Intelligent System Project

(Part VII - Software Tools)

Miquel Sànchez i Marrè

miquel@cs.upc.edu

http://kemlg.upc.edu/menu1/miquel-sanchez-i-marre

Course 2015/2016

https://kemlg.upc.edu
PART 7 – SOFTWARE TOOLS
Software Tools for an IPS Development

- WEKA
- KLASS
- R System
- GESCONDA
- Rapid Miner
- Knime
- Matlab
- Drools (Rule-Based Engine) / CLIPS
- Jcolibri (CBR engine)
- Protegé (Ontological Reasoning)
- ...
WEKA

https://kemlg.upc.edu
Weka intro

- Weka is a set of Machine Learning algorithms written in Java developed by the University of Waikato (New Zealand)
  - Initially support the book “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations” by Ian H. Witten, Eibe Fran

- Web portal with all the information, downloading, etc.:

- Weka algorithms can be used as isolated ML tools or used into your own Java code.

- There are 3 kinds of interaction:
  - SimpleCLI: command line
  - Knowledge explorer: GUI to easily integrate data loading and data selection with the application of several algorithms
  - Experimenter: to maintain partial results of learning projects.
Weka Use

Initial Data

(1) Data Splitting

(2) Loading

(3) ML Method Selection and Parameters' determination

Weka

Training Data Set (format arff)

(4) Model

COBWEB

... 

PART

ID3

C4.5

Results

(5) Testing

Test Data Set (format arff)
Data Format (arff)

- Data Set Name: @relation
- Attribute information: @attribute
- Data rows: @data
- Data: Example rows, where in each row the attribute’s values are separated with commas.
- Sparse data (most attributes have a zero value in almost all examples)
- By default, the last attribute is the one marked as the CLASS attribute (Supervised data set)
## Format Arff: exemple

```plaintext
@relation contact-lenses

{young, pre-presbyopic, presbyopic}

@attribute age {myope, hypermetrope}
@attribute spectacle-prescrip {no, yes}
@attribute astigmatism {reduced, normal}
@attribute tear-prod-rate {soft, hard, none}
```

### Data

<table>
<thead>
<tr>
<th>Age</th>
<th>Spectacle-prescrip</th>
<th>Astigmatism</th>
<th>Tear-prod-rate</th>
<th>Contact-lenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>young</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>none</td>
</tr>
<tr>
<td>young</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>soft</td>
</tr>
<tr>
<td>young</td>
<td>myope</td>
<td>yes</td>
<td>reduced</td>
<td>none</td>
</tr>
<tr>
<td>young</td>
<td>myope</td>
<td>yes</td>
<td>normal</td>
<td>hard</td>
</tr>
<tr>
<td>pre-presbyopic</td>
<td>myope</td>
<td>no</td>
<td>reduced</td>
<td>none</td>
</tr>
<tr>
<td>pre-presbyopic</td>
<td>myope</td>
<td>no</td>
<td>normal</td>
<td>soft</td>
</tr>
<tr>
<td>pre-presbyopic</td>
<td>myope</td>
<td>yes</td>
<td>reduced</td>
<td>none</td>
</tr>
<tr>
<td>pre-presbyopic</td>
<td>myope</td>
<td>yes</td>
<td>normal</td>
<td>hard</td>
</tr>
</tbody>
</table>
```

© Miquel Sànchez i Marrè, KEMLG, 2015
KLASS

https://kemlg.upc.edu
The plattform KLASS

- Tool oriented to the Knowledge discovery in ill-structured domains  [Gib91]

- Functionalities:
  - Basic descriptive statistics
  - Reciprocal neighbours clustering
  - Clustering based on rules  [Gib94]
  - Boxplot based induction rules  [Gib04]
  - Serial data management.
  - Interpretation oriented tools:
    - Class-pannel graphs  [Gib05]
    - Class Characterization  [Gib02]
    - Classifications comparison (d-test)  [Gib98]
    - Visualization of hierarchical trees
    - Visualization on trees, generation and importation of a-cuts
  - Reporting support:
    - Direct generation of LaTeX font files
R System

- R is a language and environment for statistical computing and graphics.
- It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues.
- R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.
- One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.
- R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form.
- It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.
The R Environment

- R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes:
  
  - an effective data handling and storage facility,
  - a suite of operators for calculations on arrays, in particular matrices,
  - a large, coherent, integrated collection of intermediate tools for data analysis,
  - graphical facilities for data analysis and display either on-screen or on hardcopy, and
  - a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.
RStudio

R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
GESCONDA issues

- GESCONDA is an Intelligent Data Analysis System for knowledge management in environmental databases
- Statistical & Machine Learning techniques
- Explicit management of knowledge
- Data filtering
- A common language implementation platform: Java
- Including recommendation skills
- Dynamic data analysis
GESCONDA architecture
GESCONDA tool

© Miquel Sànchez-Marrè, KEMLG, UPC  
iEMSs 2014, June 16-19, San Diego, USA
Data Base Management & Statistical Data Processing

- **Database Management**
  - Adding variables
  - Deleting variables
  - Recoding variables
  - Standarization of variables
  - Random variable generation
  - Probability values generation

- **Statistical Data Filtering**
  - Descriptive statistical analysis:
    - one-way
    - two-way
  - Graphical representation
    - TS-Plots
    - Histograms
    - Letter plots
    - Two-way charts
### Statistical Analysis

#### Data Table

<table>
<thead>
<tr>
<th>PES</th>
<th>ID</th>
<th>pet-wid</th>
<th>sep-wid</th>
<th>pet-len</th>
<th>sep-len</th>
<th>CLASSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
<td>3.5</td>
<td>1.4</td>
<td>1.4</td>
<td>5.1</td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3.0</td>
<td>1.4</td>
<td>4.9</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3.2</td>
<td>1.3</td>
<td>4.7</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>3.1</td>
<td>1.5</td>
<td>4.6</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>3.5</td>
<td>1.4</td>
<td>5.0</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>3.9</td>
<td>1.7</td>
<td>5.4</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>3.4</td>
<td>1.4</td>
<td>4.6</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>3.4</td>
<td>1.5</td>
<td>5.0</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>2.9</td>
<td>1.4</td>
<td>4.4</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>3.1</td>
<td>1.5</td>
<td>4.9</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>3.7</td>
<td>1.5</td>
<td>5.4</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>3.3</td>
<td>1.6</td>
<td>4.8</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>3.0</td>
<td>1.4</td>
<td>4.6</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>3.0</td>
<td>1.1</td>
<td>4.3</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>4.0</td>
<td>1.2</td>
<td>5.8</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>4.1</td>
<td>1.5</td>
<td>5.7</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>3.9</td>
<td>1.3</td>
<td>5.4</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>3.5</td>
<td>1.4</td>
<td>5.1</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>3.3</td>
<td>1.7</td>
<td>5.7</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>2.5</td>
<td>1.5</td>
<td>5.1</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>3.4</td>
<td>1.7</td>
<td>4.6</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
<td>4.0</td>
<td>1.5</td>
<td>5.1</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>3.5</td>
<td>1.0</td>
<td>4.6</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>24</td>
<td>3.3</td>
<td>1.7</td>
<td>5.1</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>3.4</td>
<td>1.9</td>
<td>4.8</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>3.0</td>
<td>1.6</td>
<td>5.0</td>
<td></td>
<td>setosa</td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>3.4</td>
<td>1.6</td>
<td>5.0</td>
<td></td>
<td>setosa</td>
</tr>
</tbody>
</table>
### Sessió

<table>
<thead>
<tr>
<th>Variables</th>
<th>GESP</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th><strong>ESTADÍSTICA DESCRIPTIVA</strong></th>
<th><strong>UNIVARIANT</strong></th>
<th><strong>QUANTITATIVA</strong></th>
<th><strong>DESVIACIÓ</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>N?</td>
<td>Mitjana</td>
<td>Mediana</td>
</tr>
<tr>
<td>sep-len</td>
<td>150</td>
<td>0</td>
<td>5.643</td>
</tr>
<tr>
<td>pet-len</td>
<td>150</td>
<td>0</td>
<td>3.759</td>
</tr>
<tr>
<td>sep-wid</td>
<td>150</td>
<td>0</td>
<td>3.054</td>
</tr>
<tr>
<td>pet-wid</td>
<td>150</td>
<td>0</td>
<td>1.199</td>
</tr>
</tbody>
</table>

**Intelligent Project System**

© Miquel Sànchez i Marrè, KEMLG, 2015
Clustering module

- Clustering methods
  - Nearest Neighbour
  - COBWEB/3
  - Isodata
  - K-means
  - Marata
  - Hybrid clustering based on rules

- Randomised Pseudo-bagging techniques (new !!)
  - K-means
  - Nearest Neighbour
Hierarchical clustering (COBWEB/3)
Clustering plot
Randomised Pseudo-bagging

Bagging

class_0
class_1
class_2
class_3
class_4
class_5
class_6
class_7

Utilitzar Pesos
Primer Atribut
Inèrcia
Entropia

Acceptar
Cancelar

Entrada valor

Valor K (nombre de classes)
Increment mínim
0.05

Realitzar Repeticions
Nombre de repeticions
3
Nom de classe
class

Mostrar Llavor
Llavor inicial

Acceptar
Cancelar
Decision Tree Induction module

- Decision tree classifiers
  - ID3
  - C4.5
  - CART
- Functionalities
  - Pruning
  - Validation
C4.5 induced tree

Tree data:
- Attribute selection criteria: Gain ratio
- Validation method: No validation
- Pruning method: No pruning
- Number of learning cases: 150
- Predicted errors: 0

Decision Tree Application

Data View Tree Log

Process

PETAL_WIDTH [50, 50, 50]

PETAL_WIDTH < 0.3
- Iris-setosa
  [50, 0, 0]

PETAL_WIDTH >= 0.8
  PETAL_WIDTH [0, 50, 50]

PETAL_WIDTH < 1.8
  PETAL_LENGTH [0, 49, 5]

PETAL_LENGTH <= 5.25
  PETAL_LENGTH [0, 49, 3]
Statistical Modelling module

- Multiple Linear Regression
  - Parameter estimation
  - Evaluation of fitting coefficients
  - Providing tools for graphical residuals analysis
- ANOVA
  - Parameter estimation
  - Evaluation of fitting coefficients
  - Providing tools for graphical residuals analysis
- Logistic Regression (coming soon ...)
- Multivariate Data Analysis (coming soon ...)
Linear regression residuals
Classification Rule Induction module

- Methods
  - CN2 method
  - PRI SM method
  - RISE method
  - Rules method

- Functionalities
  - Rule pruning
  - Validation (simple validation, cross validation)
  - Prediction
  - Rule exportation (CLIPS format, TXT format)
Rules method
Rule set validation

Resultats Validació

Validació Creuada (estratificada = true, NSubconjunts = 3):

ALGORISME: Rules

PERCENTATGE CORRECTEZA REGLES DE CLASSIFICACIÓ DEL SUBCONJUNT 1 : 83.72%
PERCENTATGE CORRECTEZA REGLES DE CLASSIFICACIÓ DEL SUBCONJUNT 2 : 85.42%
PERCENTATGE CORRECTEZA REGLES DE CLASSIFICACIÓ DEL SUBCONJUNT 3 : 87.5%

PERCENTATGE CORRECTEZA REGLES DE CLASSIFICACIÓ : 85.54667%
Rule set exported in CLIPS format

(defrule Regla1
  (>= ?Llargada_pètal 1.0)
  (< ?Llargada_pètal 2.18)
  =>
  (assert (classe Iris-setosa)))

(defrule Regla2
  (>= ?Amplada_pètal 0.1)
  (< ?Amplada_pètal 0.58)
  =>
  (assert (classe Iris-setosa)))

(defrule Regla3
  (>= ?Amplada_pètal 1.06)
  (< ?Amplada_pètal 1.54)
  (>= ?Llargada_pètal 3.36)
  (< ?Llargada_pètal 4.54)
  =>
  (assert (classe Iris-versicolor)))

(defrule Regla4
  (>= ?Amplada_pètal 2.02)
  (< ?Amplada_pètal 2.5)
  =>
  (assert (classe Iris-virginica)))

...
Feature Relevance module

- Unsupervised Feature weighting methods
  - Gradient method
  - UEB-1
  - UEB-2

- Functionalities
  - Weight assignment
  - Weights exportation
Determinació automàtica de pesos

Tria l'algoritme i fixa els paràmetres corresponents:

Algorismes
- GD (Gradient)
- UEB-1 (Unsupervised Entropy Based 1)
- UEB-2 (Unsupervised Entropy Based 2)

Paràmetres:
- Alfa: 0
- Mu: 0
- Threshold: 1
- Exponent: -5
- Iteracions: 5

Acceptar  Cancel·lar

Pesos assignats

<table>
<thead>
<tr>
<th>Atributs i pesos</th>
<th>Valors</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLV55-B</td>
<td>6.3243303</td>
</tr>
<tr>
<td>DQO-E</td>
<td>6.7626553</td>
</tr>
<tr>
<td>DQO-D</td>
<td>5.8280478</td>
</tr>
<tr>
<td>DBO-E</td>
<td>8.54659</td>
</tr>
<tr>
<td>DBO-D</td>
<td>5.566908</td>
</tr>
<tr>
<td>Q-E</td>
<td>5.087847</td>
</tr>
<tr>
<td>V30-B</td>
<td>3.6127439</td>
</tr>
<tr>
<td>QR-G</td>
<td>9.481031</td>
</tr>
<tr>
<td>PH-E</td>
<td>4.863965</td>
</tr>
<tr>
<td>DQO-S</td>
<td>5.803375</td>
</tr>
<tr>
<td>PH-D</td>
<td>5.577009</td>
</tr>
<tr>
<td>MCRT-B</td>
<td>10.0</td>
</tr>
<tr>
<td>DBO-S</td>
<td>8.025634</td>
</tr>
<tr>
<td>QB-B</td>
<td>7.019821</td>
</tr>
<tr>
<td>SS-E</td>
<td>7.8666644</td>
</tr>
<tr>
<td>SS-D</td>
<td>6.9835567</td>
</tr>
<tr>
<td>SSV-E</td>
<td>8.026958</td>
</tr>
<tr>
<td>SSV-D</td>
<td>7.5566745</td>
</tr>
<tr>
<td>QP-G</td>
<td>5.0625114</td>
</tr>
<tr>
<td>PH-S</td>
<td>3.7629368</td>
</tr>
<tr>
<td>FE-E</td>
<td>6.8153143</td>
</tr>
<tr>
<td>SS-S</td>
<td>9.310967</td>
</tr>
<tr>
<td>QA-G</td>
<td>0.0</td>
</tr>
<tr>
<td>SSV-S</td>
<td>9.303681</td>
</tr>
<tr>
<td>MLSS-B</td>
<td>6.4036493</td>
</tr>
</tbody>
</table>

Acceptar  Exportar a fitxer  Rebutjar
Conclusions

- Capturing knowledge implicit in databases to get a central corpus for IDSS is possible
- Reliable IDSS for a real-world complex and ill-structured domains can be constructed
- GESCONDA is a promising Intelligent Data Analysis System for building real Intelligent Decision-Support Systems
- GESCONDA integrates both statistical and artificial intelligence data mining techniques
- Future recommendation and meta-knowledge tools will improve its usability
Miquel Sàncchez i Marrè  
(miquel@cs.upc.edu)  

http://kemlg.upc.edu/