UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Grado en Inteligencia Artificial

PMAAD

Advanced Profiling

Prof. Dante Conti

Prof. Sergi Ramirez

Créditos Prof. Tomás Aluja

The problem

Profiling. What is the profile of a group of individuals?

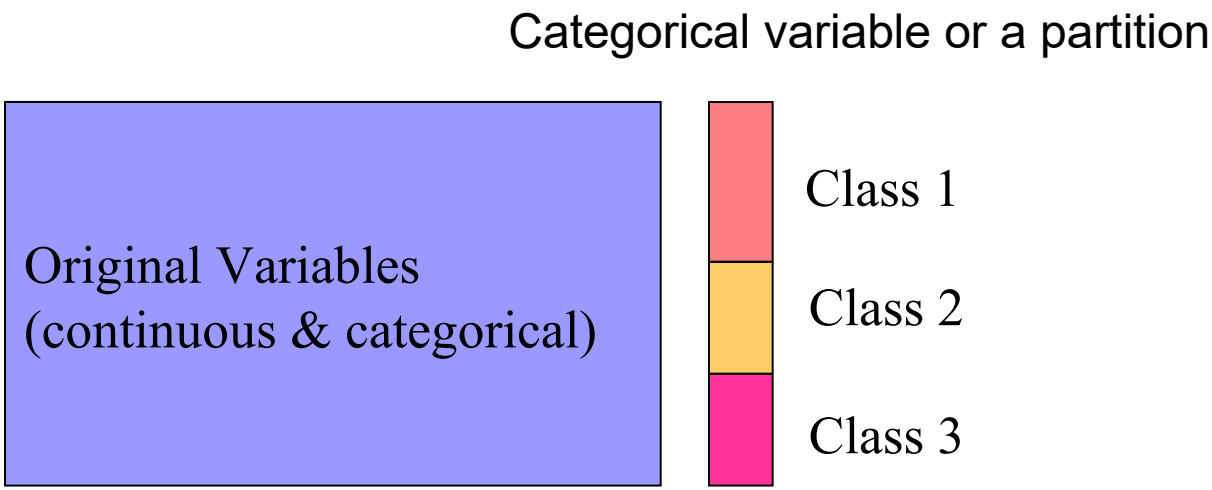
i.e.: a class resulting from a clustering algorithm, or a modality of a categorical variable. What is the profile of A buyers?

Profiling is finding the significant characteristics which make the group of individuals different than the whole set of individuals.

→ *Automatic detection of significant deviations*

Interpreting the classes

- Diferential characterisation among classes



- Statistical characterization

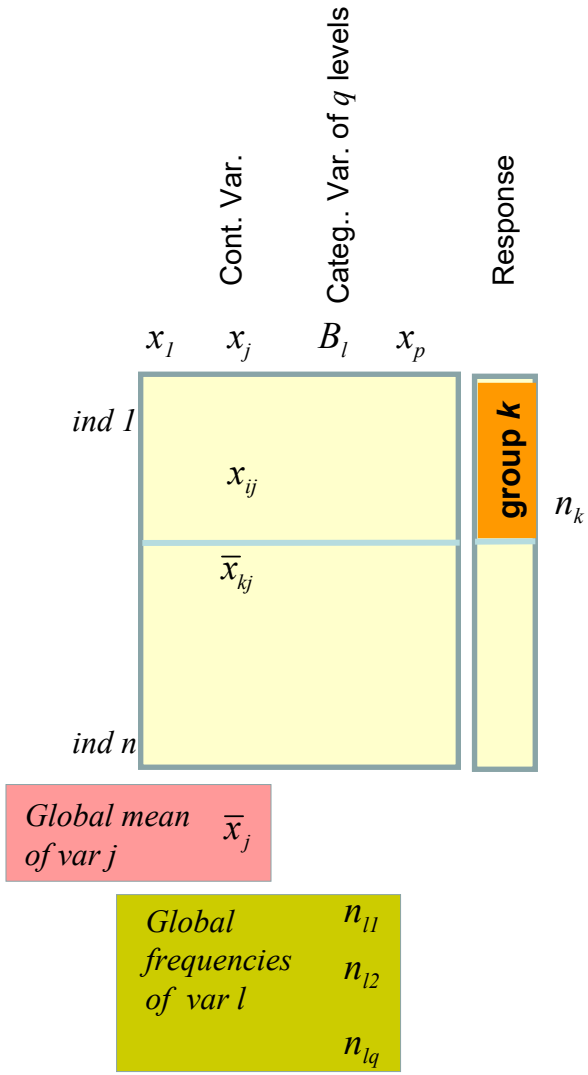
Automatic profiling of groups of individuals

We have a group of individuals defined by a level of a categorical variable.

- We take as response variable the variable identifying the groups that we want to find their profile.
- The explanatory variables are all the others and can be either categorical or continuous.

Problem: For every group of individuals detect

1. which modalities (of the categorical explanatory variables) deviate significantly from what were expected and,
2. which continuous explanatory variables, deviate significantly from what were expected.



Hypothesis test with continuous variables

groups	*means*	*counts*
1	$\bar{x}_1$	n_1
$\vdots$	$\vdots$	$\vdots$
q	$\bar{x}_q$	n_q
Global	$\bar{x}$	n

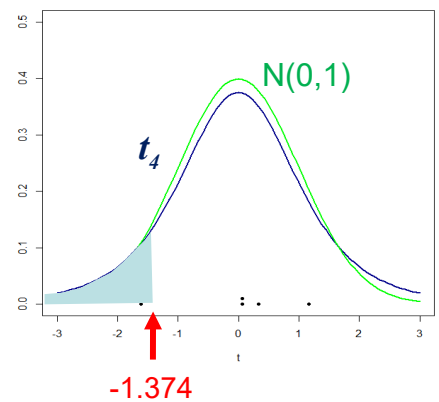
$$H_0 : \mu_k = \mu \quad k = 1, \ldots, q$$

Test statistic: Diference between the mean in group *k* and the global mean

Student's *t*

$$t = \frac{\bar{x}_k - \bar{x}}{\sqrt{(1 - \frac{n_k}{n}) \frac{s^2}{n_k}}} : \; t_{n-1}$$

William Gosset "Sudent", English, 1876, 1937

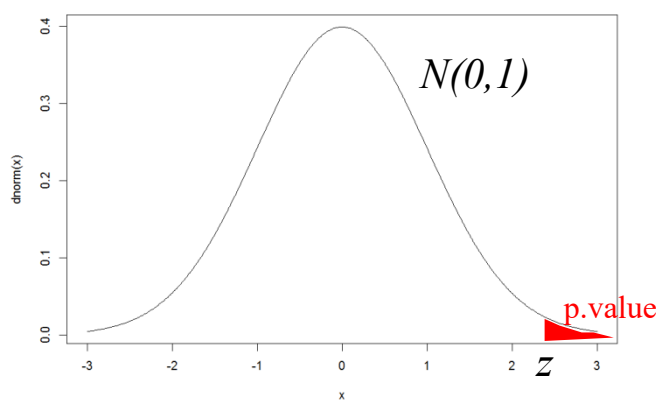


Rank the continuous variables by p.value (ascending)

Hypothesis test with categorical variables

$$
\begin{array}{c|ccc|c}
 & 1\ldots & j & \ldots J & \\
\hline
1 & & \vdots & & \\
k & \cdots & n_{kj} & \cdots & n_k \\
q & & \vdots & & \\
\hline
 & & n_j & &
\end{array}
$$

Test statistic: Difference between proportion of modality *j* in group *k* and proportion of modality *j* in whole data



$$H_0 : p_{j \cdot k} = p_j \quad k = 1,\ldots,q\,;\, j = 1,\ldots,J$$

Assumption of normality of proportions:

$$\frac{n_{kj}}{n_k} : \; N\left(p_j = \frac{n_j}{n}, \left(1 - \frac{n_k}{n}\right)\frac{p_j(1-p_j)}{n_k} \right)$$

$$z = \frac{\dfrac{n_{kj}}{n_k} - \dfrac{n_j}{n}}{\sqrt{\left(1 - \dfrac{n_k}{n}\right)\left(\dfrac{p_j\left(1-p_j\right)}{n_k}\right)}} : \; N(0,1)$$

Rank the levels of the categorical explanatory variables by p.value (ascending)

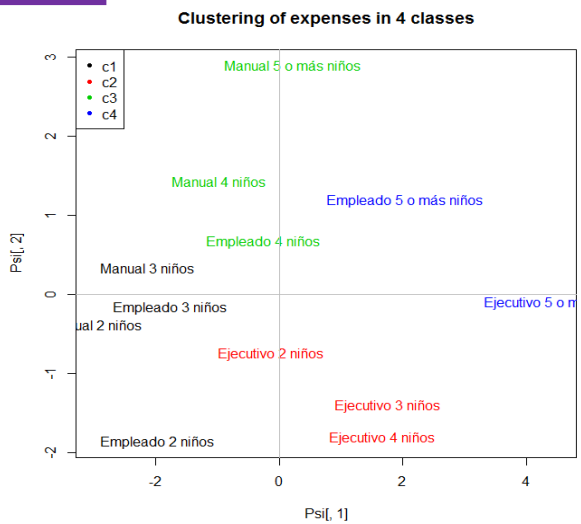
Profiling

- For each group to profile,
  1. rank the modalities of the categorical explanatory variables according their p-value (ascending).
  2. Likewise, rank the continuous variables according their p-value

- Select the most significant modalities and continuous variables by a threshold (0.05, 0.01, ..) defined a priori.
  (what matters is the ordering, actual significance depends on the number of individuals)

Interpreting the classes of expenses

```
> library(FactoMineR)
> catdes(data, num.var, proba = 0.05, row.w = NULL)
```




```
> catdes(cbind(as.factor(c1),despeses),1)

$category$`2`   Cla/Mod Mod/Cla   Global    p.value   v.test
STATUS=3            75     100 33.33333 0.01818182 2.361894

$category$`4`   Cla/Mod Mod/Cla Global    p.value   v.test
CHILDREN=5    66.66667     100     25 0.04545455 2.000424


$quanti$`1`    v.test Mean in category Overall mean sd in category Overall sd    p.value
fruits     -2.006717           369.75     505.0000       22.89514   158.0638 0.04477978
bread      -2.112728           354.25     446.6667       44.53299   102.5860 0.03462408
meat       -2.437108          1493.00    1886.7500       33.73426   378.9022 0.01480525
vegetables -2.492476           539.50     732.0000       67.18817   181.1261 0.01268558
poultry    -2.497153           548.75     803.1667       15.48184   238.9361 0.01251951

$quanti$`2`     v.test Mean in category Overall mean sd in category Overall sd    p.value
poultry      2.175824         1074.667     803.1667       104.4169   238.9361 0.02956844

$quanti$`4`     v.test Mean in category Overall mean sd in category Overall sd    p.value
vegetables   2.571343           1046.0       732.00           51.0   181.1261 0.01013051
milk         2.397322            539.5       358.25           21.5   112.1406 0.01651540
fruits       1.994058            717.5       505.00          169.5   158.0638 0.04614569
```

Dependence of a partition respect to a cont. Var.

Sir Ronald A. Fisher
English,1890-1962

To be used for feature selection

levels	*means*	*counts*
1	$\bar{x}_1$	n_1
$\vdots$	$\vdots$	$\vdots$
q	$\bar{x}_q$	n_q

$$H_0 : \mu_1 = \cdots = \mu_q = \mu$$

$\bar{x}, s^2$ global mean and global variance

$$F = \frac{S_B^2 / q-1}{S_W^2 / n-q} : F_{q-1,n-q}$$

$$S_W^2 = \sum_{k=1}^{q} \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)^2$$

$$S_B^2 = \sum_{k=1}^{q} n_k (\bar{x}_k - \bar{x})^2$$

Ranking of variables by p.values (ascending)

Equivalent statistic: $$\eta^2 = \frac{S_B^2}{S_W^2 + S_B^2}$$

Dependence of a partition respect to a categ. Var.

Karl Pearson
English, 1857, 1936

To be used for feature selection

$$
\begin{array}{c|ccc|c}
 & 1\ldots & j & \ldots J & \\
\hline
1 & & \vdots & & \\
k & \cdots & n_{kj} & \cdots & n_k \\
q & & \vdots & & \\
\hline
 & & n_j & &
\end{array}
$$

$$H_0 : n_{kj} = np_k p_j \quad \forall kj$$

$$X^2 = \sum_{k=1}^{p}\sum_{j=1}^{q} \frac{\left(n_{kj} - {n_k n_j}/{n}\right)^2}{{n_k n_j}/{n}} : \chi^2_{(p-1)(q-1)}$$

Ranking of variables by p.values (ascending)

Improvement when Interpreting Clusters Results

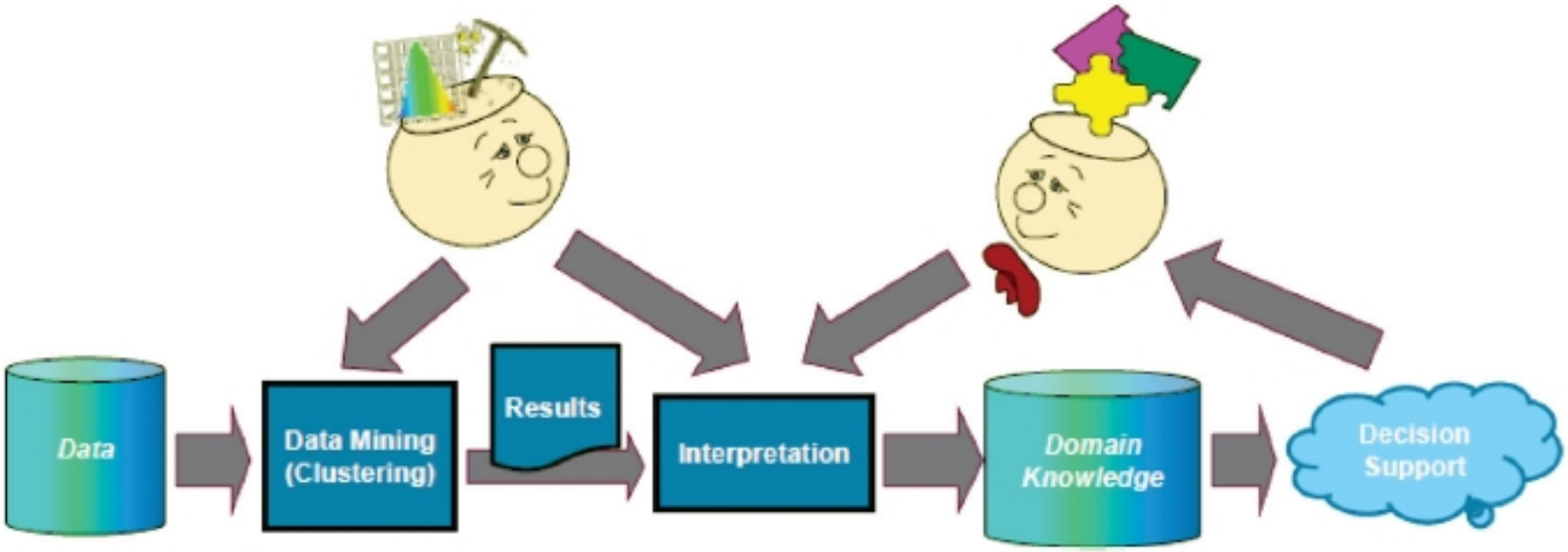
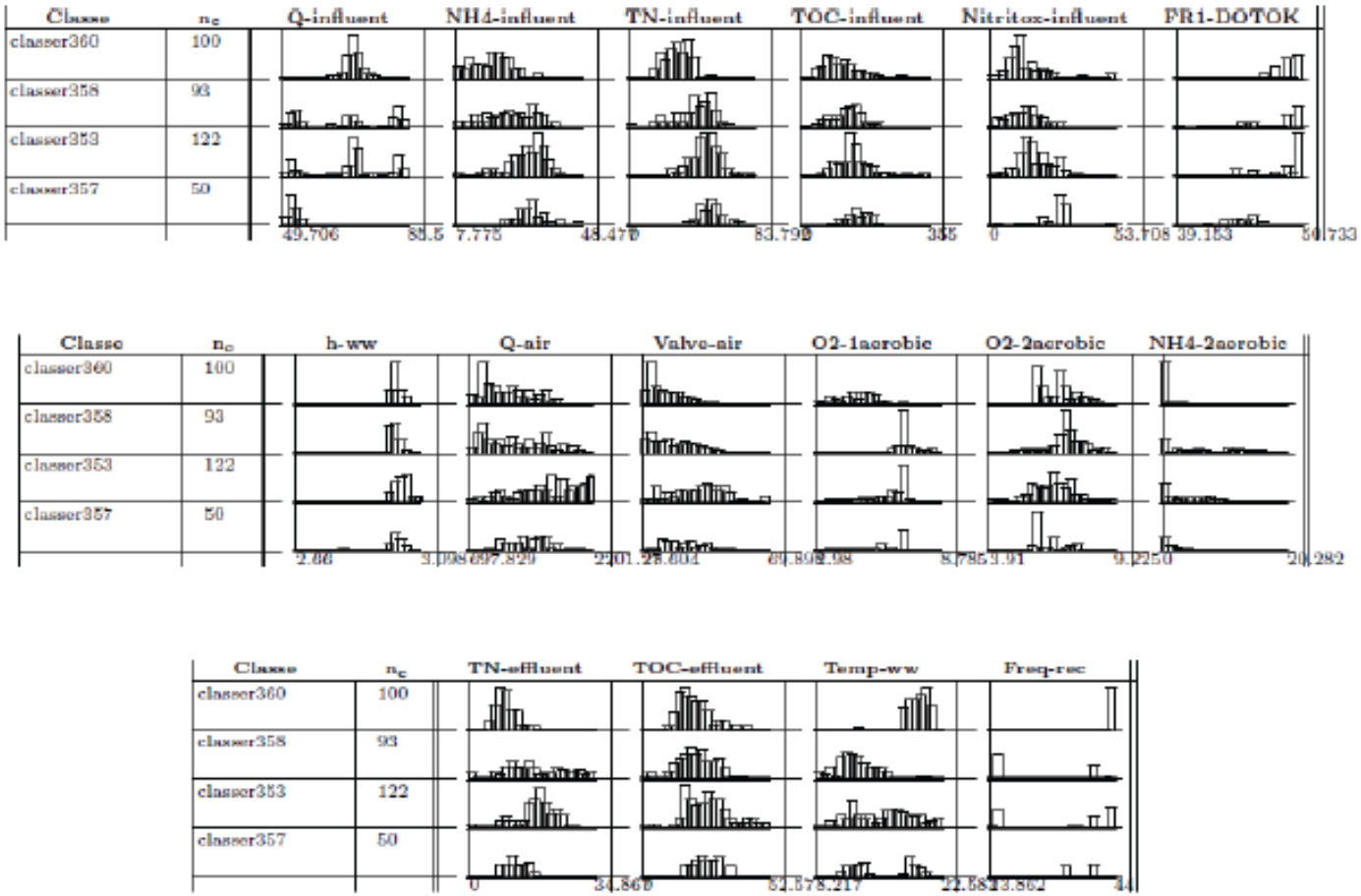


Figure 1. Interpretation support tools bridging the gap between DM and decision support.

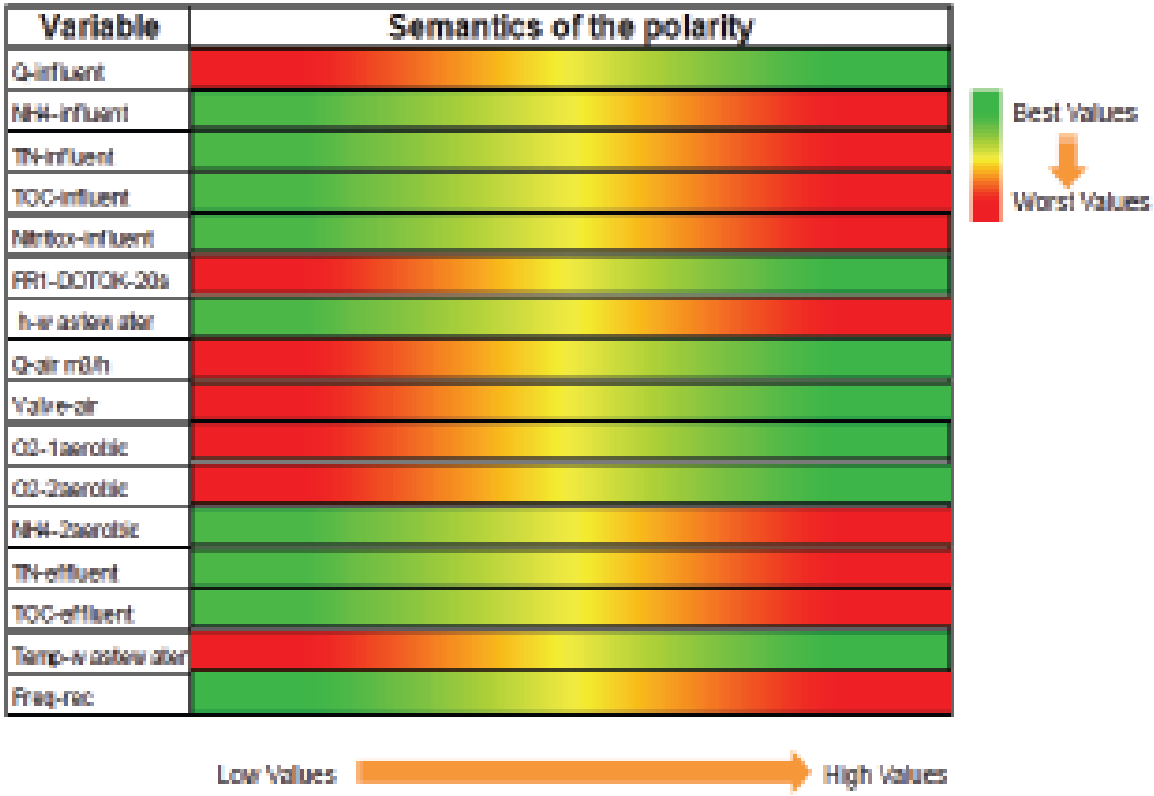
Advanced Profiling

CPG → Class Panel Graph



UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Semantics and Polarity



TLP → Traffic Light Panel

The analyst has to read the CPG, mark the characteristic values of the variables for the different classes and assign qualitative levels to those values. The dominant level of a variable in a given class can be determined in two ways:

a) Identify the qualitative level of the mean or median of the variable in the class,

b) Identify the qualitative level of the mode of the variable in the class.

When the variables have only 3 qualitative levels assigned, it is very interesting to assign the colours of a traffic light to those levels (red for the bad or negative value, yellow for the medium or neutral value and green for the good or positive value), being the *bad* values the higher or lower values of the variable depending on the variable's semantics. It is

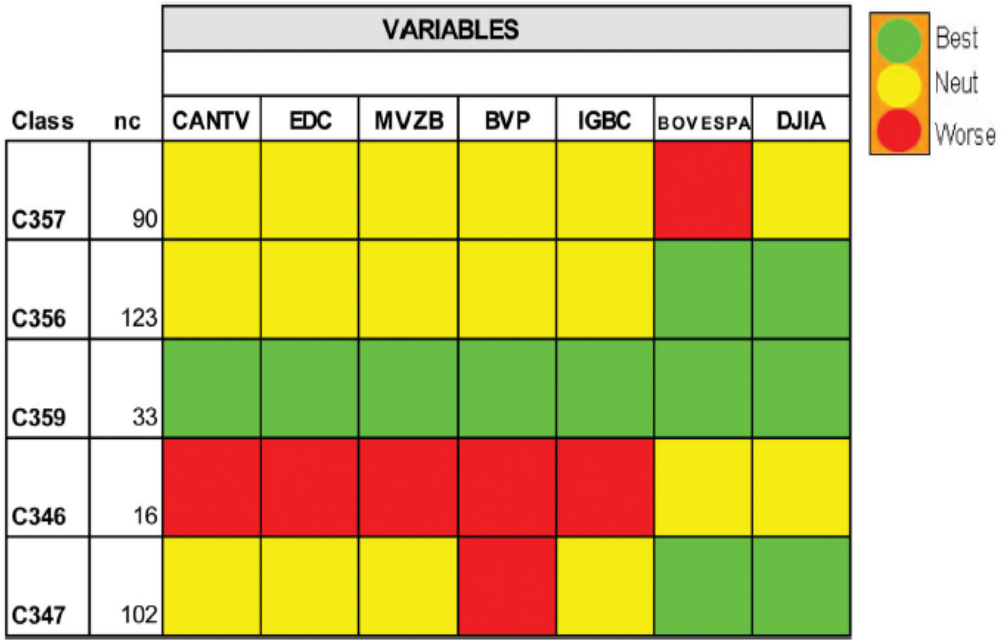
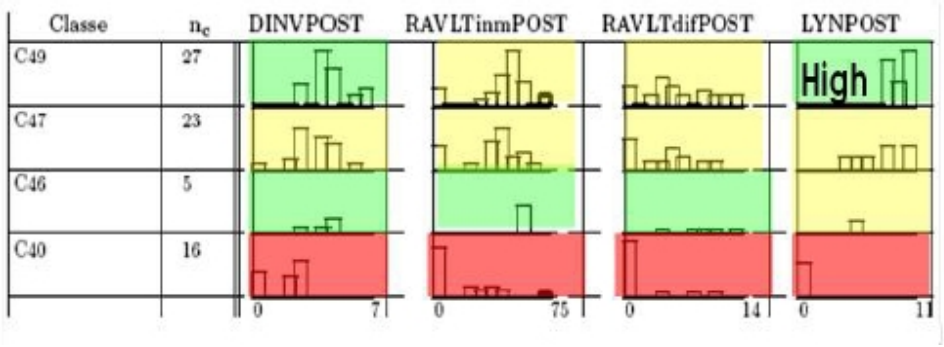


Figure 2. Traffic light panel (colour online only).

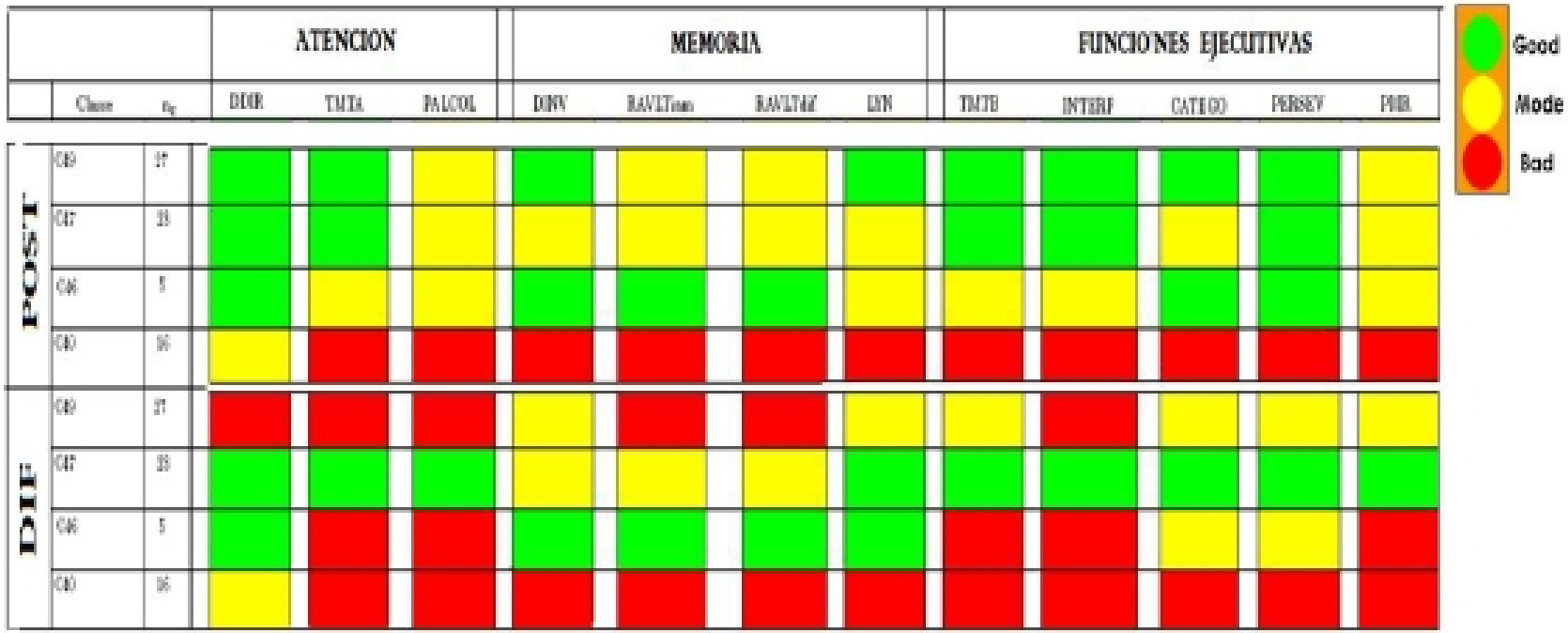
TLP → Traffic Light Panel (How to get it?)

1. Perform the Class panel graph (CPG) of all the variables versus the discovered classes
2. Calculate the basic statistics per class
3. Using materials from steps 1 and 2 identify variables or combination of variables with specific ranges of values in a class that distinguish the class from the others.
4. Assign qualitative levels to the variables implied in step 3 by detecting the area where the mass of the distribution is placed.
5. Perform a TLP for the variables, using the qualitative values assigned in step 4
6. Show the Traffic lights panel to the expert and ask him to select a label for the class. The expert is conceptualizing the class in this step, on the basis of the Traffic lights panel
7. Perform significance tests assessing relevance of differences for the variables implied in step 6, for instance: ANOVA, Kruskall-Wallis or χ^2 independence tests depending on the item.



Advanced Profiling

TLP → Improvements (Grouping variables)



UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

aTLP →
annotated Traffic Light Panel

$$\mathrm{VC}_k|C = \frac{{}^{s}X_k|C}{\bar{X}_k|C}.$$

Low VC			
Intermediate VC			
High VC			

Advanced Profiling

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

aTLP →
annotated Traffic Light Panel