Decision tree methods for discovering (cross-national) interactions in longitudinal data with imbalanced data

ROUSSEAUX, Emmanuel, RITSCHARD, Gilbert


Available at:
http://archive-ouverte.unige.ch/unige:121253
Decision tree methods for discovering (cross-national) interactions in longitudinal data with imbalanced data

Emmanuel Rousseaux  
Emmanuel.Rousseaux@unige.ch

Gilbert Ritschard  
Gilbert.Ritschard@unige.ch

NCCR LIVES – Overcoming vulnerability: life course perspectives  
Institute of Demography and Socioeconomics  
University of Geneva  
Switzerland
We aim at studying similarity/dissimilarity across countries

- Common strategy: finding interactions in covariates
  - Interaction effect $\Rightarrow$ dissimilarity
  - No interaction effect $\Rightarrow$ similarity
- Decision tree models: discovering interactions
- Regression models: confirming interactions
Introduction
Entropy-based decision tree modeling
Entropy measures for imbalanced data
The R package trim
Future work

Motivation
Decision trees: an example
Methodological issue

Figure: Decision tree analysis from Guarin and Rousseaux (2015). Method CHAID.
Legend: $E$ = employed, $U$ = unemployed, $REL3$ = respondent educational level,
$FEL3$ = father educational level.
Methodological issue: The unfrequent response class problem

- Generally appears when:
  - Studying critical life situations (e.g. long-term unemployment or critical health problems)
  - Working with longitudinal data (e.g. data in person-period format)

- Consequences
  - Can prevent to discover relevant underlying interactions in the data (when using a statistical criterion)
  - Can lead to a poor prediction performance of the rare event (e.g. for entropy-based decision trees)
Growing decision trees: A recursive partitionning process

- Starting point (1st node): The entire population
- For every covariate: Assessment of the *best* split
- We select the *best* split of the *best* covariate.
- Is a stopping criterion satisfied? If yes, we stop.
- If not, we split and repeat the process in every child nodes
Figure: Binary Shannon entropy (Shannon, 1948).
Selected literature review: (Zighed, Ritschard, and Marcellin, 2010)

Figure: Asymmetric entropy for some levels $w_m$ of data imbalance, $N = 1000$. 
Selected literature review: (Lallich, Lenca, and Vaillant, 2007)

Figure: Off-centered entropy for some levels $w_m$ of data imbalance.
The off-centering residual effect

Figure: $h(0.1, 0.9) < h(0.3, 0.7)$. 
Note: the choice of the best split is made by looking at the entropy reduction.

⇒ New strategy: we will design the measure to ensure that reducing the value of the entropy improve the visibility of the rare class.
First approach: the $w$-normalized entropy.

We apply the following change in variable $p \mapsto \frac{1}{|p/w|} p/w$.

The full formula is:

$$h_n(p_1, \ldots, p_\ell; w_1, \ldots, w_\ell) = h \left( \frac{p_1}{w_1 \sum_j \frac{p_j}{w_j}}, \ldots, \frac{p_\ell}{w_\ell \sum_j \frac{p_j}{w_j}} \right)$$

with

- $p = (p_1, \ldots, p_\ell)$: the distribution in a node.
- $w = (w_1, \ldots, w_\ell)$: the distribution to avoid.
- $h$: a classical entropy measure.
Figure: $w$-normalized entropy for some levels $\omega_m$ of data imbalance.
Second approach: the rco-score.

We ensure that a same move from the \( w \) distribution will not favour the majority classe more than the rare class.

The full formula is:

\[
h_{\text{rco}}(p_1, \ldots, p_\ell; w_1, \ldots, w_\ell) = h\left(\frac{p_1 + \bar{w}_1}{\ell + 1}, \ldots, \frac{p_\ell + \bar{w}_\ell}{\ell + 1}\right)
\]

with

\[
\bar{w} = (\bar{w}_1, \ldots, \bar{w}_\ell) = (1 + (\frac{1}{\ell} - w_1), \ldots, 1 + (\frac{1}{\ell} - w_\ell))
\]
Figure: rco-score for some levels $w_m$ of data imbalance.
The trim package

https://r-forge.r-project.org/projects/trim/

- Make available these methods especially designed for working with imbalanced data
- Data preparation: designed to be combined with the Rsocialdata package (Rousseaux, Bolano, and Ritschard, 2013)
- Take weights into account
- Exploring interactions: interactive plot
- Automatically confirm interactions with regression modeling
The Rsocialdata project – Key points

- Storing survey data in R
  (missing values, weighting, etc.)
- Documenting data:
  questions, values, variables, and surveys
- Saving time when preparing data to focus quickly on analyses
- Facilitating the share of survey data
Future work

- Comparative study of the entropy measures presented
- Release of the trim package on the CRAN