DATA MINING USING PENTaho / WEKA

Yannis Angelis
Channels & Information Exploitation Division
Application Delivery Sector
EFG Eurobank
Agenda

• BI in Financial Environments

• Pentaho Community Platform

• Weka Platform

• Integration with Pentaho Platform

• Data mining using Weka
A brief (BI) history...

- **1970**: Mainframe OLAP
- **1980**: Spreadsheets
- **1990**: Desktop OLAP
- **2000**: Open-source OLAP
- **2010**: Open-source BI suites

- **APL**
- **Express**
- **Mondrian**
- **Pentaho**
- **JPivot**
- **Business Objects**
- **Cognos PowerPlay**
- **Comshare**
- **Microsoft OLAP Services**
- **Informix MetaCube**
- **Oracle partitions, bitmap indexes**
- **MySQL**
- **Sybase IQ**
- **JRubik**
- **BEE**
- **OpenI**
- **Palo**
- **Codd’s “12 Rules of OLAP”**
- **Essbase**
- **JPivot**
- **Palo**
- **Express**
- **APL**
Banking flavors of BI

- Data warehouse
  - Data integration / cleansing (ETL)
- Reporting
  - Operational
  - Analytical
- Data Mining
  - Risk related areas
    - Fraud detection / prediction
    - Credit scoring / scorecards
  - Marketing related areas
    - Next best activity
    - Churn management
    - Anti-attrition
Eurobank specificities

- IBM DW implementation (blueprint & localized)
- Essbase & MS OLAP engines
- Cognos & MS Reporting Services (BO in some cases)
- Different DB flavors in data marts (MS SQL, Oracle, DB2)
- SPSS & SAS analytics
- Excel – mania...
- Open Source BI Solutions
Pentaho... explained

- Pentaho Community
  - BI Platform
  - Reporting (PRD)
  - Mondrian (OLAP engine)
  - Kettle (PDI - ETL)
  - Weka (Data mining)

- All build under Java platform
Pentaho in action...
The *Weka* or *woodhen* (*Gallirallus australis*) is an endemic bird of New Zealand. (Source: *WikiPedia*)
Weka platform

- **WEKA (Waikato Environment for Knowledge Analysis)**

- Funded by the NZ government (for more than 10 years)
  - Develop an open-source state-of-the-art workbench of data mining tools
  - Explore fielded applications
  - Develop new fundamental methods

- Became part of Pentaho platform in 2006 (PDM – Pentaho Data Mining)
Weka Functionality & Architecture

• Comprehensive set of data pre-processing tools, learning algorithms and evaluation methods
  – data pre-processing tools
  – classification/regression algorithms
  – clustering algorithms
  – attribute/subset evaluators & search algorithms for feature selection
  – algorithms for finding association rules

• Graphical user interfaces (incl. data visualization)

• Environment for comparing learning algorithms

• Modular, object-oriented architecture
  – Packages for different types of algorithms (filters, classifiers, clusterers, associations, attribute selection etc.)
  – Sub-packages group components by functionality or purpose E.g. classifiers.bayes, filters.unsupervised.attribute
  – Algorithms are Java Beans (implementing specific interface)
  – GUIs use introspection/reflection to dynamically generate editor dialogs at runtime
Weka UI

<table>
<thead>
<tr>
<th>Component</th>
<th>Parameters</th>
<th>Time</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>KnowledgeFlow</td>
<td></td>
<td>0:0:39</td>
<td>Welcome to the Weka Knowledge Flow</td>
</tr>
</tbody>
</table>
• **Extensibility**
  – Plugin mechanisms allow WEKA to be extended without modifying the classes in the WEKA distribution
  – New algorithms can be implemented and added

• **Standards & Interoperability**
  – PMML Support (import / export)
  – LibSVM / SVM-Light data format support
Weka / Pentaho Integration

• Main point of integration is with Kettle (PDI) project
  – Kettle can export data in ARFF format
    • High-volume, low memory consumption
  – Kettle gets WEKA-specific transformation steps
    • **WekaScoring**: score data using a pre-constructed WEKA model (classification, regression or clustering) or PMML model as part of an ETL transformation
    • **KnowledgeFlow**: execute arbitrary Knowledge Flow processes as part of an ETL transformation
    • **WekaForecasting**: load or import a time series forecasting model (created in Weka’s time series analysis and forecasting environment) as part of an ETL transformation

• Only WekaScoring can be used (is free and downloadable) with the community edition of Pentaho
Data mining with Weka

• (One-of-the-many) **Definition**: Extraction of implicit, previously unknown, and potentially useful information from data

• **Goal**: improve marketing, sales, and customer support operations, risk assessment etc.
  – Who is likely to remain a loyal customer?
  – What products should be marketed to which prospects?
  – What determines whether a person will respond to a certain offer?
  – How can I detect potential fraud?

• **Central idea**: historical data contains information that will be useful in the future (patterns => generalizations)

• **Data Mining** employs a set of algorithms that automatically detect patterns and regularities in data
Eurobank’s Case

- **Problem**: Prediction (Probability Score) of a Corporate Customer Delinquency (or default) in the next year

- Customer historical data used include:
  - Customer footings behavior (assets & liabilities)
  - Customer delinquencies (rates and time data)
  - Business Sector behavioral data
Implementation Thoughts...

- Variable selection using the *Information Value* (IV) criterion.

\[
IV = \sum_i (G_i - B_i) \ln \left( \frac{G_i}{B_i} \right)
\]

<table>
<thead>
<tr>
<th>Value of IV</th>
<th>Statistical strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>less than 0.02</td>
<td>a very weak statistical relation</td>
</tr>
<tr>
<td>0.02 – 0.1</td>
<td>a weak statistical relation</td>
</tr>
<tr>
<td>0.1 – 0.3</td>
<td>an average statistical relation</td>
</tr>
<tr>
<td>0.3 – 0.5</td>
<td>a strong statistical relation</td>
</tr>
<tr>
<td>greater than 0.5</td>
<td>an extremely strong statistical relation</td>
</tr>
</tbody>
</table>

- Automatic Binning of continuous data variables was used (Chi-merge). Manual corrections were made to address particularities in the data distribution of some variables (using again IV)
Weka Case Screenshots

Attribute selection output:

```
10 Ratio_l_12_2
19 GROUP_HDADS_LAST_PERIOD
33 LOANS_L06_1

Merit of best subset found : 0.0537
```

Attribute Subset Evaluator (supervised, Class (nominal): 1 default):

CFS Subset Evaluator
including locally predictive attributes

Selected attributes: 2,7,10,11,13,15,17,20,23,25,28,31,35,36,37,38 : 15

FLAG_1
cispr_12_AMOUNT_1
apli1_2_2
apli1_6_2
Max_Bucket_WHOL_12_1
TimeInTwoPlus_WHOL_12_1
TimeSinceTwoPlus_WHOL_12_1
E_SN_1
Times_IN_6_1
Time_rumice_6_1
Trend_D_DC_1_0_2
0_NO_16_1
0_NO_112_1
0_NT_lA_1
0_NT_l4_1
0_NT_l4_2
```
Visualizations
Weka Limitations...

- Traditional algorithms need to have all data in (main) memory
  => big datasets are an issue

- Solution:
  - Incremental schemes
  - Stream algorithms

MOA (Massive Online Analysis)
http://moa.cs.waikato.ac.nz/
Conclusion...

• Try it yourself...

• WEKA is available at
  http://www.cs.waikato.ac.nz/ml/weka

• A list of projects based on WEKA include applications for natural language processing, text mining etc.