Computational Statistics and Data Analysis 80 (2014) 129–139

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Variable selection by Random Forests using data with missing values☆

A. Hapfelmeier*, K. Ulm

Institut für Medizinische Statistik und Epidemiologie, Technische Universität München, Ismaninger Str. 22, 81675 München, Germany

ARTICLE INFO

Article history:
Received 28 January 2013
Received in revised form 18 June 2014
Accepted 19 June 2014
Available online 1 July 2014

Keywords:
Random Forests
Variable importance
Variable selection
Missing data
Multiple imputation
Complete case analysis

ABSTRACT

Variable selection has been suggested for Random Forests to improve data prediction and interpretation. However, the basic element, i.e. variable importance measures, cannot be computed straightforward when there are missing values in the predictor variables. Possible solutions are multiple imputation, complete case analysis and the use of a self-contained importance measure that is able to deal with missing values. Simulation and application studies have been conducted to investigate the properties of these procedures when combined with two popular variable selection methods. Findings and recommendations: Complete case analysis should not be used as it led to inaccurate variable selection. Multiple imputation is the method of choice if the selection of a variable is supposed to reflect its potential relevance in a complete data setting. However, Random Forests are commonly used without any preprocessing of the data as they are known to implicitly deal with missing values. In such a case, the application of the self-contained importance measure permits the selection of variables that are of relevance in these actual prediction models.

1. Introduction

In many research fields, Random Forests (Breiman, 2001; Breiman and Cutler, 2008) are appreciated for their ability to implicitly deal with missing values, high dimensional data and complex relations among variables. Moreover, they provide variable importance measures that rate a variable's relevance for prediction. These measures are also used for variable selection (Altmann et al., 2010; Archer and Kimes, 2008; Díaz-Uriarte and Alvarez de Andrés, 2006; Genuer et al., 2010; Hapfelmeier and Ulm, 2013; Jiang et al., 2004; Tang et al., 2009; Yang and Gu, 2009; Rodenburg et al., 2008; Sandri and Zuccolotto, 2006; Schwarz et al., 2007; Svetnik et al., 2004) meant to distinguish relevant from non-relevant variables. The benefit for prediction though is still ambiguous (Yang and Gu, 2009; Zhou et al., 2010; Altmann et al., 2010; Díaz-Uriarte and Alvarez de Andrés, 2006; Svetnik et al., 2004).

A major issue which is in main focus of this work is how to perform variable selection when there are missing values in the predictor variables. Existing methods are based on importance measures that cannot be computed straightforward in such a case. Possible solutions are investigated in this work: Complete case analysis as a fast and easy way to deal with missing values (though it is known to lead to biased inference when values are not missing completely at random, cf. Schafer and Graham, 2002; Horton and Kleinman, 2007). Multiple imputation by chained equations (MICE; van Buuren et al., 2006; White et al., 2011) which enables the simultaneous imputation of multiple variables with missing values. And a self-contained importance measure (Hapfelmeier et al., 2014b) which is able to handle missing values without the need to omit or replace them before analysis. Two representative variable selection methods are used in combination with these

☆ R-Code for the simulation study is given as supplementary material (see Appendix B).

* Corresponding author. Tel.: +49 89 4140 4340; fax: +49 89 4140 4850.
E-mail address: Alexander.Hapfelmeier@tum.de (A. Hapfelmeier).

http://dx.doi.org/10.1016/j.csda.2014.06.017
0167-9473/

approaches. A simulation and an application study is performed to investigate their properties for variable selection and prediction (cf. Hapfelmeier et al., 2014a, for similar studies about the computation of importance measures).

2. Random Forests

Random Forests are made up by several single trees (e.g., using the CART or C4.5 algorithm; Breiman et al., 1984; Quinlan, 1993). The latter are fit to bootstrapped data while splits are found in random sets of variables. Predictions are given by averaged values or majority votes of each tree's prediction. The so called 'out of bag' (OOB) samples, i.e. observations not used to fit the respective trees, can be used for an unbiased estimate of the prediction error, viz. the 'OOB-error'. However, when there are missing values, surrogate splits need to be employed. They mimic the initial primary splits but provide decision rules based on other variables. In the following, Random Forests that consist of conditional inference trees (Hothorn et al., 2006) will be used. This algorithm guarantees unbiased variable selection (cf. Strobl et al., 2007a; White and Liu, 1994; Kim and Loh, 2001; Dobra and Gehrke, 2001; Kung et al., 2012; Lee and Shih, 2006; Shih and Tsai, 2004, for corresponding discussions) and variable importance measures when combined with subsampling (as opposed to bootstrap sampling; cf. Strobl et al., 2007b,c).

2.1. An importance measure for variables with missing values

The permutation importance measure is a popular means to assess a variable's relevance for prediction in a Random Forest. It is computed by the difference of a tree's OOB-error before and after random permutation of a predictor variable. If the latter is related to the outcome and further predictors, and if it is of relevance for prediction in the forest, the error is supposed to rise by permutation. The mean value of all trees is the actual importance of a variable.

There are two versions of this measure. The original importance measure (as described above) and the conditional importance measure introduced by Strobl et al. (2008). These assess the marginal and conditional relation of a variable to the outcome, respectively. The former will therefore take higher values for variables that have the same conditional relation to the outcome as others, but are (cor-)related to further informative variables. Both measures can be of specific value depending on the research question (Nicodemus et al., 2010; Altmann et al., 2010). In this work the original version is preferred for its sensitivity to relations between variables which are supposed to be uncovered.

A general limitation is that the permutation importance measure cannot be computed straightforward when there are missing values. It is unclear how to appropriately handle surrogate splits, and the corresponding surrogate variables, that contribute to the computation of the OOB-error but are not part of the permutation scheme. A solution to this problem was introduced earlier (Hapfelmeier et al., 2014b). It is closely related to the existing methodology, and therefore retains appreciated properties. Yet it differs in one substantial aspect: the null hypothesis that a variable X is not related to the outcome and other predictors is not simulated by permutation any more. Alternatively, observations are randomly sent to the daughter nodes when a parent node k is split in X (Ishwaran, 2007, suggest a similar method to avoid permutation). In doing so, the respective probability, e.g., that an observation is sent the left way, is given by the relative frequency $\hat{p}_k$ of observations that initially went in the same direction. The algorithm to compute this importance measure is:

1. Compute the OOB-error of a tree.
2. Randomly assign each observation with $\hat{p}_k$ to the child nodes if the parent node k is split in X.
3. Recompute the OOB-error of the tree (following step 2).
4. Compute the difference between the original and recomputed OOB-errors.
5. Repeat steps 1–4 for each tree and use the average difference over all trees as the overall importance score.

This procedure circumvents the necessity to directly process missing values and solves any problems associated with permutation. In case of complete data it produces similar values to the original approach (Hapfelmeier et al., 2014b). Also, Random Forests are usually appreciated for their ability to implicitly deal with missing values in the predictor variables. This has not been true for the computation of importance measures, though. The self-contained method described above closes this gap and permits applicants to gain more insight in their prediction models. For all these reasons it will be used in the following simulation studies.

2.2. Variable selection

A very general suggestion for variable selection is to investigate the performance of prediction models that are fit to different sets of variables and to pick the best one (Guyon and Elisseeff, 2003). Such 'performance-based', wrapper methods have also been proposed for Random Forests (Díaz-Uriarte and Alvarez de Andrés, 2006; Genuer et al., 2010; Jiang et al., 2004; Svetnik et al., 2004). A very popular representative, which will be used in the following analyses, is the one of Díaz-Uriarte and Alvarez de Andrés (2006): the algorithm computes the importances of all variables in an initial step. In subsequent steps, the least important variables are sequentially rejected and the OOB-errors of corresponding Random Forests are recorded. The final model is chosen to be the one that uses the smallest variable set to produce an error within a range of u standard errors to the best performing model ($u = 1$ equals the 'one-standard-error' rule ('1 s.e.' rule); Breiman et al., 1984; Hastie et al., 2009).