Donor selection by multiple Tree-Based Models for Imputation

Jorge Aramendi (j-aramendi@eustat.eus), Javier San Vicente (j-sanvicente@eustat.eus)

Keywords: Imputation, Tree-Based Models, R&D Activities Survey

1. INTRODUCTION

Missing data is a pervasive problem in official statistics. In the research that motivates this paper, we were required to produce a workable method to impute records in the Research and Development Activities Survey, conducted by the Basque Statistical Institute (EUSTAT) **[1]**. This is an annual survey answered by well over a thousand entities (private companies, research institutes, hospitals, universities...) which conduct research activities.

Although not widespread, non response exists: for the period 2009-2013, almost 11% of the questionnaires were not returned and could not be directly obtained, even though there was certainty that the entity surveyed was engaged in R&D activities. In some particular activity sectors, non response was considerably higher. This poses a considerable problem for a survey which is meant to provide figures disaggregated by sectors.

Unlike ordinary production activities, expenditure in R&D activities is very erratic: investment can boom one year as a new facility is created and endowed, or a lab is retooled, then taper off. There is not an stable relationship of R&D expenditure and personnel to firm size or turnover.

Another feature of the survey is that many of the variables investigated, both qualitative and quantitative, are related by constraints which would be hard, if at all feasible, to embed in a model. This project is a cooperative work with researchers of the University of the Basque Country (EHU-UPV).

2. METHODS

Given the facts sketched above, a donor-based imputation method was thought the best choice. The main advantage is that imputation is performed *en bloc* by copying (after suitable re-scaling) all of the imputable fields from the donor to the imputed record; the fact that these fields belong to an observed donor guarantee their internal consistency.

The choice of donor-based imputation requires a method to choose donors which in turn requires a way of specifying "likeness" between observations. This is hard with multivariate observations of mixed type (qualitative and quantitative). Alternatives such as Gower distance are purely additive in their components, a feature which does not quite reflect the constraints and interactions between variables in our problem.

What has been done instead is inspired in work on weak learners in the last twenty years. For each of the response variables we want to impute, we train one (regression or classification) tree using the fully observed cases and predictors restricted to variables that are observed in all cases, even for the non respondents. This includes directory information, such as location, type of business, size or sector of activity. Then, to impute a missing observation,

1 We drop each case to impute down each of the trees and note the leave where it ends. Each of the fully observed cases ending in the same leave are "voted" as

potential donors, on the ground that they are similar to the case to impute as far as the tree can ascertain.

2 The sum of votes is taken as an indicator of "likeness" or similarity, and the fully observed case with the largest value is used as a donor.

It should be stressed that the similarity is not computed on the basis of values of the responses observed for recipient and potential donors, but rather on the reconstruction that can be made of the same using the available predictors: this is what makes the imputation feasible even for entities that have never before been observed.

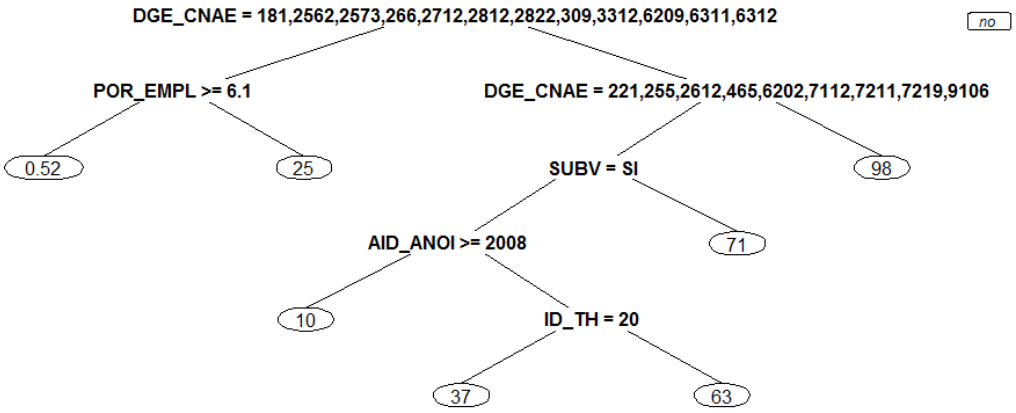

Figure 1. Example of one of the trees

We can (and do) weight the voting power of each tree, since some response variables are of higher importance than others. We also use predictors in several time periods, which adds a temporal dimension to the similarity measure. In all, 57 trees (one per response variable) are trained, each contributing from 0.5 to 3 "votes" to the similarity measure between an unobserved case and each potential donor. The most similar case in the pool of potential donors is used in the imputation, all the imputable fields being copied from donor to recipient, re-scaling to account for their possibly different sizes.

3. RESULTS

Results have been satisfactory. The method affords easy a fast imputation in an otherwise rather intractable problem. With the weights chosen, the algorithm picks frequently, but not always, a donor in the same sector than the recipient. Beyond that, the choice takes into account the 57 response variables in a way that would be difficult for a human imputer to account for. An R package, idimp, specific to the problem but easy to generalise, has been written which performs all necessary computations. It also provides ancillary functions which track and explain the computation of similarities and may help in refining the weights given to each tree. This package uses rpart[2] and partykit[3] R packages.

4. CONCLUSIONS

A procedure has been described and implemented providing a method to impute a mixture of qualitative and quantitative variables, restricted by many constraints. The method uses coincidence in the leaves of each of the trees as a criterion for proximity in a given variable: adding the weighted votes of all trees gives a global measure of similarity.

There are some questions which require further research. One of them is to what extent it is desirable to let the trees grow. We have taken the choice of growing very large trees, subject only to the constraint of at least five observations in each leave. This makes each tree a weak learner: the method relies on the fact that there are a large number of them. It is clear that tree size should be chosen in some way dependent on the number of trees.

Another option which we presently investigate is that of fractional voting. At present, cases sharing a leave with the recipient are all given one (weighted) vote, and those not in the same leave zero votes. This raises the question of near matches: cases that share most of the same branch than the recipient but end up in a different leave. We consider the option of giving partial votes to such cases, at the expense of a far more complex software implementation, but with the benefit of making the method less sensitive to tree size.

References

[1] EUSTAT. Research and Development Activities Survey , data and documents.

http://en.eustat.eus/estadisticas/tema_426/opt_0/ti_Scientific_research_and_technological_development_RD_in_biotechnology/temas.html

[2] Terry Therneau, Beth Atkinson and Brian Ripley (2015). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10. https://CRAN.R-project.org/package=rpart

[3] Torsten Hothorn, Achim Zeileis (2015). partykit: A Modular Toolkit for Recursive Partytioning in R. Journal of Machine Learning Research, 16, 3905-3909. URL http://jmlr.org/papers/v16/hothorn15a.html