International Statistical Review

International Statistical Review (2014), 82, 3, 349–352 doi:10.1111/insr.12059

Discussions

Carolin Strobl

Universität Zürich, Zurich, Switzerland
E-mail: carolin.strobl@psychologie.uzh.ch

With 'Fifty Years of Classification and Regression Trees', Wei-Yin Loh has given a concise historical overview of the central developments in recursive partitioning. It is interesting to read how the successive improvements were triggered by their predecessor algorithms—especially as the author has (co-)authored many of the milestones in our field. Moreover, I appreciate his emphasis on open-source implementations, which make the methodology available to scientists from all disciplines and all around the world.

I have two questions to the author and would like to add a short comment about variable importance measures.

1 Missing Value Handling

As has been pointed out for several algorithms in the paper, the treatment of missing values is an interesting aspect that distinguishes recursive partitioning techniques from other statistical methods. The two approaches specific for recursive partitioning are (i) surrogate variables, which are correlated with the primary splitting variable and can thus be used to replace it for processing observations with missing values in the primary variable, and (ii) creation of separate nodes for missing values.

Both approaches are distinct from ad hoc approaches classically—while often unreflectedly—used in statistical analyses, such as case-wise deletion, but also from more advanced approaches like (multiple) imputation techniques. While the advantage of preserving observations with missing values and thus avoiding data loss is straightforward, I wonder how these approaches relate to the concepts of missing completely at random (MCAR), missing at random (MAR) and missing not at random (cf., e.g. Little & Rubin, 1986).

Hapfelmeier *et al.* (2014) systematically investigate the effects of MCAR, MAR and missing not at random on a random forest variable importance measure modified to be able to deal with missing values (including surrogate variables), but I am not aware of any such studies for the approach of creating separate nodes for missing values.

Do you know of any or can you infer how this approach performs under the different missingness mechanisms? In particular, is creating a separate node for missing values informative for predicting the response variable under the MCAR and MAR schemes, and/or could it be informative for narrowing down the missingness mechanism itself?

2 Ignorance of Variable Selection Bias

Because both Wei-Yin Loh and I have worked in this area, I assume we both find the development of unbiased split selection criteria to be one of the most important improvements over the early recursive partitioning algorithms. As pointed out in the paper, variable selection

© 2014 The Authors. International Statistical Review © 2014 International Statistical Institute. Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

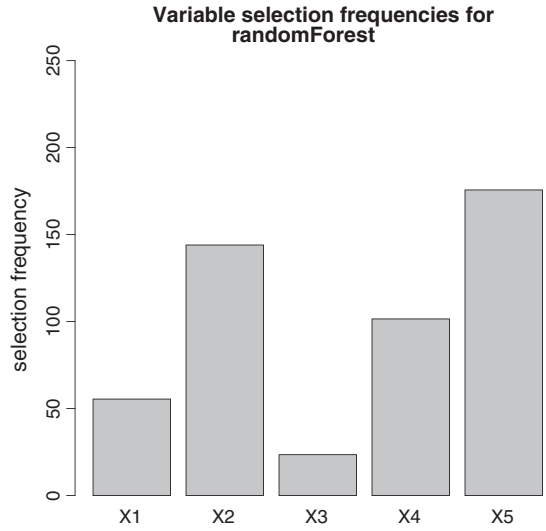

Figure 1. *Variable selection frequencies for a random forest algorithm with biased split selection (*`randomForest`*, Liaw & Wiener, 2002). In the underlying simulation design, only the second predictor variable X_2 is informative, but its selection frequency is outperformed by the irrelevant variable X_5, which is preferred only because it has more categories.*

bias is defined as an artificial preference for variables offering more cutpoints—even if all variables are noise variables containing no information.

What we should probably point out more clearly though is that this also means that when a predictor variable offering few cutpoints is in fact associated with the response—and thus should be found relevant by any reasonable statistical learning technique—it may still be outperformed by a less informative or even irrelevant competitor, just because the latter offers more cutpoints. This is illustrated in Figure 1 from Strobl *et al.* (2007) (for the selection frequencies of a random forest, but of course, the same can be observed for the selection frequencies of single trees).

In the simulation design underlying Figure 1, the predictor variables systematically differ in the number of cutpoints they offer: X_1 was generated from a normal distribution (thus offering a high number of different cutpoints), X_2 from a binomial distribution (offering only one cutpoint) and X_3 to X_5 from a multinomial distribution with 4, 10 and 20 categories, respectively (offering again an increasing number of cutpoints). Only X_2 was simulated to have a strong effect on the response class, whereas X_1, X_3, X_4 and X_5 are entirely uninformative noise variables.

Yet, we can clearly see in Figure 1 that the relevant variable X_2 is outperformed by the irrelevant noise variable X_5, which is preferred solely because it has more categories. (If the effect size of X_2 is modelled to be more moderate, it is also outperformed by noise variables with less categories.)

One should think that the results shown here, and in many previous studies that Wei-Yin Loh has summarized in his paper, are so clear that any statistically educated person should never want to use a biased recursive partitioning algorithm again. Yet I encounter so many cases where biased recursive partitioning algorithms are still employed in both applied and methodological publications—including some of those cited in 'Fifty Years of Classification and Regression Trees'.

I really wonder why this is the case. Does it mean that the authors of those publications do not consider variable selection bias an issue of concern or willingly ignore decades of

© 2014 The Authors. International Statistical Review © 2014 International Statistical Institute

research? Or rather that we have not managed to bring our results to the attention of a broader scientific audience? Or maybe even that—at first sight—recursive partitioning looks so easy that anyone can do it without bothering to read up on it? I would be very interested to hear your opinion.

3 Variable Importance Measures

When giving up the interpretability of single trees for the stability of ensemble methods, variable importance measures are the only means to tease out at least some information from the otherwise black box. As Wei-Yin Loh has pointed out, these variable importance measures only provide a summarized impression and cannot be interpreted with respect to the direction or actual form of the relationship between predictor variables and response. Still, they can serve as a valuable tool in application areas where exploratory screening (cf., e.g. Lunetta *et al.*, 2004; Bureau *et al.*, 2005) is the only way to narrow down the number of candidate variables that need to be considered in more detail.

Even though many disciplines with a strong tradition in hypothesis-driven research, such as psychology, are still somewhat shy about these types of procedures, they have their right to exist as one legitimate means of generating hypotheses when no other means is available, as pointed out by Strobl (2013). What is crucial to note, however, is that when machine learning or other statistical techniques are used for screening or automated variable selection (cf., e.g. Diaz-Uriarte & de Andrés, 2006; Rodenburg *et al.*, 2008, for random-forest-based approaches), a newly drawn sample must be used to later conduct statistical significance tests on the selected variables. In some cases, it might even be possible to experimentally test their effects (e.g. by 'knocking out' a previously identified candidate gene).

To conclude with the issue of variable importance measures, let me add a short specification of the works of Strobl *et al.* (2007, 2008).

In Strobl *et al.* (2007), it is shown that—unsurprisingly—random forests built from trees with biased split selection criteria also show variable selection bias (as was illustrated in Figure 1) and that this bias also transfers to the Gini and permutation variable importance measures. However, what was very surprising to us was that even when random forests are built from trees with unbiased split selection criteria, like in the `cforest` function available in the R-package `party`, the widely used bootstrap sampling induces another source of bias, which again affects the variable selection frequencies but also results in an increased variance for the permutation variable importance. This is the reason that we discourage the use of bootstrap sampling and employ subsampling as the default in `cforest`.

Strobl *et al.* (2008), on the other hand, consider the consequences of correlations between predictor variables, which had previously been noted by Archer & Kimes (2008) and Nicodemus & Shugart (2007). In this situation, it is not *per se* clear how a good variable importance measure should behave, and even for parametric models like multiple linear regression, a variety of variable importance measures have been suggested (cf., e.g. Azen & Budescu, 2003), which vary in their particular treatment of correlated variables.

It is a matter of taste or philosophy—rather than an objectively defined bias in the statistical sense—how a variable importance measure should behave in the presence of correlated variables. My impression from speaking to applied researchers was, however, that they were interpreting the random forest permutation importance similar to the coefficients of a multiple regression model, which reflect the impact of a variable given all other variables in the model—which is not how Breiman's original permutation variable importance works. Therefore, we developed a conditional permutation scheme available for `cforest`, which more closely mimics the behaviour of multiple regression coefficients (that is, however, computationally only feasible if the number of correlated variables is not too high).

© 2014 The Authors. International Statistical Review © 2014 International Statistical Institute

References

Archer, K.J. & Kimes, R.V. (2008). Empirical characterization of random forest variable importance measures. *Comput. Stat. Data Anal.*, **52**(4), 2249–2260.

Azen, R. & Budescu, D.V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychol. Methods*, **8**(2), 129–48.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P. & Eerdewegh, P.V. (2005). Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.*, **28**(2), 171–182.

Diaz-Uriarte, R. & de Andrés, S.A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**(3).

Hapfelmeier, A., Hothorn, T., Ulm, K. & Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Stat. Comput.*, **24**(1), 21–34.

Liaw, A. & Wiener, M. (2002). Classification and regression by `randomForest`. *R News*, **2**(3), 18–22.

Little, R. & Rubin, D. (1986). *Statistical Analysis with Missing Data.* New York: John Wiley & Sons, Inc.

Lunetta, K.L., Hayward, L.B., Segal, J. & Eerdewegh, P.V. (2004). Screening large-scale association study data: Exploiting interactions using random forests. *BMC Genetics*, **5**(32).

Nicodemus, K. & Shugart, Y.Y. (2007). Impact of linkage disequilibrium and effect size on the ability of machine learning methods to detect epistasis in case–control studies. In *Proceedings of the Sixteenth Annual Meeting of the International Genetic Epidemiology Society*, Vol. 31, North Yorkshire, UK, pp. 611.

Rodenburg, W., Heidema, A.G., Boer, J.M., Bovee-Oudenhoven, I.M., Feskens, E.J., Mariman, E.C. & Keijer, J. (2008). A framework to identify physiological responses in microarray based gene expression studies: selection and interpretation of biologically relevant genes. *Genet. Epidemiol.*, **33**(1), 78–90.

Strobl, C. (2013). Data mining. In *The Oxford Handbook on Quantitative Methods*, Ed. T. Little, pp. 678–700. USA, Chapter 29: Oxford University Press.

[*Received March 2014, accepted March 2014*]

International Statistical Review (2014), 82, 3, 352–357 doi:10.1111/insr.12065

Antonio Ciampi

Department of Epidemiology and Biostatistics, McGill University,1020 Pine Ave. West, Montreal H3A 1A2, Quebec, Canada
E-mail: antonio.ciampi@mcgill.ca

Fifty years ago, Morgan & Sonquist (1963) introduced the now famous AID algorithm. AID uses a sample to construct a tree-structured predictor for a specified continuous variable y given a specified vector of covariates x. The result is therefore, in contemporary language, a *regression* tree. The first paper to introduce *classification* trees in the modern sense appeared 9 years later, in 1972: it presented the algorithm THAID, as mentioned in the review (REF). We have to wait 12 more years to see both kinds of trees reunited in the very influential book by Breiman *et al*. (1984), 'Classification and Regression Trees'. This title is often abbreviated as CART, which is somewhat confusing, because the acronym CART™ also denotes the proprietary software associated to the book. Notwithstanding the great merits of the CART book, it is more than fair to consider the 1963 paper as the seminal work for the *research area* known as 'Classification and Regression Trees', the object of Prof. Loh's excellent review. One could argue that the 1963 paper, together with the CART book, is also at the origin of flexible statistical modelling (beyond variable selection algorithms in regression) such as MARS (Friedman, 1991) and PIMPLE (Breiman, 1991), and, indeed, of *statistical* machine learning (Hastie *et al.*, 2009) . Yet one cannot disagree with the author's decision of restricting the review to

© 2014 The Authors. International Statistical Review © 2014 International Statistical Institute