Introduction to Data Mining

José Hernández-Orallo

Dpto. de Sistemas Informáticos y Computación Universidad Politécnica de Valencia, Spain

jorallo@dsic.upv.es

Roma, 14-15th May 2009

1



- Motivation. BI: Old Needs, New Tools.
- Some DM Examples.
- Data Mining: Definition and Applications
- The KDD Process
- Data Mining Techniques
- Development and Implementation

- CRISP-DM (www.crisp-dm.org) (*CRoss-Industry* Standard Process for Data Mining)
 - A company consortium (initially under the funding of the European Commission), which includes SPSS, NCR and DaimlerChrysler.



- Business Understanding:
 - Understand the project goals and requirements from a business perspective. Substages:
 - establishment of business objectives (initial context, objectives and success criteria),
 - evaluation of the situation (resource inventory, requirements, assumptions and constraints, risks and contingences, terminology and costs and benefits),
 - establishment of the data mining objectives (data mining objectives and success criteria) and,
 - generation of the project plan (project plan and initial evaluation of tools and techniques).

- Data understanding:
 - Collect and familiarise with data, identify the data quality problems and see the first potentialities or data subsets which might be interesting to analyse (according the business objectives from the previous stage). Substages:
 - initial data gathering (gathering report),
 - data description (description report),
 - data exploration (exploration report) and
 - data quality verification (quality information).

- Data preparation:
 - The goal of this stage is to obtain the "minable view". Here we find: integration, selection, cleansing and transformation. Substages:
 - data selection (inclusion/exclusion reasons),
 - data cleansing (data cleansing report),
 - data construction (derived attributes, generated records),
 - data integration (mixed data) and
 - data formatting (reformatted data).

- Data modelling:
 - It is the application of modelling techniques or data mining to the previous minable views.
 Substages:
 - selection of the modelling technique (modelling technique, modelling assumptions),
 - evaluation design (test design),
 - model construction (chosen parameters, models, model description) and
 - model evaluation (model measures, revision of the chosen parameters).

Evaluation:

- It is necessary to evaluate (from the view point of the goal) the models of the previous stage. In other words, if the model is useful to answer some the business requirements. Substages:
 - result evaluation (evaluation of the data mining results, approved models),
 - revise the process (process revision) and,
 - establishment of the following steps (list of possible actions, decisions).

Deployment:

- The idea is to exploit the potential of the extracted models, integrate them in the decisionmaking processes of the organisation, spread reports about the extracted knowledge, etc.
 Substages:
 - deployment planning (deployment plan),
 - monitoring and maintenance planning (monitoring and maintenance plan),
 - generation of the final report (final report, final presentation) and,
 - project revision (documentation of the experience).

Progressive implementation on an organisation:





Elder Research, www.dataminglab.com

11

- Links to Commercial and non-commercial DM Software:
 - http://www.kdnuggets.com/software/index.html
- Free:
 - WEKA (http://www.cs.waikato.ac.nz/~ml/weka/) (Witten & Frank 1999, 2006)
 - Rproject: free tool for statistical analysis (http://www.R-project.org/)

EXAMPLE: Clementine

www.spss.com

- Tool that includes:
 - Several data sources (ASCII, XLS and many DBMS through ODBC).
 - Visual interface.
 - Several data mining techniques: neural networks, decision trees, rules, a priori, regression, ...
 - Data processing (pick & mix, combination and separation).
 - Report and batch facilities.

EXAMPLE: Clementine (www.spss.com)



14

EXAMPLE: Clementine

- Drug study
 - A number of hospital patients suffer a pathology which can be treated with a wide range of drugs.
 - 5 different drugs are available. Patients respond differently to these drug.
 - Problem:

Which drug is the most appropriate one for a new patient?

EXAMPLE: Clementine.

First step: DATA ACCESS:

- Read the data: e.g. a textfile with delimiters.
- The fields are named:

age	age			
sex	gender			
BP	blood pressure (High, Normal, Low)			
Cholesterol	cholesterol (Normal, High)			
Na	Sodium concentration in blood.			
Κ	Potasium concentration in blood.			
drug	drug to which the patient reacted			
	satisfactorily.			

The attributes/variables can be combined: E.g. A new attribute (Na/K), can be added.

EXAMPLE: Clementine

Second Step: Familiarisation with the data. We visualise the records:

table1								
Table								
Age	Sex	BP	Chole strol	Na	К	Orug	Na _to_ K	\square
23	F	HIGH	HIGH	0,79	0,03	drug¥	25,35	
47	Н	LOM	HIGH	0,74	0.06	drugC	13.09	
47	Н	LOH	HIGH	0.7	0.07	drugC	10.11	
28	F	NORMAL	HIGH	0,56	0,07	drugX	7,8	
61	F	LOM	HIGH	0.56	0,03	drug¥	18.04	
22	F	Normal	HIGH	0 .68	0.08	drugX	8.61	
49	F	NORMAL	HIGH	0,79	0,05	drug¥	16.28	
41	Н	LOH	HIGH	0,77	0.07	drugC	11.04	
60	Н	NORMAL	HIGH	0 .78	0.05	drugY	15,17	
43	Н	LOM	NORMAL	0.53	0.03	drug¥	19.37	∇

EXAMPLE: Clementine

- Allows field selection and filtering.
- Can show graphically some data properties. E.g. : Which is the proportion of cases which reacted well to the drug?

Drug						
Distribution						
Yalue	Proportion	2	Occurences			
drugA		11,5	23			
drugB		8.0	16			
drugC		8.0	16			
drugX		27.0	54			
drug¥		45,5	91 🔽			

EXAMPLE: Clementine

• Can find relations. E.g:

The relation between sodium and potasium is shown in a plot.



We observe an apparently random distribution (except from drug γ)

EXAMPLE: Clementine

- We can clearly observe that the patients with high Na/K quotient respond better to drug Y.
- But we want a classification model for every new patient, i.e.:

Which is the best drug for each patient?

Third step: Model construction

Tasks performed in Clementine:

- Filter non-desired (irrelevant) attributes.
- Type the fields.
- Construct models (rules, decision trees, neural networks₂₀...)

EXAMPLE: Clementine

This process is performed and graphically visualised in



EXAMPLE: Clementine	Rule	Folding	Select	Generate	View
EXAMPLE: Clementine Models can be browsed:	Rule Na_to_ BF	Folding K < 16.08 HIGH Age < 40 Choi Age >= 4 Age Age LOW Cholestr Na_f Cholestr	Select 4 6 lestrol H 46 < 60 >= 60 rol HIGH to_K < 19 to_K >= 1 rol NORMA	Generate HIGH -> dru WORMAL 5.013 -> dr L5.013 -> d HL -> drugX	9A ugC rugY
	Na_to_	Na_to_K Na_to_K _K >= 16.08	< 14.884 >= 14.88 84 -> dru	F −> drugX 34 −> drugY 49Y	

The rules extend the same criterion which was discovered previously, i.e., drug *Y* for the patient with high Na/K ratio. But it also gives rules for the rest.

EXAMPLE: SAS ENTERPRISE MINER (EM)

- Suite that includes:
 - Database connection (through ODBC and SAS datasets).
 - Sampling and inclusion of derived variables.
 - Data evaluation through dataset split into: training, validation (in case) and test.
 - Different data mining techniques: decision trees, regression, neural network, clustering, ...
 - Model comparisons.
 - Model conversion into SAS code.
 - Graphical interface.
- Also includes tools for all the process flow: the stages can be repeated, modified and stored.

EXAMPLE:

M Enterprise Miner - MY DM Project File Edit View Actions Options Window Help

Running 3 nodes

🗚 🖾 💹 🌺 🦣 🔯 🍉 🗵 🏢 🥠 Sample Explore Modify Model Assess Utility MY DM Project SegLoan Analysis E Bata Sources German Credit Home Equity Fraud Purchase ~£ੋਜ਼• Decision Tree 🔄 Diagrams Reg Fraud Detection 77 DMNeural 🍇 Loan Analysis 🐺 Purschase Propsensity Model Packages 🗄 <u> n</u> Users •**E**f@• 1 ≁≣ • Wayne Thompson ά. Credit risk Data Partition Model Score Impute Regression • Comparison Property Value NUGETO amoorph Imported Data Exported Data $\sqrt{\sqrt{2}}$ ≥: Variables Variable Selection AutoNeura Decile Bin 20 ScoreDist Bin 20 ROC Chart Yes ROC Epsilon 0.01 Selection Statistic Default Segment Profile , <u>w</u> Time of Creation 11/10/05 10:25 AM Cluster Run Id 07f3fdbb-a9b1-4bb0--Last Error -Last Status Complete -Needs Updating Yes Needs to Run No Time of Last Run 11/10/05 10:25 AM 0 2 -MBR SAS Code ScoreDist Bin Number of bins for assessment score distribution data set. -•

ENTERPRISE MINER (EM)

(process flow, KDD)

SAS

💘 Connected to SASMain - Logical Workspace Server

- 🗆 🗙

Weka, University of Waikato, NZ. (cs.waikato.ac.nz)



Weka, University of Waikato, NZ. (cs.waikato.ac.nz)



Angoss Knowledge Seeker:

le GE	Whole Dataset	Income+'<=50K'	Income=">50K"
unents and Si nus] census Data workclass	T Fraktings Dostigs Destropather Withorkpay		
frikesgt	transverritering	bactory utility	tersor rucad tersor rucad tersor rucad tersor rucad tersor rucad tersor rucad tersor rucad tersor rucad tersor rucad
education	Accessed Acc		
education-num		1	

27

 Oracle: Tools "Business Intelligence" and "Data Mining"



OracleBI Data Miner

😻 Oracle Data Miner - Table : CD_BUYERS				📽 Result Viewer: "DM4JSCD_BUYER19890_TM"			
File View Data Activity Tools Help				File Publish Help			
Navigator Build	Structure Date			Predictive Confidence Accuracy ROC Lift Test Settings Task			
Cracle_CB Apply	Eatoh Size: 100	Entch Next	Patrash	Name: "DM4.IST798810292813. R" Confusion Matrix:			
E L Mining Act	Petch Size. 100	Fetch Next	Reiresn	Chart Others 1			
E Association Rules	CUST_ID CD_BUY	ER AGE MAR	TA ANNUAL_IN	10 Others 816 87			
Attribute Importance	162 0	53 Marri 48 Marri	ad 90624	1 107 186			
E Classification	164 1	48 Marri	ed 199590	Hint: Rows = Actual; Columns = Predi			
E Clustering	165 1	30 Marri	ed 202051	0.8 True Positive Rate: 0 coverages			
Feature Extraction	166 0	29 Neve	M 220419	False Positive Rate: 0.0540122000			
E Cate Common	167 0	51 Sepa	M 113635	© 0.6 Threshold Avg Accuracy: 0.7692333859			
CRERCER	169 0	71 Marri	ed 205011	Diagonal Overall Accuracy: 0.8377926421			
te control views	170 0	21 Neve	M 111676	g 0,4 Cost: 194			
E Tables	171 0	34 Marri	ed 227359	Probability Threshold: 0,4705882353			
🖾 ABN_APPLY_OUTPUT_J	172 1	45 Divor	c. 148549	0.2 Derived Cost Matrix			
BABN_TEST_APPLY_OUT	174 0	20 Neve 27 Marri	ed 140863	Others 1			
ABNBUILDSETTINGS_JDN	175 0	30 Sepa	. 295612				
ABNCOMPUTETESTMETR	176 1	41 Marri	ed 189956	0.0 0.2 0.4 0.6 0.8 1.0			
A ADNOCOMPOLETESTMETR	177 0	25 Neve	M 266668	False Positive Rate Hint: Rows = Actual; Columns = Predi			
	179 0	45 Divor 27 Neve	M 213842	Area Under Curve:0.874251			
Activity tasks	180 0	36 Neve	M 212856				
	181 0	34 Marri	ed 289731	Detail: False Positive Cost: 1 False Negative Cost: 1 Compute Cost			
Sesult Viewer: CD_BUYERS20881_DT				reshold False Positive False Neg True Positi True Nega Accuracy Avg Accuracy Cos			
<u>File Publish Help</u>				3 87 1107 1186 1816 10.8377926410.769233381194.			
Tree Results Build Settings Task				Service Attribute			
tu a Charun Laguras Onlu		~		Deta Source: CBERGER. CD BUYERS Attribute: AGE			
Node ID Predicate	Predicted V	Confidence	Cases Sun				
E0 true	0	0.7600	1,804 1.00	Histogram for:AGE Statistics:			
E1 RELATIONSHIP is in { Husb	and 0	0.5440	818 0.45	4 Sample count: 3000			
13 PAYROLL_DEDUCTION <=	97.5 U 69.5 0	0.7852	298 0.16	2 S24.3 Minimum value: 17			
14 PAYROLL_DEDUCTION > 6	9.5 0	0.7027	185 0.10	243-31,8 Maximum value: 90			
E3 PAYROLL_DEDUCTION > 9	7.5 1	0.5942	520 0.28	2 31.0 - 38.9 Average value: 38.5			
	. <= 0	0.6746	126 0.06	38.9 40.2 Variance: 186.88			
16 CAPITAL_GAIN > 5715.5	1	1.0000	10 0.00	5 Sigma: 13.67			
S AVE_CHECKING_BALANCE	≥ 1 1	0.6802	394 0.21	4 C 63.5 - 60.8 Skewness: 0.61			
⊟6 OCCUPATION is in { ? Clent 18 CAPITAL GAIN <= 5463.0	c. Cr 1	0.5337	193 U.10 170 0.09	U 60.8 - 68.1			
19 CAPITAL_GAIN > 5463.0	1	1.0000	23 0.01	68.1 - 75.4			
17 OCCUPATION is in (Armed-	F Ex 1	0.8209	201 0.11	4 75.4-82.7			
DCI (THONOUID is is (Natis		111204	1301 1021	* > 82.7			
Predicted Target Value: 1				Null value			
Confidence 1.0000				0 100 200 300 400 500 600 700			
Cases: 23				Bin Count Binning Strategy:			
Levec 5				Group Value(s) Bin Count % of Total Equal Width			
Split Rules: Full Rule Surrogete				0 < 24.3 518 17.27			
CAPITAL_GAIN > 5463.0 AND OCCUPATION is in { ? Cleric, Crafts Farming Handler House	-s Machine Other Protec	Sales TechSun T	ransp. } AND	1 24.3 - 31.6 527 17.57 Graph orientation:			
AVE_CHECKING_BALANCE > 101.5 AND		and the second se		2 31.6 - 38.9 573 19.1 Vertical			
RELATIONSHP is in (Husband Wife)				4 46.2 - 53.5 368 12.27 Horizontal			

29

MS SQL SERVER: Analysis Services

- OLAP Services in SQL Server 97 was extended in SQL Server 2000 with DM features. This was called "Analysis Services". Much more techniques included in the new SQL Server (2005).
- In SQL Server 2007, three different interfaces available, extended DMX language.
- It is based on the "OLE DB for Data Mining": an extension of the DB access protocol: OLE DB.
- Implements an SQL extension which works with DMM (Data Mining Model).

Mining Non-structured Data

- Web Mining refers to the "global process of discovering information and knowledge which can be potentially useful and which is previously unknown from data on the web". (Etzioni 1996)
- Web Mining combines goals and techniques from different areas:
 - Information Retrieval (IR)
 - Natural Language Processing (NLP)
 - Data Mining (DM)
 - Databases (DB)
 - WWW research
 - Agent Technology
- There are several kinds of web mining:
 - web content mining.
 - web structure mining.
 - web use mining.

To know more... Some pointers

- General resources:
 - www.kdnuggets.com
- Associations:
 - ACM SIGKDD (and the journal: "explorations")
 - http://www.sigkdd.org/explorations/issue.php?issue=current
- Some books:
 - Berry M.J.A.; Linoff, G.S. "Mastering Data Mining" Wiley 2000.
 - Berthold, M.; Hand, D.J. (ed) "Intelligent Data Analysis. An Introduction" Second Edition, Springer 2002.
 - Dunham, M.H. "Data Mining. Introductory and Advanced Topics" Prentice Hall, 2003.
 - Han, Jiawei; Micheline Kamber "Data Mining: Concepts and Techniques" Morgan Kaufmann, April 2000.
 - Witten, I.H.; Frank, E. "Tools for Data Mining", Morgan Kaufmann, 2005.