Computational Statistics and Data Analysis 96 (2016) 57–73

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Random forest for ordinal responses: Prediction and variable selection

Silke Janitza [a,*], Gerhard Tutz [b], Anne-Laure Boulesteix [a]

[a] Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, D-81377 Munich, Germany
[b] Department of Statistics, University of Munich, Akademiestr. 1, D-80799 Munich, Germany

ARTICLE INFO

Article history:
Received 1 December 2014
Received in revised form 9 October 2015
Accepted 11 October 2015
Available online 19 October 2015

Keywords:
Random forest
Ordinal regression trees
Ordinal response
Prediction
Feature selection
Variable importance

ABSTRACT

The random forest method is a commonly used tool for classification with high-dimensional data that is able to rank candidate predictors through its inbuilt variable importance measures. It can be applied to various kinds of regression problems including nominal, metric and survival response variables. While classification and regression problems using random forest methodology have been extensively investigated in the past, in the case of ordinal response there is no standard procedure. Extensive studies using random forest based on conditional inference trees are conducted to explore whether incorporating the ordering information yields any improvement in both prediction performance or variable selection. Two novel permutation variable importance measures are presented that are reasonable alternatives to the currently implemented importance measure which was developed for nominal response and makes no use of the ordering in the levels of an ordinal response variable. Results based on simulated and real data suggest that predictor rankings can be improved in some settings by using new permutation importance measures that explicitly use the ordering in the response levels in combination with ordinal regression trees. With respect to prediction accuracy, the performance of ordinal regression trees was similar to and in most settings even slightly better than that of classification trees.

1. Introduction

In many applications where the aim is to predict the response or to identify important predictors, the response has an inherent ordering. Examples of ordinal responses in biomedical applications are tumor stages I–IV, disease severity, for example from mild to moderate to severe disease state, and artificially created scores combining several single measurements into one summary measure, like the Apgar score, which is used to assess the health of a newborn child. Statistical models for ordinal responses such as the proportional odds, the continuation ratio and the adjacent category model have been investigated extensively in the literature (see Agresti, 2002). However, these methods are not suitable for applications where the association between predictors and the response is of a complex nature, including higher-order interactions and correlations between predictors. Moreover, the models rely on assumptions (such as proportional odds) that are frequently not realistic in practical applications. Further, parameter estimation typically faces the problem of numerical instability if the number of predictors is high compared to the number of observations.

* Corresponding author. Tel.: +49 89 440077755.
E-mail address: janitza@ibe.med.uni-muenchen.de (S. Janitza).

http://dx.doi.org/10.1016/j.csda.2015.10.005
0167-9473/

The random forest (RF) method by Breiman (2001) is a commonly used tool in bioinformatics and related fields for classification and regression purposes as well as for ranking candidate predictors (see Boulesteix et al., 2012b, for a recent overview). It has been used in many applications involving high-dimensional data. As a nonparametric method, RF can deal with nonlinearity, interactions, correlated predictors and heterogeneity, which makes it especially attractive in genetic epidemiology (Briggs et al., 2010; Chang et al., 2008; Liu et al., 2011; Nicodemus et al., 2010; Sun et al., 2007). The RF method can be applied for classification (in the case of a nominal response) as well as for regression tasks (in the case of a numeric response). By using an ensemble of classification or regression trees, respectively, one can obtain predictions and identify predictors that are associated with the response via RF's inbuilt variable importance measures (VIMs).

For nominal and numeric response the application of RF has been well investigated. However, in the case of ordinal response there is no standard procedure. While in the classical RF algorithm by Breiman (2001) the ordering of a predictor is taken into account by allowing splits only between adjacent categories, the ordering information in the response is ignored (i.e., the response is treated as a nominal variable), and an ensemble of classification trees is constructed. However, ignoring the ordering information results in a loss of information which might lead to less accurate predictions. Except for the study of Archer and Mas (2009), approaches for ordinal regression problems have only been discussed for CART (e.g., Piccarreta, 2001) and we are not aware of any study or implementation which extends these approaches to RF.

The unbiased RF version of Hothorn et al. (2006b) is based on a unified framework for conditional inference and, in contrast to the classical RF version of Breiman (2001), in which certain types of variables are favored for a split (Strobl et al., 2007; Nicodemus, 2011; Boulesteix et al., 2012a; Nicodemus and Malley, 2009), it provides unbiased variable selection when searching for an optimal split (Strobl et al., 2007; Hothorn et al., 2006b). This RF version is a promising tool for constructing trees with ordinal response because, in contrast to the standard RF implementation by Breiman (2001), where splitting is based on the Gini index, it provides the possibility of taking the ordering information into account when constructing a tree, which may possibly yield improved predictions.

A further issue which is investigated in this paper is the appropriate handling of the ordering information in the response when computing the importance of variables by using a VIM. The importance for each predictor is derived from the difference in prediction accuracies of the single trees resulting from the random permutation of this predictor. An appropriate prediction performance measure is essential for a good performance of the VIM, as demonstrated by Janitza et al. (2013).

The design of an appropriate VIM in the common case of ordinal response variables, however, has to our knowledge never been addressed in the literature. The currently used VIM based on the error rate as a prediction accuracy measure does not seem suitable in the case of an ordinal response because the error rate does not differentiate between different kinds of misclassification.

In this paper we investigate whether incorporating the ordering information contained in the response improves RF's prediction accuracy and predictor ranking through RF. To improve predictor ranking for ordinal responses, we investigate the use of three alternative permutation VIMs which are based on the mean squared error, the mean absolute error and the ranked probability score, respectively, that all take the ordering information into account. While the VIM based on the mean squared error is an established VIM that is frequently used for RF in the context of regression problems, the latter two VIMs are novel and have not been considered elsewhere. Finally we explore the impact of the choice of scores on prediction accuracy and on predictor rankings.

This article is structured as follows. In Section 2 we introduce the methods. The first part of the methods section reviews established performance measures that can be used to assess the ability of a classifier to predict an ordinal response. The second part gives an introduction to tree construction and prediction by RF based on conditional inference trees. Thereafter we outline the concept of variable importance and introduce the two existing VIMs as well as our two novel VIMs that we propose for predictor rankings through RF and ordinal response data. In Sections 3 and 4 we present our studies on simulated and real data, respectively. In both sections we report on the studies of prediction accuracy first. Here we compare the prediction accuracy of a RF constructed from classification trees with that of a RF constructed from ordinal regression trees. Subsequently we show the studies on the performance of VIMs. Here we compare the performance of the standard error rate based VIM to those of the three alternative permutation VIMs when computed on classification and ordinal regression trees. In Section 5 we summarize our findings and give recommendations to applied researchers working with RF and ordinal response data.

2. Methods

2.1. Performance measures

In the following we give definitions of established performance measures that are used in our studies for two purposes: (i) to evaluate the prediction accuracy of RF for predicting an ordinal response and (ii) for use in the proposed alternative permutation VIMs.

The error rate (ER) for the classification of observations $i = 1, \ldots, n$ with true classes Y_i into predicted classes $\hat{Y}_i$ is given by

$$ER = \frac{1}{n} \sum_{i=1}^{n} I(\hat{Y}_i \neq Y_i), \tag{1}$$