A knowledge discovery and management framework for mining rare life course patterns

Emmanuel Rousseaux

Supervisors: Gilbert Ritschard and Michel Léonard

Handling life course survey data: The Dataset project

Motivation

♢ Storing efficiently survey data
♢ Specific design for longitudinal data
♢ Assist the user on the pre-processing steps for a specific analysis
♢ Exporting directly data and user manual for sharing

General specifications

♢ Advanced managment of missing values
♢ Native weights handling
♢ Consistency checks
♢ Representativeness checks
♢ User-oriented functions

Synthetic outputs in PDF

♢ "Ready to publish" formatted outcome
♢ Synthetic, easy to read
♢ For outcome of any provided analysis tool

Availability and use

♢ Released on the `R-Forge`
♢ Under active development
♢ Used by about 5-10 LIVES PhD Candidates
♢ Used in two Master courses at the Unige
♢ Two LIVES data bases are stored in this format

Function	Description
`as.valid`	Turn missing values in valid cases
`as.missing`	Turn valid cases in missing values
`bivan`	Bivariate analysis providing different measures of association (e.g. Pearson's Chi-squared, Cramer's V, Goodman and Kruskal's lambda etc.)
`checkvars`	To specify variables to use for controlling the representativeness when filtering out cases
`contains`	To search keywords in the label of all the variable stored in the database
`cut`	Turn a scale variable to an ordinal variable according to break points specified by the user
`frequencies`	Shown the frequency table or the density of a specific variable
`map`	Plot a scale variable on a map
`recode`	To recode levels of a categorical variable
`reglog`	Logistic regression method
`summaryToPDF`	Produce a summary of a Dataset object in a PDF file
`traminer.seqdef`	Create a TraMineR state sequence object [2]
`tree.chaid`	To construct a classification tree using CHAID algorithm
`tree.cart`	To construct a classification tree using CART algorithm
`weighting`	Return the name of the variable used as weight variable
`weights`	To define a variable as weight variable

Table 1: Some key methods provided by the Dataset toolbox

Representativeness checks

♢ Performed automatically
♢ Useful when interpreting results

```
weighting(shp) <- "weights"
checkvars(shp) <- c("sex", "working.status")
shp.sport <- subset(shp, sport.assoc == "Active member")

## => control on sex:  warning, p-value < 0.05
## man are oversampled
## woman are undersampled
## => control on working.status:  warning, p-value <
0.05
## active occupied are oversampled
## not in labor force are undersampled
```

Rendering spatial data

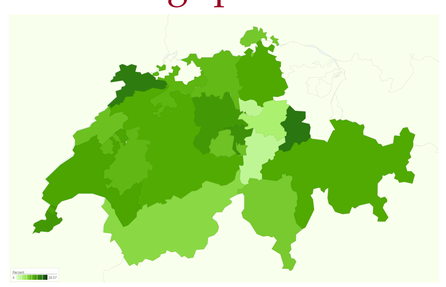

Figure 1: Poor/Good SRH ratio. SHP 2011, wave 2010 [4].

Decision-tree-based methods for the discovering of vulnerable profiles

Goal: Discovering interactions between social determinants

Entropy-based trees

♢ Example C4.5 [3], look for purity of nodes
♢ 50/50 is worst (highest uncertainty) situation.
♢ Trees are efficient tools for finding discriminating predictors

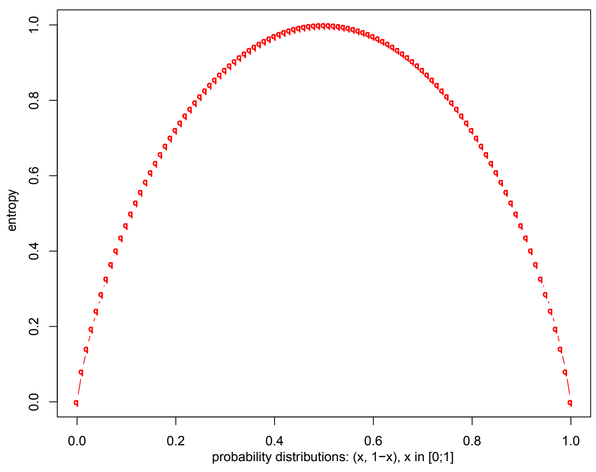

Figure 2: Shannon's entropy.

Rare classes are difficult to learn

♢ Rare classes have low recall
♢ Same majority class in most nodes

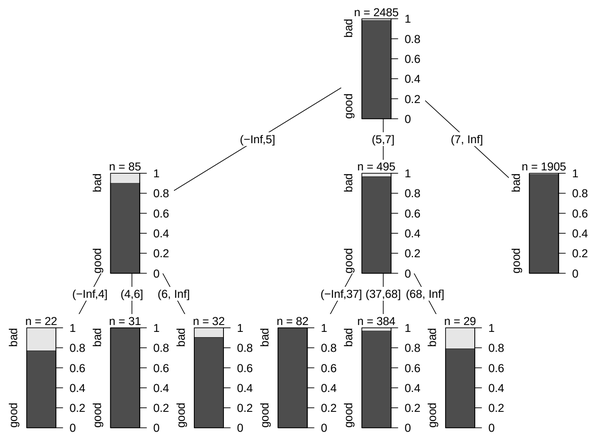

Figure 3: Entropy-based decision tree example on Poor/Good SRH. SHP wave 2010.

Accounting for the prior distribution

♢ If the overall frequency of a class is only 2%, then a node with 50/50 would be an improvement
♢ Testing for distribution differences and accounting for individual variability

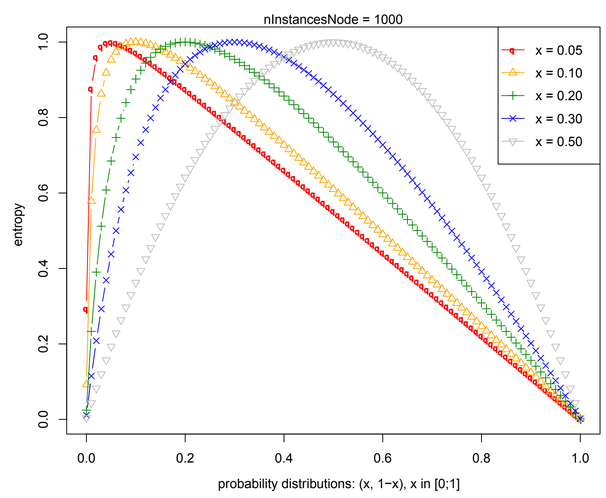

Figure 4: Zighed et al.'s asymmetric entropy [5].

Temporal multi-channel association rules mining

Goal: Discovering new hypothesis

Mining rules on several trajectories

Discovering temporal rules: $A^{t_1} \Rightarrow B^{t_2}$ [1]

♢ When A occurs at time t_1, then generally B occurs at t_2
♢ After experiencing A, there is a strong risk to fall in poor health state in the next 2 years
♢ We look on the whole life course

$$A^{t_1}_{work\ traj.} \wedge B^{t_2}_{family\ traj.} \Rightarrow C^{t_3}_{health\ traj.}$$

Interest in event Z which prevents B to occur after A

♢ Then we want to find exclusions to rules we founded
♢ $A^{t_1} \wedge Z^{t_2} \Rightarrow \bar{B}$
♢ If I experience A in time t_1 but I experience Z in time t_2, I won't experience B

$$A^{t_1}_{work\ traj.} \wedge B^{t_2}_{family\ traj.} \wedge Z^{t_3}_{family\ traj.} \Rightarrow \bar{C}_{health\ traj.}$$

Vizualisation of association rules

⇒ Plot of association rules between life course patterns
⇒ Able to account for rule redundancy

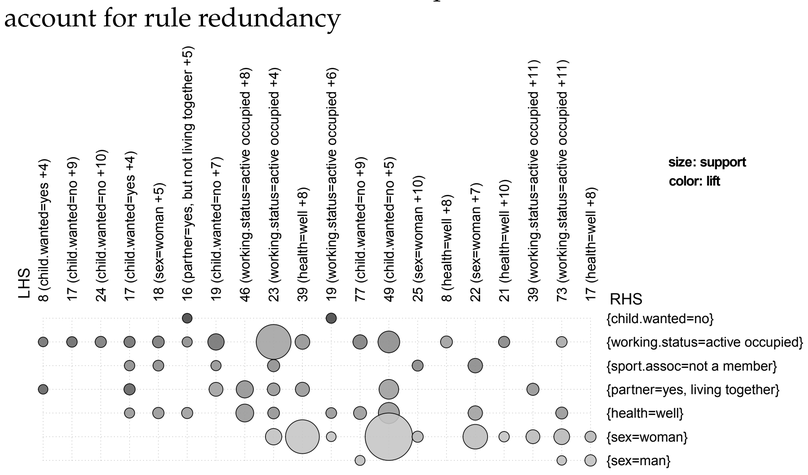

Figure 5: Grouped matrix example, sup. = 0.01, conf. = 0.6, 577 rules. SHP wave 2010.

[1] Agrawal, R. and Srikant, R. 1995. Mining sequential patterns. In *Eleventh International Conference on Data Engineering*, P. S. Yu and A. S. P. Chen, Eds. IEEE Computer Society Press, Taipei, Taiwan, 3–14.

[2] Gabadinho, A., Ritschard, G., Müller, N.S. & Studer, M. (2011), Analyzing and visualizing state sequences in R with TraMineR, *Journal of Statistical Software*. Vol. 40(4), pp. 1-37.

[3] Quinlan, J. R. (1993) C4.5: Programs for machine learning. San Mateo, CA: *Morgan Kaufmann Publishers Inc*.

[4] Voorpostel, M., Tillmann, R, Lebert, F., Weaver, B., Kuhn, U., Lipps, O., Ryser, V.-A., Schmid, F., Rothenbühler, M., & Wernli, B. (2011). *Swiss Household Panel User guide (1999-2010), Wave 12, October 2011*. Lausanne: FORS.

[5] Zighed, D.A., G. Ritschard and S. Marcellin (2010), Asymmetric and Sample Size Sensitive Entropy Measures for Supervised Learning in Ras, Z.W. and L.-S. Tsay (eds) *Advances in Intelligent Information Systems* , Series: *Studies in Computational Intelligence*, Vol. 265, Berlin: Springer. pp. 345-450.

SWISS NATIONAL SCIENCE FOUNDATION
The National Centres of Competence in Research (NCCR)
are a research instrument of the Swiss National Science Foundation

SNSF Site Visit,
University of Geneva
November 12-13, 2012

Contact: emmanuel.rousseaux@unige.ch
IP 14 – Measuring life sequences and the disorder of lives
FACULTY OF ECONOMIC AND SOCIAL SCIENCES