surveyor	Data Distilleries http://www.datadistilleries.com/	Amplio Abanico.	UNIX	ODBC
GainSmarts	Urban Science www.urbanscience.com	Especializado en gráficos de ganancias en campañas de clientes (sólo Decision Trees, Linear Statistics y Logistic Regression).	UNIX/NT	
Intelligent Miner	IBM http://www.ibm.com/software/data/iminer	Decision Trees, Association Rules, ANN, RBF, Time Series, K Means, Linear Regression.	UNIX (AIX)	IBM, DB2
Microstrategy	Microstrategy www.microstrategy.com	Datawarehouse sólo	Win NT	Oracle
Polyanalyst	Megaputer http://www.megaputer.com/html/polyanalyst4.0.html	Symbolic, Evolutionary	Win NT	Oracle, ODBC
Darwin	Oracle http://www.oracle.com/ip/analyze/warehouse/datamining/index.html	Amplio Abanico (Decision Trees, ANN, Nearest Neighbour)	UNIX/NT	Oracle
Enterprise Miner	SAS http://www.sas.com/software/components/eminer.html	Decision Trees, Association rules, ANN, regression, clustering.	UNIX (Sun), NT, Mac	Oracle, ODBC
SGI MineSet	Silicon Graphics http://www.sgi.com/software/mineset/	association rules and classification models, used for prediction, scoring, segmentation, and profiling	UNIX (Irix)	Oracle, Sybase, Informix.
Wizsoft/Wizwhy	http://www.wizsoft.com/			

Esta tabla se actualiza en: http://www.dsic.upv.es/~jorallo/master/

Sistemas

- Más software comercial DM:
  http://www.kdcentral.com/Software/Data_Mining/
  http://www.the-data-mine.com/bin/veiw/Software/WebIndex
- Algunos Prototipos No Comerciales o Gratuitos:
  - Kepler: sistema de plug-ins del GMD (http://ais.gmd.de/KD/kepler.html).
  - Rproject: herramienta gratuita de análisis estadístico (http://www.R-project.org/)
  - Librerías WEKA (http://www.cs.waikato.ac.nz/~ml/weka/) (Witten & Frank 1999)

Sistemas

EJEMPLO: **Clementine**

www.spss.com

- Herramienta que incluye:
  - fuentes de datos (ASCII, Oracle, Informix, Sybase e Ingres).
  - interfaz visual.
  - distintas herramientas de minería de datos: redes neuronales y reglas.
  - manipulación de datos (pick & mix, combinación y separación).

Sistemas

EJEMPLO: **Clementine**

Ejemplo Práctico: Ensayo de Medicamentos

http://www.pcc.qub.ac.uk/tec/courses/datamining/ohp/dm-OHP-final_3.html

- Un número de pacientes hospitalarios que sufren todos la misma enfermedad se tratan con un abanico de medicamentos.
- 5 medicamentos diferentes están disponibles y los pacientes han respondido de manera diferente a los diferentes medicamentos.
- Problema:

¿qué medicamento es apropiado para un nuevo paciente?

Sistemas

EJEMPLO: **Clementine**. Ejemplo Práctico: Ensayo de Medicamentos

Primer Paso: ACCEDIENDO LOS DATOS:

- Se leen los datos. Por ejemplo de un fichero de texto con delimitadores.
- Se nombran los campos:

age	edad
sex	sexo
BP	presión sanguínea (High, Normal, Low)
Cholesterol	colesterol (Normal, High)
Na	concentración de sodio en la sangre.
K	concentración de potasio en la sangre.
drug	medicamento al cual el paciente respondió satisfactoriamente.

SE PUEDEN COMBINAR LOS DATOS:
P.ej. se puede añadir un nuevo atributo: Na/K

Sistemas

EJEMPLO: **Clementine**

Segundo Paso: Familiarización con los Datos. Visualizamos los registros:

table1

Table

Age	Sex	BP	Cholestrol	Na	K	Drug	Na_to_K
23	F	HIGH	HIGH	0.79	0.03	drugY	25.35
47	M	LOW	HIGH	0.74	0.06	drugC	13.09
47	M	LOW	HIGH	0.7	0.07	drugC	10.11
28	F	NORMAL	HIGH	0.56	0.07	drugX	7.8
61	F	LOW	HIGH	0.56	0.03	drugY	18.04
22	F	NORMAL	HIGH	0.68	0.08	drugX	8.61
49	F	NORMAL	HIGH	0.79	0.05	drugY	16.28
41	M	LOW	HIGH	0.77	0.07	drugC	11.04
60	M	NORMAL	HIGH	0.78	0.05	drugY	15.17
43	M	LOW	NORMAL	0.53	0.03	drugY	19.37

Sistemas

EJEMPLO: **Clementine**

- Permite seleccionar campos o filtrar los datos
- Permite mostrar propiedades de los datos. Por ejemplo:
  ¿Qué proporción de casos respondió a cada medicamento?

Drug

Distribution

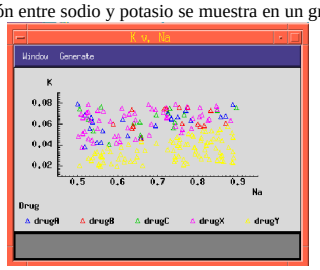
Value	Proportion	%	Occurences
drugA		11,5	23
drugB		8,0	16
drugC		8,0	16
drugX		27,0	54
drugY		45,5	91

Sistemas

EJEMPLO: **Clementine**

- Permite encontrar relaciones. Por ejemplo:

La relación entre sodio y potasio se muestra en un gráfico de puntos.



Se observa una dispersión aparentemente aleatoria (excepto para el medicamto. *Y*)

Sistemas

EJEMPLO: **Clementine**

Se puede observar a simple vista que los pacientes con alto cociente Na/K responden mejor al medicamento Y.

Pero queremos una clasificación para todos los medicamentos. Es decir, nuestro problema original:

¿Cuál es el mejor medicamento para cada paciente?

Tercer Paso: Construcción del Modelo

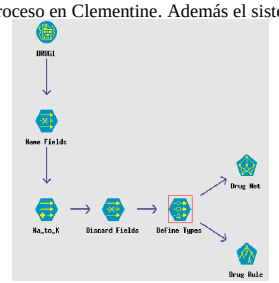
Tareas a realizar en Clementine:

- Filtrar los campos no deseados.
- Definir tipos para los campos.
- Construir modelos (reglas y redes)

Sistemas

EJEMPLO: **Clementine**

Se sigue este proceso en Clementine. Además el sistema lo visualiza:

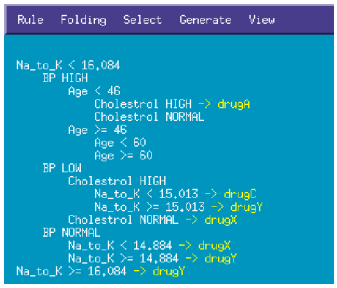


A partir de 2000 ejemplos entrena la red y construye las reglas.

Sistemas

EJEMPLO: **Clementine**

Permite examinar las reglas:



Las reglas extienden el mismo criterion que se había descubierto previamente: es decir, medicamento *Y* para los pacientes con alto cociente Na/K. Pero además añaden reglas para el resto.

Sistemas

EJEMPLO: SAS **ENTERPRISE MINER** (EM)

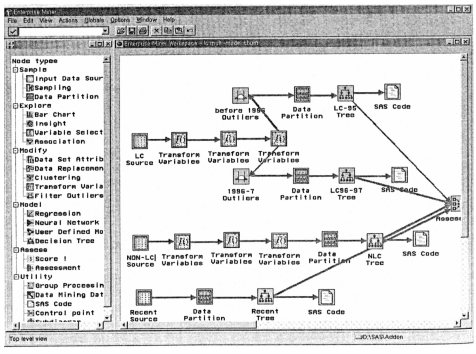
- Herramienta completa. Incluye:
  - conexión a bases de datos (a través de ODBC y SAS datasets).
  - muestreo e inclusión de variables derivadas.
  - partición de la evaluación del modelo respecto a conjuntos de entrenamiento, validación y chequeo.
  - distintas herramientas de minería de datos: varios algoritmos y tipos de árboles de decisión, redes neuronales, regresión y clustering.
  - comparación de modelos.
  - conversión de los modelos en código SAS.
  - interfaz gráfico.
- Incluye herramientas para flujo de proceso: trata en el proceso KDD como un proceso y las fases se pueden repetir, modificar y grabar.

Sistemas

EJEMPLO:

SAS ENTERPRISE MINER (EM)

(flujo del proceso KDD)

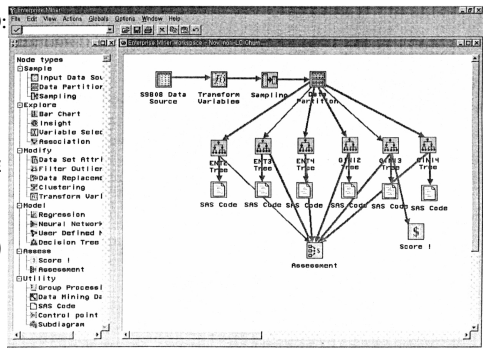


Sistemas

EJEMPLO:

SAS ENTERPRISE MINER (EM)

Selección (assessment) de modelos



Sistemas

Oracle: Herramientas "Business Intelligence" y "Data Mining"
http://www.oracle.com/ip/analyze/warehouse/bus_intell/index.html

Tienen una orientación más empresarial y de sistemas de información. Herramientas de OLAP, Datawarehouse e Informes Avanzados:

- Oracle Express Server.
- Sales Analyzer and Financial Analyzer.
- Oracle Express Objects and Oracle Express Analyzer.
- Oracle Discoverer and Oracle Reports.

Herramientas propias de Minería de Datos:
- Oracle **Darwin**. (incluido ya en Oracle9i)
  http://www.oracle.com/ip/analyze/warehouse/datamining/index.html

20

Sistemas

MS SQL SERVER: Analysis Services

- OLAP Services de SQL Server 97 se amplió a partir de SQL Server 2000 con características de DM en el llamado "Analysis Services".

- Se fundamenta en el "OLE DB for Data Mining": 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

21

Tendencias

- 80s y principios 90s:
  - OLAP: consultas predefinidas. El sistema OLAP como sistema para extraer gráficas y confirmar hipótesis. Técnicas fundamentalmente **estadísticas**.
  - Se usa exclusivamente información interna a la organización.
- Finales de los 90
  - Data-Mining: descubrimiento de patrones. Técnicas de **aprendizaje automático** para generar patrones novedosos.
  - El Data-Warehouse incluye Información Interna fundamentalmente.
- Principios de los 00
  - Técnicas de "scoring" y **simulación**: descubrimiento y uso de modelos globales. Estimación a partir de variables de entrada de variables de salida (causa-efecto) utilizando simulación sobre el modelo aprendido.
  - El Data-Warehouse incluye **Información** Interna y **Externa** (parámetros de la economía, poblacionales, geográficos, etc.).

22

Minería de Datos no Estructurados

- ***Web Mining*** se refiere al proceso global de descubrir información o conocimiento potencialmente útil y previamente desconocido a partir de datos de la Web. (Etzioni 1996)

Web Mining combina objetivos y técnicas de distintas áreas:
- Information Retrieval (IR)
- Natural Language Processing (NLP)
- Data Mining (DM)
- Databases (DB)
- WWW research
- Agent Technology

Se puede distinguir entre:
- *web content mining.*
- *web structure mining.*
- *web use mining.*

23

Para saber más... Direcciones

Recursos Generales:
- KDcentral (www.kdcentral.com)
- The Data Mine (http://www.the-data-mine.com)
- Knowledge Discovery Mine (http://www.kdnuggets.com)

Mailing list:
- KDD-nuggets: moderada y con poco ruido:
  Para suscribirse, enviar un mensaje a "kdd-request@gte.com" con "subscribe kdnuggets" en la primera línea del mensaje (el resto en blanco).

Revistas:
- Data Mining and Knowledge Discovery. (http://www.research.microsoft.com/)
- Intelligent Data Analysis (http://www.elsevier.com/locate/ida)

Asociaciones:
- ACM SIGDD (y la revista "explorations",
  http://www.acm.org/sigkdd/explorations/instructions.htm)

24

Bibliografía

Muy recomendables (generales y asequibles)

- **Berry M.J.A.; Linoff, G.S. "Mastering Data Mining" Wiley 2000.**
- **Berthold, M.; Hand, D.J. (ed) "Intelligent Data Analysis. An Introduction" Springer 1999. (Nueva edición a aparecer en 2002).**
- **Dunham, M.H. "Data Mining. Introductory and Advanced Topics" Prentice Hall, 2003.**
- **Fayyad, U.M.; Piatetskiy-Shapiro, G.; Smith, P.; Ramasasmy, U. "Advances in Knowledge Discovery and Data Mining", AAAI Press / MIT Press, 1996.**
- **Han, Jiawei; Micheline Kamber "Data Mining: Concepts and Techniques" Morgan Kaufmann, April 2000.**
- **Hand, David J.; Heikki Mannila and Padhraic Smyth "Principles of Data Mining", The MIT Press, 2000.**
- **Witten, I.H.; Frank, E. "Tools for Data Mining", Morgan Kaufmann, 1999.**

Bibliografía

Más específicos o técnicamente más duros:

- Dzeroski, S.; Lavrac, N. "Relational Data Mining" Springer 2001.
- Etzioni, O. "The World-Wide Web. Quagmire or Gold Mine" Communications of the ACM, November 1996, Vol. 39, nº11, 1996.
- Fayyad, U.M.; Grinstein, G.G.; Wierse, A. (eds.) "Information Visualization in Data Mining and Knowledge Discovery" Morgan Kaufmann 2002.
- Mena, Jesus "Data Mining Your Website", Digital Press, July 1999.
- Pyle, D. "Data Preparation for Data Mining", Morgan Kaufmann,1999.
- Thuraisingham, B. "Data Mining. Technologies, Techniques, Tools, and Trends", CRC Press, 1999.
- Wong, P.C. "Visual Data Mining", Special Issue of *IEEE Computer Graphics and Applications*, Sep/Oct 1999, pp. 20-46.