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

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

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

where $I(\cdot)$ denotes the indicator function. The error rate does not take the ordering of the classes into account since it only distinguishes between a correct classification ($\hat{Y} = Y$) and an incorrect classification ($\hat{Y} \neq Y$).

With the mean squared error (MSE) not all misclassifications are regarded as equally bad as is the case for the error rate. A higher penalty is put on a classification into a class which is more distant from the true class Y than on a classification into a class which is closer to Y. Let Y be an ordinal response that falls into ordered categories arbitrarily coded as $r = 1, \ldots, k$. To measure the distance between ordinal response classes we use scores $s(r) \in \mathbb{R}$ with $s(1) < s(2) < \cdots < s(k)$. The distance between the true class Y and the predicted class $\hat{Y}$ is then computed from the difference in the corresponding scores, $s(\hat{Y}) - s(Y)$. Treating an ordinal variable as interval scaled by attributing scores might be problematic. However, it has the advantage that loss functions for interval scaled variables like the mean squared error in the form

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(s(\hat{Y}_i) - s(Y_i))^2 \tag{2}$$

might be used (see e.g. Tutz, 2011, p. 474, Fürnkranz and Hüllermeier, 2010, p. 134, and Hechenbichler and Schliep, 2004). When using the simple scores $s(r) = r$, Eq. (2) yields $\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$.

The mean absolute error (MAE) used for our studies on ordinal regression problems is very similar to the mean squared error, with the difference that classification into a distant class is not penalized as much. Using the same notation as before, the mean absolute error for ordinal regression problems is given by

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|s(\hat{Y}_i) - s(Y_i)|. \tag{3}$$

For metric response Y the mean absolute error takes the form $\frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i|$ which directly results from Eq. (3) when using the simple scores $s(r) = r$.

The ranked probability score (RPS) originally introduced by Epstein (1969), is a generalization of the Brier score to multiple categories. It can be computed as the sum of Brier scores for all two-class problems that arise when splitting the sample on all possible thresholds made between two adjacent categories. The RPS has been shown to be particularly appropriate for the evaluation of probability forecasts of ordinal variables (Murphy, 1970). It is defined as

$$RPS = \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{k}(\hat{\pi}_i(r) - I(Y_i \leq r))^2, \tag{4}$$

where k denotes the number of response classes and $\hat{\pi}_i(r)$ denotes the predicted probability of observation i belonging to classes $\{1, \ldots, r\}$. The RPS measures the discrepancy between the predicted cumulative distribution function and the true cumulative distribution function (Murphy, 1970). The predicted cumulative distribution function can be computed from class probabilities that are predicted by a model, that is the estimated probabilities of an observation belonging to classes $r = 1, \ldots, k$. The true cumulative distribution function simplifies to a step function with a step from 0 to 1 at the true value Y_i for observation i. A graphical illustration of the RPS is given in Fig. 1 for an observation i with observed category $y_i = 6$. Fig. 1 shows the true cumulative distribution function (solid gray line) with step from 0 to 1 at the true value $y_i = 6$ and the cumulative distribution function (solid black line) that is obtained from class predictions of a model. Predicted distribution functions are given for two different models. The dashed lines correspond to the distance between the predicted and the true cumulative distribution functions (i.e., $\hat{\pi}_i(r) - I(6 \leq r)$) for a specific category r. These distances are squared when computing the RPS as in Eq. (4). The predicted cumulative distribution function in the left panel indicates that Model 1 does not seem to be very accurate in predicting the value for observation i. Here distances between the true and the predicted cumulative distribution functions are large and the RPS for observation i takes the value $0.21^2+0.41^2+0.52^2+0.61^2+0.64^2+(0.69-1)^2+(0.77-1)^2+(0.89-1)^2+(1-1)^2 = 1.4254$. A much better prediction is obtained when using Model 2. This model assigns the greatest probabilities for values of or around the true value $y_i = 6$. Accordingly, the distances between the true and the predicted cumulative distribution functions are rather small, which is reflected by an RPS of $0.02^2+0.02^2+0.09^2+0.18^2+0.32^2+(0.61-1)^2+(0.85-1)^2+(0.96-1)^2+(1-1)^2 = 0.3199$. It is clear from this illustration that the RPS is smaller (indicating a better prediction) if the predicted probabilities are concentrated near the observed class and is minimal if the predicted probability for the observed class is 1. From its definition it is clear that the RPS uses solely the ordering of the categories and does not require information on the distances between categories.

2.2. Random forests and ordinal regression trees

The RF method is a classification and regression tool that combines several decision trees. An individual tree is fit using a random sample of observations. Breiman (2001) proposed using bootstrapping, in which observations are sampled with replacement from the original data. Later it was shown that bootstrapping induces certain biases and that sampling without replacement from the original data should be preferred (Strobl et al., 2007). For each split in a tree, *mtry* randomly drawn

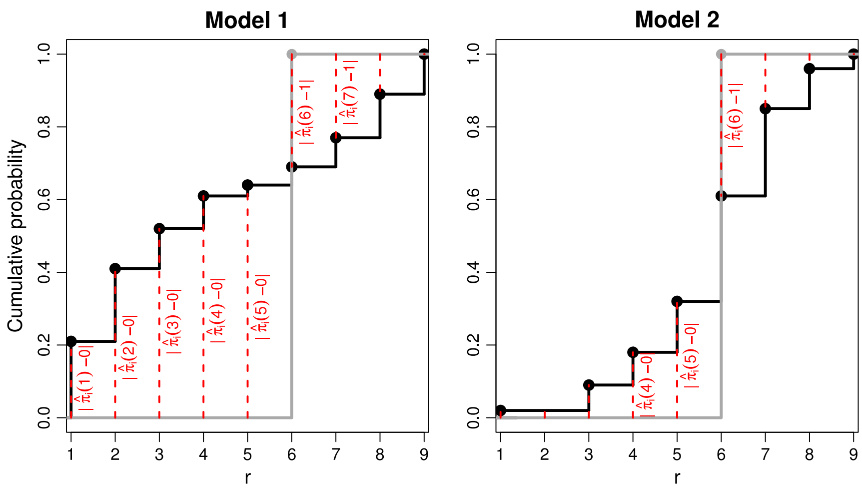

Fig. 1. Predicted (solid black line) and true (solid gray line) cumulative distribution functions for an individual with observed category $y_i = 6$ for two different models. Dashed lines indicate the difference between the predicted and the true cumulative distribution functions, that is $|\hat{\pi}_i(r) - I(y_i \le r)|$, for $r = 1, \ldots, k$ and $y_i = 6$.

predictors are assessed as candidates for splitting and the predictor that yields the best split is chosen. The default values for *mtry* are $\sqrt{p}$ for classification and $\frac{p}{3}$ for regression.

In the RF version of Hothorn et al. (2006b) that we use throughout this paper, conditional inference tests are performed for selecting the best split in an unbiased way. For each split in a tree, each candidate predictor from the randomly drawn subset is tested for its association with the response, yielding a p-value. The predictor with the smallest p-value is selected, and within the selected predictor the best split is chosen. This methodology utilizes a permutation test framework and is thus applicable to problems where both predictors and response can be measured on arbitrary scales, including nominal, ordinal, discrete and continuous variables.

In the case of ordinal response, the response is transformed to a metric scale by attributing scores to the levels of the response. The transformed response is then used to test the association with candidate predictors. If $s(r) \in \mathbb{R}$ denotes the score for category $r \in \{1, 2, \ldots, k\}$ and Y_i denotes the ordinal response of observation i with covariates $X_{ij}, j = 1, \ldots, p$, then the statistic that is used for testing the association between the ordinal response and a predictor variable X_j of arbitrary scale using observations $i = 1, \ldots, n$ is defined as

$$T_j = \sum_{i=1}^{n} g_j(X_{ij})s(Y_i) \tag{5}$$

with $g_j : \mathcal{X}_j \to \mathbb{R}^{p_j}$ being a non-random transformation of the predictor variable X_j from the one-dimensional vector space to a p_j-dimensional vector space. For a numeric predictor variable the transformation is usually the identity function such that $g_j(X_{ij}) = X_{ij}$ and $p_j = 1$. For a nominal categorical predictor variable taking levels in $1, \ldots, m$, g_j is the unit vector of length m with the lth element being equal to one and $p_j = m$. Note that in this case the statistic T_j itself is an m-dimensional vector which is then mapped onto the real line, for example by taking the component that has maximal absolute standardized value; see Hothorn et al. (2006b). For an ordinal categorical predictor variable the class levels are transformed to a metric scale through attributing scores — but now scores are attributed to the levels of the ordinal predictor X_j. If both response and predictor are ordinal variables this test is also known under the name *linear-by-linear association test* (Agresti, 2002). Note that notations and the formula for the statistic (5) are a special case of Hothorn et al. (2006b, p. 8). More precisely, the formula for the statistic arises from the special case in which the response is univariate and all observations $i = 1, \ldots, n$ are used for deriving the test statistic (thus omitting observation weights).

Test statistics are derived by standardizing statistics of the form (5), and from the standardized statistics p-values can be derived. For detailed information we refer the reader to the original literature by Hothorn et al. (2006a,b). The variable with the smallest p-value is then selected for splitting a node. For implementing a split in the selected variable a cutpoint has to be chosen which partitions the data into two subsets. In conditional inference trees the cutpoint is chosen so that the discrepancy in the scores between the two groups of observations (which are defined by the binary split) is maximal. In brief, for each possible split of the sample space into subsets A_1 and A_2, a two-sample statistic is used which measures the discrepancy in the scores between observations with $X_{ij} \in A_1$ and observations with $X_{ij} \in A_2$. This two-sample statistic emerges as a special case of Eq. (5), in which the function g is the indicator function which takes value 1 if X_{ij} is contained in A_1, and 0 otherwise. Among all possible splits the split which maximizes the two-sample test statistic is chosen.

Trees that are constructed based on the statistic (5) are denoted by *ordinal regression trees* to indicate that the trees were constructed by using the ordering information. Tests which take the ordering of a variable into account have higher power compared to tests which ignore the underlying ordering because some degrees of freedom are saved by restricting the possible parameter space (Agresti, 2002, p. 98).

Note that statistics of the form (5) for an ordinal response coincide with statistics for a numeric response with values $s(Y_1), \ldots, s(Y_n)$. This leads to the selection of the same variables and cutpoints in ordinal regression trees and regression trees. Though ordinal regression trees and regression trees have the same tree structure, predictions by the trees are different because the aggregation schemes are different. For an observation that was not used to construct the RF, each tree in the RF makes a prediction. When using regression trees the final prediction is then the average over all tree predictions, which results in a real-valued prediction. For ordinal responses, real-valued predictions are difficult to interpret and there is no standard procedure how to obtain class predictions from these values. Thus the application of regression trees for ordinal responses might not be advisable. Classification and ordinal regression trees in contrast yield estimates for class probabilities, $\hat{P}(Y = r), r \in \{1, \ldots, k\}$. In the RF version of Hothorn et al. (2006b) $\hat{P}(Y = r)$ is estimated by averaging the tree-specific class probabilities. This is in contrast to the classical RF version of Breiman (2001) in which predicted class probabilities are directly computed from the number of trees voting for a class. The class probabilities can then be used to obtain class predictions. In both RF versions the currently implemented strategy for obtaining class predictions for an ordinal response is to classify into the most likely class:

$$\hat{Y} = r \Leftrightarrow \hat{P}(Y = r) = \max_{l=1,\ldots,k} \hat{P}(Y = l).$$

The predicted class thus corresponds to the *mode* of the predicted class probability distribution.

So far we have assumed that predictions are obtained for separate test data which was not used to construct the forest. A special feature of RF is that we can use the same data which was used for constructing the forest for obtaining predictions and for evaluating the prediction performance of the forest. Since each tree is built from a random sample of the data, there are some observations from the data which were not used in its construction ("out-of-bag"). These observations are denoted by *OOB observations*. In a forest each tree is built from a different sample from the original data, so each observation is "out-of-bag" for some of the trees. The prediction for an observation can then be obtained by using only those trees for which the observation was not used for the construction. In this way, a classification is obtained for each observation and the error rate or a different performance measure (like those introduced in Section 2.1 in the case of an ordinal response) can be estimated from these predictions in an unbiased way, in the sense that the resulting estimate reflects the performance expected on independent test data not used for training. When computing the error rate in this way, the resulting error rate is often referred to as *out-of-bag (OOB) error*.

The OOB observations have not only proven useful for estimating the predictive accuracy of a RF but also for computing the RF's permutation VIMs, as outlined in the next section.

2.3. Variable importance measures

RF provides measures that can be used for obtaining a ranking of predictors that reflects the importance of these variables in the prediction of the response and can, for example, be used to select the variables with the best predictive ability. The standard VIMs implemented in the classical RF version of Breiman (2001) are permutation based VIMs and the Gini VIM. The latter has been shown to favor certain types of predictors (Strobl et al., 2007; Nicodemus and Malley, 2009; Nicodemus, 2011; Boulesteix et al., 2012a) and therefore its predictor rankings should be treated with caution. Here we focus on permutation VIMs, which give essentially unbiased rankings of the predictors.

We use a general definition of a permutation VIM which is based on an arbitrary error measure M (e.g., the error rate). It is defined as

$$VIM_j^M = \frac{1}{ntree} \sum_{t=1}^{ntree} (MP_{tj} - M_{tj}), \tag{6}$$

where

- $ntree$ denotes the number of trees in the forest,
- M_{tj} denotes the error of tree t when predicting all observations that are OOB for tree t *before* permuting the values of predictor variable X_j, and,
- MP_{tj} denotes the error of tree t when predicting all observations that are OOB for tree t *after* randomly permuting the values of predictor variable X_j.

The idea underlying this VIM is the following: if the predictor is not associated with the response, the permutation of its values has no influence on the classification, and thus no influence on the error made by the trees. Then the prediction accuracy of the forest is not substantially affected by the permutation and the importance of the predictor takes a value close to zero, indicating that there is no association between the predictor and the response. In contrast, if response and predictor are associated, the permutation of the predictor values destroys this association. "Knocking out" this predictor by permuting its values results in worse prediction. Then the difference in errors before and after randomly permuting the predictor takes a positive value, reflecting the high importance of this predictor.

The two established permutation VIMs for RF arise when using the error rate (for classification trees) or the mean squared error (for regression trees) as the error measure M in Eq. (6). Throughout this paper we will term these measures the *error rate based VIM* and the *MSE-based VIM*, respectively. These VIMs have been explored in the literature in the context of classification and regression tasks, respectively, and are often applied in the literature (e.g., Steidl et al., 2010; Karamanian et al., 2014; Harrington et al., 2014).

In the R package party (Hothorn et al., 2012), the VIM for ordinal regression trees is the error rate based VIM. However, there are no studies that have shown that the error rate is appropriate for ordinal regression trees or that the error rate based VIM gives better rankings than, for example, the MSE-based VIM.

In this paper we introduce two novel permutation VIMs which might be, in addition to the MSE-based VIM, promising for ordinal response data. These VIMs are based on the performance measures introduced in Section 2.1. More precisely, we propose VIMs of the form (6) where the ranked probability score (cf. Eq. (4)) or the mean absolute error (cf. Eq. (3)) is used as the error measure M. These VIMs will be termed the *RPS-based VIM* and the *MAE-based VIM*.

The computation of VIMs involves obtaining tree predictions for the OOB observations. The computational complexity for computing the predictions for a tree is directly related to the tree's depth. Louppe (2014) shows that in the worst case this is $\mathcal{O}(n^2)$, which corresponds to a tree where all splits put a single training observation in one daughter node, while all other training observations are put in the other daughter node. In the best case always half of the training observations are put in each daughter node, so that time complexity for obtaining tree predictions is at best $\mathcal{O}(n\log(n))$. According to Louppe (2014) "the analysis of the average case shows however that pathological cases are not dominant and that, on average, complexity behaves once again as in the best case". The computational complexity of the computation of the RPS is $\mathcal{O}(nk)$, so that for the average case the total time complexity for computing the RPS-based VIM amounts to $\mathcal{O}(n\log(n)+nk)$. Since the number of response classes k is usually much smaller than $\log(n)$, computing the RPS-based VIM is of order $n\log(n)$. The three other VIMs are of the same time complexity, $\mathcal{O}(n\log(n))$.

Our implementation of the two novel VIMs can be obtained from the website http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/070_drittmittel/janitza/rf_ordinal/index.html. Note that the implementation allows the computation of the VIMs from either ordinal regression or classification trees, if constructed using the R package party. In addition to the RPS- and MAE-based VIMs, an implementation of the MSE-based VIM is provided that enables one to compute the MSE-based VIM from ordinal regression trees and from classification trees as well, a feature which is not currently possible using the R package party.

Note that while the error rate based VIM does not take the ordering information of the response levels into account, the three other VIMs do. In our studies we investigate and compare the performances of the four VIMs.

3. Simulation studies

3.1. Studies on prediction accuracy

Using the RF version based on conditional inference trees we compared two RF variants with respect to their ability to predict an ordinal response:

1. *RF ordinal*. RF consisting of ordinal regression trees. Simulations were performed using default scores (i.e., $s(r) = r$, $r = 1, \ldots, k$). Additional studies with quadratic scores $s(r) = r^2$, $r = 1, \ldots, k$, were also performed.
2. *RF classification*. RF consisting of classification trees. The ordinal response is treated as nominal, meaning that the information regarding the natural ordering of the levels of the response is ignored.

Prediction accuracy of a RF variant was assessed using the ranked probability score (RPS; see Eq. (4)) and the error rate (see Eq. (1)) computed for a large independent test dataset of size $n = 10000$ that followed the same distribution as the training set on which the RFs were fit. Note that the RPS and the error rate do not necessarily come to the same conclusion, meaning that the error rate might be lower for one RF variant than for the other but its RPS is higher. Since the error rate does not consider how "severe" a misclassification is, we consider the RPS to be a more appropriate performance measure for evaluating a model that predicts an ordinal response. Thus we will focus on the results that are obtained when using the RPS for assessing prediction accuracy.

3.2. Studies on variable importance

Permutation VIMs based on the different performance measures described in Section 2.3 were applied to see which VIMs are most appropriate in the case of ordinal response. The four different VIMs were computed for RF constructed from ordinal regression trees (*RF ordinal*) as well as for RF using classification trees (*RF classification*; see 3.1).

VIMs give a ranking of the predictors according to their association with the response. To evaluate the quality of the rankings of the VIMs, the area under the curve (AUC) was used. Let the predictor variable indices $B = \{1, \ldots, p\}$ be partitioned into two disjoint sets $B = B_0 \cup B_1$, where B_0 represents the *noise predictors* (without any effect) and B_1 represents the *signal predictors* (with effect). The AUC is computed as follows:

$$AUC = \frac{1}{|B_0|\,|B_1|}\sum_{j\in B_0}\sum_{i\in B_1} I(VI_j < VI_i) + 0.5I(VI_j = VI_i) \quad (7)$$

Table 1
Intercepts for the proportional odds model (9) with $\gamma_{0rg} = \gamma_{0r}$.

Number of response levels	γ_{01}	γ_{02}	γ_{03}	γ_{04}	γ_{05}	γ_{06}	γ_{07}	γ_{08}	γ_{09}
$k = 3$	−1.80	1.80	∞	–	–	–	–	–	–
$k = 6$	−4.50	−1.50	0.00	1.50	4.50	∞	–	–	–
$k = 9$	−5.90	−3.41	−1.55	−0.31	0.31	1.55	3.41	5.90	∞

Table 2
Effects of predictors on the cumulative odds of the proportional odds model (9) for mixture components $g = 1, 2$.

Mixture component	Coefficient vector $\boldsymbol{\gamma}_g^T = (\gamma_{g,1}, \dots, \gamma_{g,65})$															
$g = 1$	(1	1	1	1	1	0.75	0.75	0.75	0.75	0.75	0.5	0.5	0.5	0.5	0.5	0	0, . . . , 0)
$g = 2$	(1	1	−1	−1	0	1	1	−1	−1	0	1	1	−1	−1	0	0	0, . . . , 0)

where $|B_l|$ denotes the cardinality of B_l with $l \in \{0, 1\}$, and $I(\cdot)$ denotes the indicator function (see, e.g., Pepe, 2004). Note that the AUC is often used for evaluating the ability of a method (which may be for example a diagnostic test or a prediction model) to correctly discriminate between observations with binary outcomes (often diseased versus healthy). In our studies, in contrast, the AUC is computed considering the predictor variables $X_1, \dots, X_p$ as the units to be predicted (as noise or signal variables) rather than the observations $i = 1, \dots, n$. The AUC here corresponds to an estimate of the probability that a randomly drawn signal predictor has a higher importance score than a randomly drawn noise predictor. Thus the AUC was computed in our studies to assess the ability of a VIM to differentiate between signal and noise predictors. An AUC value of 1 means that each of these signal predictors receives a higher importance score than any noise predictor, thus indicating perfect discrimination by the VIM. An AUC value of 0.5 means that a randomly drawn signal predictor receives a higher importance score than a randomly drawn noise predictor in only half of the cases, indicating no discriminative ability by the VIM.

3.3. Data simulation

The data were simulated from a mixture of two proportional odds models. Let $P(Y \le r|\boldsymbol{x})$ denote the cumulative probability for the occurrence of a response category equal to or less than r for an individual with covariate vector $\boldsymbol{x}$. This probability is derived from a mixture of two proportional odds models

$$P(Y \le r|\boldsymbol{x}) = \zeta P_1(Y \le r|\boldsymbol{x}) + (1 - \zeta)P_2(Y \le r|\boldsymbol{x}), \tag{8}$$

where ζ is the mixture proportion and $P_1(Y \le r|\boldsymbol{x})$ and $P_2(Y \le r|\boldsymbol{x})$ are the cumulative probabilities that arise from two independent proportional odds models. The proportional odds model for mixture component $g \in \{1, 2\}$ has the form

$$P_g(Y \le r|\boldsymbol{x}) = \frac{\exp(\gamma_{0rg} + \boldsymbol{x}^T\boldsymbol{\gamma}_g)}{1 + \exp(\gamma_{0rg} + \boldsymbol{x}^T\boldsymbol{\gamma}_g)}, \quad r = 1, \dots, k, \tag{9}$$

where the category-specific intercepts satisfy the condition $\gamma_{01g} \le \dots \le \gamma_{0kg} = \infty$. In contrast to the intercepts, the coefficients $\boldsymbol{\gamma}_g$ do not vary over categories. In this case the comparison of two individuals with respect to their cumulative odds $P_g(Y \le r|\boldsymbol{x})/P_g(Y > r|\boldsymbol{x})$ for mixture component g does not depend on the category r, giving the model its name, "proportional odds model" (see e.g., Tutz, 2011).

In our studies, the intercepts do not differ between the two mixture components; that is $\gamma_{0r1} = \gamma_{0r2} = \gamma_{0r}$. The intercepts for the categories were chosen such that the difference between the intercepts of adjacent categories is larger for more extreme categories. Concrete values for the intercepts are provided in Table 1. The simulation setting comprises both predictors not associated with the response (noise predictors) and associated predictors (signal predictors). Predictors $X_1, X_2, \dots, X_{15}$ had an effect on the cumulative odds of the first mixture component. The first five predictors each had a large effect, with corresponding parameter coefficients $\gamma_{1,1} = \gamma_{1,2} = \dots = \gamma_{1,5} = 1$; the second set of five predictors each had a moderate effect, with coefficients $\gamma_{1,6} = \gamma_{1,7} = \dots = \gamma_{1,10} = 0.75$; and the last set of five signal predictors each had a small effect, with coefficients $\gamma_{1,11} = \gamma_{1,12} = \dots = \gamma_{1,15} = 0.5$. The remaining predictors $X_{16}, X_{17}, \dots, X_{65}$ had no effect on the cumulative odds of the first mixture component and their respective coefficients were zero. For the second mixture component fewer predictors had an effect but all effects were large (coefficient of either 1 or −1). Almost all predictors which had an effect for the first component, had an effect for the second — with the exceptions of X_5, X_{10} and X_{15}, which had no effect for the second component. For predictors $X_5, X_{10}, X_{15}, X_{16}, X_{17}, \dots, X_{65}$ the corresponding coefficients were set to zero, while for the other predictors the parameter coefficients were $\gamma_{2,1} = \gamma_{2,2} = \gamma_{2,6} = \gamma_{2,7} = \gamma_{2,11} = \gamma_{2,12} = 1$ and $\gamma_{2,3} = \gamma_{2,4} = \gamma_{2,8} = \gamma_{2,9} = \gamma_{2,13} = \gamma_{2,14} = -1$. Table 2 shows the coefficients for both mixture components. To summarize, there are predictors that have no effect at all, predictors that have an effect for both mixture components and predictors that have an effect for only one mixture component.

Data was generated for sample sizes $n = 200$ and $n = 400$. Let $\boldsymbol{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{i,65})$ denote the covariate vector for the observation i. For the generation of the response value y_i the cumulative probability for the occurrence of a response category equal to or less than r was computed according to (8). Probabilities for classes $r = 1, \ldots, k$ were derived and a multinomial experiment was performed for each observation using its response class probabilities.

For each setting (specified in the subsequent section) 100 datasets were generated.

We also conducted a further study to see if results differ in high-dimensional data settings. This study is not shown here and is provided as supplementary material to this article (see Appendix A).

3.4. Simulation and parameter settings

Various settings were simulated that differed in

- the value for the mixture proportion ζ. Settings were simulated for $\zeta = 0.6$ (data generation based on a mixture of two proportional odds models), $\zeta = 1$ (data generation based on the proportional odds model specified by mixture component $g = 1$) and $\zeta = 0$ (data generation based on the proportional odds model specified by mixture component $g = 2$),
- the number of ordered response levels, chosen as $k = 3$, $k = 6$ and $k = 9$, and,
- the generation of predictor variables. For settings without correlations, $\boldsymbol{x}_i$, $i = 1, \ldots, n$, were drawn from $\mathrm{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$, with $\boldsymbol{I}_p$ denoting the identity matrix of dimension $(p \times p)$ and p denoting the number of predictors. For settings with correlations, $\boldsymbol{x}_i$, $i = 1, \ldots, n$, were drawn from $\mathrm{N}(\boldsymbol{0}_p, \boldsymbol{\Sigma}_p)$ with block diagonal covariance matrix

$$\boldsymbol{\Sigma}_p = \begin{bmatrix} \boldsymbol{A}_{\text{signal}} & 0 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{A}_{\text{noise}_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{A}_{\text{noise}_2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{A}_{\text{noise}_3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{A}_{\text{noise}_4} & 0 \\ 0 & 0 & 0 & 0 & 0 & \boldsymbol{A}_{\text{noise}_5} \end{bmatrix}.$$

The first block matrix $\boldsymbol{A}_{\text{signal}} \in \mathbb{R}^{(15\times 15)}$ determined the correlations among the signal predictors $X_1, \ldots, X_{15}$. It was defined as $\boldsymbol{A}_{\text{signal}} = (a_{ij})$ with

$$a_{ij} = \begin{cases} 1, & i = j \\ 0.8, & i \neq j;\, i, j \in \{1, 3, 6, 8, 11, 13\} \\ 0, & \text{otherwise} \end{cases}$$

in this way generating uncorrelated and also strongly correlated signal predictors. The matrices $\boldsymbol{A}_{\text{noise}_j} \in \mathbb{R}^{(10\times 10)}$ for $j = 1, \ldots, 5$ were given by

$$\boldsymbol{A}_{\text{noise}_j} = \begin{bmatrix} 1 & \rho_j & \cdots & \rho_j \\ \rho_j & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_j \\ \rho_j & \cdots & \rho_j & 1 \end{bmatrix},$$

and determined correlations among a set of 10 noise predictor variables with $\rho_1 = 0.8$, $\rho_2 = 0.6$, $\rho_3 = 0.4$, $\rho_4 = 0.2$ and $\rho_5 = 0$.

Simulation studies were performed using the unbiased RF version based on conditional inference trees which is implemented in the `R` package party. For our studies, the setting for unbiased tree construction was used as suggested by Strobl et al. (2007). In this setting no p-value threshold is applied when selecting the optimal split (by setting the parameter `mincriterion` in `cforest_control` to zero). No other stopping criteria such as a minimum number of observations in a terminal node or a minimum number of observations required for a node to be split were applied by setting `minbucket` and `minsplit` to their lower limits. The number of randomly drawn candidate predictors `mtry` was set to the default value $\lfloor\sqrt{p}\rfloor$, where p denotes the number of predictors (here $p = 65$) and the number of trees `ntree` was set to 1000.

3.5. Results

Prediction accuracy

Fig. 2 shows the results of the simulation studies on the comparison of *RF ordinal* and *RF classification* with respect to their predictive accuracy (measured in terms of RPS) for the sample size of $n = 200$ (results for $n = 400$ are very similar and thus not shown). For a direct comparison, we show the ratio of the RPS for *RF ordinal* to that for *RF classification*. Values of the RPS ratio below 1 mean that the prediction accuracy as measured by RPS is smaller for *RF ordinal* and thus are in favor of *RF ordinal*. Conversely, values above 1 mean that the prediction accuracy as measured by RPS is larger for *RF ordinal* and

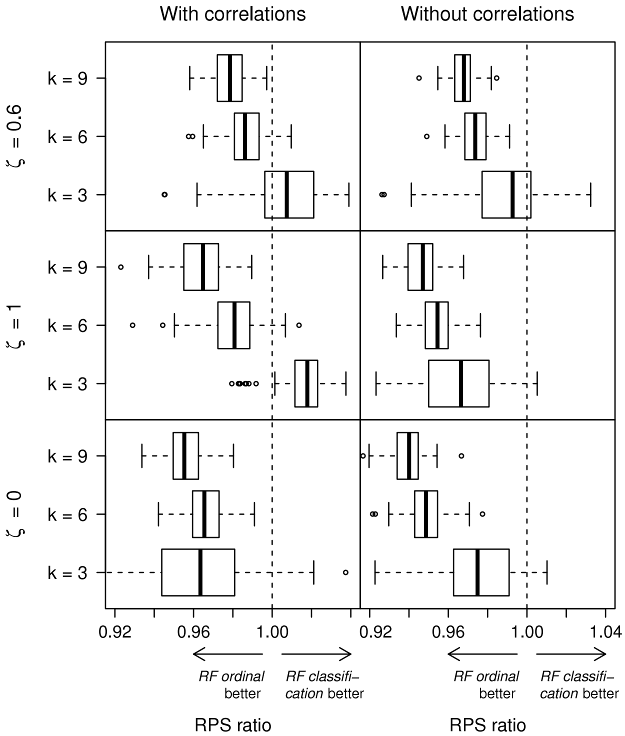

Fig. 2. Performance ratio for *RF ordinal* versus *RF classification* for simulated data. A ratio of the ranked probability scores (RPS) below 1 indicates a better prediction accuracy of *RF ordinal* and a ratio above 1 indicates a better prediction accuracy of *RF classification*. Data was generated for $n = 200$ from a mixture of proportional odds models (8) with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row). Data was generated for $k \in \{3, 6, 9\}$ ordered response levels and for settings in which predictors correlate (left column) and in which all predictors are uncorrelated (right column).

advocate the use of *RF classification* for prediction purposes. For values close to 1 prediction accuracies of *RF ordinal* and *RF classification* are comparable. In all settings the ratio of RPS is in the range [0.92; 1.04] and thus is very close to 1, so there are no large differences between the prediction accuracies of the forest types in our simulation studies. However, one can observe a trend towards better performance of *RF ordinal* for a larger number of response levels. Overall, the performance is better for *RF ordinal* in most of the settings, except for $k = 3$, in which the performance of *RF classification* is better in two of six settings. Similar results were obtained when performance was measured in terms of the error rate (results not shown). Note that the results presented here were obtained by using equally spaced scores. The results are very similar when using quadratic scores, which suggests that our conclusions do not depend on the specific choice of scores for *RF ordinal*.

Performance of variable importance measures

Figs. 3–5 show the results of our simulation studies on the performance of VIMs for $n = 200$ when using the four VIMs outlined in Section 2.3, computed for both *RF ordinal* and *RF classification*. Results for $n = 400$ are comparable and thus not shown. Here we only show the results when using default (i.e., equally spaced) scores for tree construction and MSE- and MAE-based VIM computation. Very similar results were obtained when specifying quadratic scores. This suggests that specific values for the scores do not seem to have a significant impact as long as the scores reflect the correct ordering of the levels.

In the settings with 9 response levels (Fig. 3) the performances of the MSE-based VIM and our two novel permutation VIMs are consistently better than that of the error rate based VIM, independent of the type of trees used (ordinal regression or classification trees). Obviously, making use of the ordering is advantageous when deriving the importance of variables for these settings. Interestingly, in some settings the difference is rather small and in others it is more pronounced. Similar results are obtained for the setting with 6 response levels (Fig. 4). However, the difference between the error rate based VIM and the other VIMs is less pronounced than for the settings with a 9-category response variable. In all settings in which the response has only 3 levels the differences between the VIMs are marginal (Fig. 5), though overall our novel VIMs and the MSE-based VIM remain superior. In our studies the RPS-based, MSE-based and MAE-based VIM show comparable performances.

The results suggest that the performances of all VIMs can in some settings be further improved by making use of the ordering in the construction of trees, through the application of ordinal regression trees. If used in combination with ordinal regression trees, our novel VIMs and the MSE-based VIM achieved the most accurate predictor rankings. The worst rankings in contrast were obtained for the classical error rate based permutation VIM (which is currently in use for ordinal responses

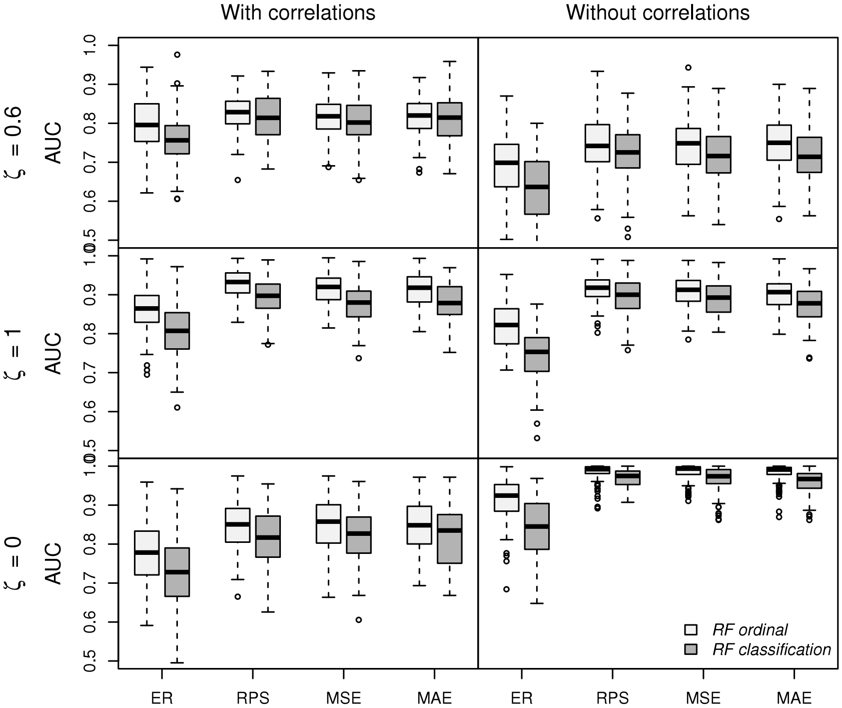

Fig. 3. Performance of different VIMs for *RF ordinal* and *RF classification*: settings for a 9-category ordinal response. VIMs are computed using the error rate (ER), the ranked probability score (RPS), the mean squared error (MSE) and the mean absolute error (MAE). Data was generated for $n = 200$ using a mixture of proportional odds models (8) with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row).

in the R package party) computed from classification trees. This indicates that predictor rankings are worst when making no use of the ordering at all, neither in tree construction nor in the computation of the variables' importance scores.

A plausible explanation for the improvement in the ranking by using ordinal regression trees is that in ordinal regression trees it is more likely that a predictor associated with the response is selected for a split. A predictor that is often selected in a tree and occurs close to the root node of the tree is likely to receive a high importance score. The advantage when applying ordinal regression trees is that the power of the statistical test to correctly detect an association between a predictor and the ordinal response is higher. It is thus less likely that a noise predictor yields a lower p-value just by chance and is selected for the split. Results obtained for the described simulation studies provide evidence for this. One can, for example, inspect the trees of a forest and compute the number of trees for which an influential predictor was chosen for the first split. If the fraction of trees is significantly higher for the forest consisting of ordinal regression trees, this is an indication that ordinal regression trees are more accurate in selecting predictors for a split compared to classification trees. For our simulation studies we calculated the fraction of trees where a signal predictor was selected for the first split for both *RF ordinal* and *RF classification*; the results are displayed in Fig. 6. The results confirm our hypothesis that *RF ordinal* is more accurate in selecting important predictors for a split than *RF classification*. Since the power of a test that takes into account the ordering increases with the number of ordered categories, the discrepancy between *RF ordinal* and *RF classification* is most pronounced for $k = 9$ and least pronounced for $k = 3$.

4. Real data applications

In our studies we considered five publicly available real datasets with an ordinal response. Note that we did not perform a selection of the datasets depending on the obtained results but instead report results for all datasets that we analyzed.

4.1. Data

The Very Low Birth Weight Data was analyzed by O'Shea et al. (1998) for identifying perinatal events from sonographical and echodensity measurements. The data can be obtained from the website http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets. In our analyses we aimed to predict the Apgar score (a score for the physical health status of a newborn measured on a 9-point scale) from diverse factors such as medication the mother took during pregnancy, weight and sex of the newborn and the type of delivery.

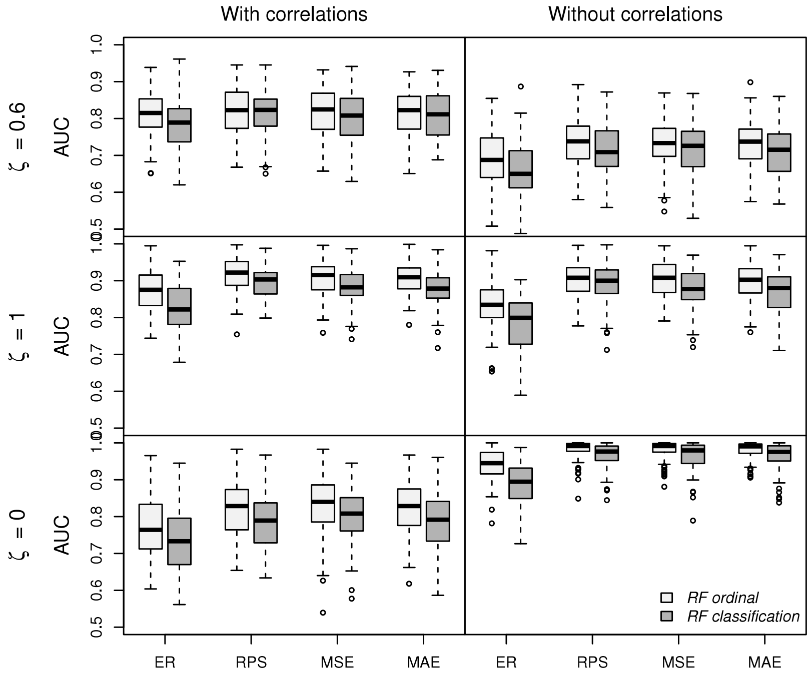

Fig. 4. Performance of different VIMs for *RF ordinal* and *RF classification*: settings for a 6-category ordinal response. VIMs are computed using the error rate (ER), the ranked probability score (RPS), the mean squared error (MSE) and the mean absolute error (MAE). Data was generated for $n = 200$ using a mixture of proportional odds models (8) with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row).

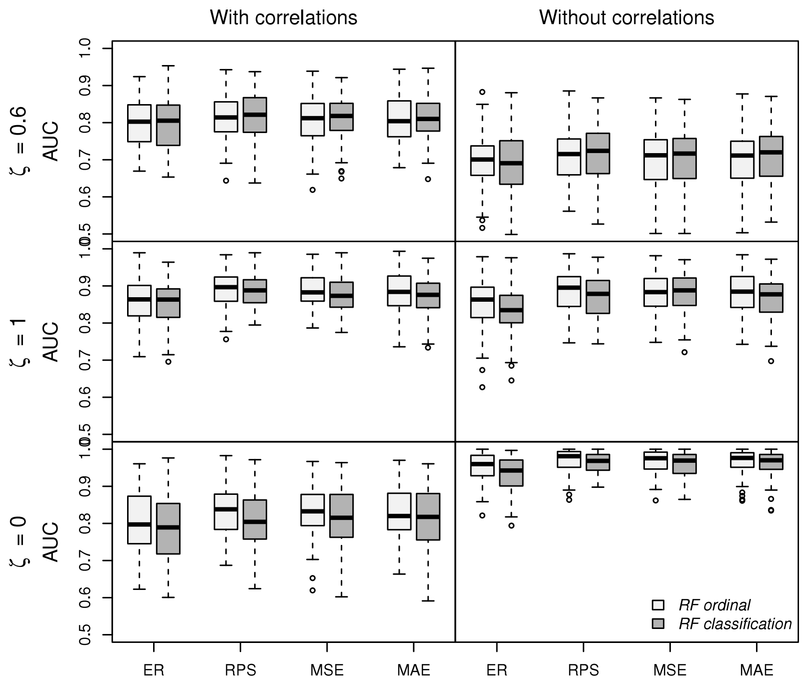

Fig. 5. Performance of different VIMs for *RF ordinal* and *RF classification*: settings for a 3-category ordinal response. VIMs are computed using the error rate (ER), the ranked probability score (RPS), the mean squared error (MSE) and the mean absolute error (MAE). Data was generated for $n = 200$ using a mixture of proportional odds models (8) with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row).

The Wine Quality Data is available from the UCI repository (http://archive.ics.uci.edu/ml/datasets.html); see also Cortez et al. (2009) for details on the data. The response to be predicted from physicochemical measurements (like alcohol

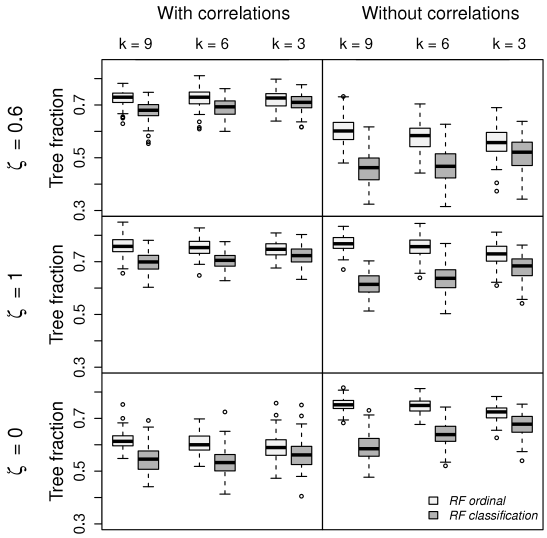

Fig. 6. Fraction of trees in *RF ordinal* and *RF classification* where an influential predictor was selected for the first split. Distributions arise from 500 replications of the simulation setting described in Section 3.1 with $k = 3$ response levels, $k = 6$ and $k = 9$. Data was generated for $n = 200$ using a mixture of proportional odds models (8) with mixture proportions $\zeta = 0.6$ (upper row), $\zeta = 1$ giving weight 1 to the first mixture component $g = 1$ (middle row), and $\zeta = 0$ giving weight 1 to the second mixture component $g = 2$ (lower row).

concentration or residual sugar) was the quality of a wine, measured on a scale from 0 (poorest quality) to 10 (highest quality). There were no observations with the highest quality (i.e., a score of 10) and very poor quality (score from 0 - 2). Due to their small number ($n = 5$), we removed observations with a score of 9 from the data.

The National Health and Nutrition Examination Survey (NHANES) is a series of cross-sectional surveys of the US population (National Center for Health Statistics, 2012). The data can be obtained from the institution's homepage. We chose a subset of the data that had been previously analyzed by Janitza et al. (2015). We considered the self-reported general health status as the outcome variable to be predicted from demographical and health-related factors. The response is categorized into five categories (1: excellent, 2: very good, 3: good, 4: fair, 5: poor).

The data for the SUPPORT Study can be obtained from the website http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets. The considered dataset is a random sample of 1000 patients from phases I & II of the **S**tudy to **U**nderstand **P**rognoses and **P**references for **O**utcomes and **R**isks of **T**reatment (SUPPORT) (Knaus et al., 1995). Several outcomes in seriously ill hospitalized adults have been considered. We focused on the prediction of functional disability, which is categorized into 5 ordered categories from slight to severe (see Table 3 for details).

The Mammography Experience Data was analyzed by Hosmer and Lemeshow (2004, p. 264), who studied the relationship between mammography experience (have never had a mammography, have had one within the last year, last mammography greater than one year ago) and the attitude towards mammography based on a study questionnaire. The data is part of the R package TH.data.

For all datasets (except for the Very Low Birth Weight Data) we excluded covariates for which more than 10% of the observations had missing values. Observations with missing values in any of the included covariates were deleted. An overview of the number of response levels, predictor variables and observations for the datasets (as used for our analysis) is given in Table 4. Table 3 gives an overview of the response variables considered in our analyses. Note that we had types of responses ranging from different scoring systems (Wine Quality Data, NHANES Data and Very Low Birth Weight Data), to categorizations of functional disability (SUPPORT Study), to the recentness of events, as grouped into 3 categories (Mammography Experience Data).

All RF parameters were defined as described for the simulated data in Section 3 ($\texttt{mtry} = \lfloor\sqrt{p}\rfloor$, $\texttt{ntree} = 1000$, no early stopping). Default (i.e., equally spaced) scores were used in our analysis.

4.2. Studies on prediction accuracy

The ranked probability score (RPS; see Eq. (4)) and the error rate (see Eq. (1)) were used to assess prediction accuracies by *RF ordinal* and *RF classification*. Prediction accuracies were assessed using 10-fold cross-validation. The cross-validation was repeated 500 times to obtain more stable results.

4.3. Studies on variable importance

When using real data one usually faces the problem that it is unknown which of the variables are important and which are not. As we know from our investigations (not shown), for all datasets there are at least some variables which improve

Table 3
Response variables of the five real datasets and their frequency in the analyzed data.

Data	Considered response variable	Levels
Very low birth weight	Apgar score	1 (life-threatening) ($n = 33$) 2 ($n = 16$) 3 ($n = 19$) 4 ($n = 15$) 5 ($n = 25$) 6 ($n = 27$) 7 ($n = 35$) 8 ($n = 36$) 9 (optimal physical condition) ($n = 12$)
Wine quality	Wine quality score[a]	3 (moderate quality) ($n = 20$) 4 ($n = 163$) 5 ($n = 1457$) 6 ($n = 2198$) 7 ($n = 880$) 8 (high quality) ($n = 175$)
NHANES	Self-reported health status	1 — excellent ($n = 198$) 2 — very good ($n = 565$) 3 — good ($n = 722$) 4 — fair ($n = 346$) 5 — poor ($n = 83$)
SUPPORT study	Functional disability	1 — patient lived 2 months, and from an interview (taking place 2 months after study entry) there were no signs of moderate to severe functional disability ($n = 310$) 2 — patient was unable to do 4 or more activities of daily living 2 months after study entry; if the patient was not interviewed but the patient's surrogate was, the cutoff for disability was 5 or more activities ($n = 104$) 3 — Sickness Impact Profile total score is at least 30 2 months after study entry ($n = 57$) 4 — patient intubated or in coma 2 months after study entry ($n = 7$) 5 — patient died before 2 months after study entry ($n = 320$)
Mammography experience	Last mammography visits	1 — never ($n = 234$) 2 — within a year ($n = 104$) 3 — over a year ($n = 74$)

[a] There were no observations with categories 0, 1, 2, 9, 10 in the analyzed dataset.

Table 4
Characteristics of the five real datasets.

Data	No. response levels k	No. predictors p	Sample size n
Very low birth weight	9	10	218
Wine quality	6	11	1599
NHANES	5	26	1914
SUPPORT study	5	16	798
Mammography experience	3	5	412

response prediction since the predictions by the constructed forests were always more accurate than the predictions by the null model (i.e., that without covariates). If we assume that we had an additional set of variables which are not associated with the response, we would be able to investigate and compare the discriminative abilities of the VIMs: a well-performing VIM is namely expected to attribute higher importance scores to the original (and potentially important) predictors than to the noise predictors.

We proceeded as follows:

- We augmented the original data by a set of noise predictors. This was done by duplicating the set of original predictor variables and then randomly permuting the rows of this duplicated predictor set. In this way we made sure that each predictor within this duplicated predictor set was unrelated to the response variable, while preserving realistic correlation structures within the duplicated predictor set.
- We fit *RF ordinal* and *RF classification* to this augmented data and derived the variables' importance scores using each of the four permutation VIMs described in Section 2.3.
- We used the area under the curve (AUC) to measure the performance of VIMs. The AUC is an estimate of the probability that a randomly drawn predictor from the original (i.e., unpermuted) set of predictors would obtain a higher importance score than a randomly drawn predictor from the permuted set of predictors (see Section 3.2).

This process was repeated 500 times. Note that while in Section 3.2 an AUC value of 1 indicated perfect discrimination between signal and noise predictors, here we expect that perfect discrimination can already be obtained for AUC values lower than 1: since it is likely that not all of the original variables are truly influential predictors, some of them actually should be regarded as noise predictors instead. However, this does not pose a problem for our studies because our aim is to

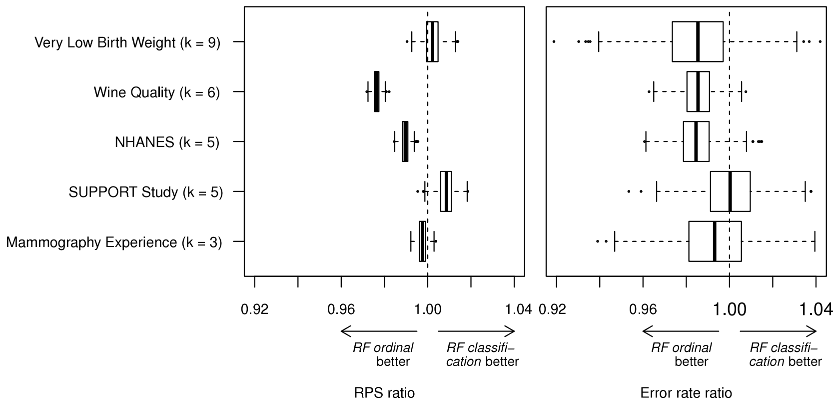

Fig. 7. Performance ratio for *RF ordinal* versus *RF classification* for the five real datasets. Values below 1 indicate a better performance of *RF ordinal* and values above 1 indicate a better performance of *RF classification*. Prediction accuracy was measured by ranked probability score (left) and error rate (right) using 10-fold cross-validation repeated for 500 random splits.

compare the VIMs with respect to discriminative ability, so we are interested in the differences in their AUC values rather than the absolute AUC values.

4.4. Results

Prediction accuracy

The results on prediction accuracy of *RF ordinal* and *RF classification* based on the five real datasets are shown in Fig. 7. For a direct comparison of *RF ordinal* and *RF classification* we computed the RPS ratio (left panel) and the error rate ratio (right panel). The results shown in Fig. 7 are in line with the results obtained from our simulation studies in Section 3; overall the differences in prediction accuracies are rather small. The ratios are even closer to 1 than the ratios obtained for the simulated data (cf. Fig. 2). In contrast to the simulated data, we do not observe a trend with respect to the number of response levels. Instead, which RF variant performs better seems to highly depend on the considered dataset as well as on which performance measure is used; when using the RPS as the performance measure (which we consider to be more appropriate than the error rate) for three of the datasets (Wine Quality Data, NHANES Data, Mammography Experience Data) an at least marginally better predictive accuracy was obtained by *RF ordinal*, while for the other two datasets (the Very Low Birth Weight Data and the SUPPORT Study) *RF classification* gave slightly more accurate predictions. In contrast, *RF ordinal* is for all datasets at least as good as *RF classification* when the error rate is used as the performance measure.

Performance of variable importance measures

Fig. 8 shows the AUC values over the 500 repetitions. Very marginal differences in performance can be observed when the importance of variables is derived from ordinal regression trees compared to classification trees. The performance of a VIM seems to highly depend on the nature of the response variable since results differ between the datasets. While for the Very Low Birth Weight Data and for the NHANES Data all three VIMs that take into account the ordering in response levels have better discriminative ability than the error rate based VIM, there is hardly any difference between the error rate based VIM and our two novel VIMs (based on the RPS and MAE) for the other three datasets. Note that for the Wine Quality Data we obtain perfect discrimination for all VIMs, which indicates that all variables in the original dataset are associated with the quality of a wine. Interestingly, in these studies, compared to our two novel VIMs based on the RPS and the MAE, the MSE-based VIM always performs worse or has equal performance at best.

5. Discussion

The use of the ordering in the levels of an ordinal response variable in tree construction is not supported by the classical RF version of Breiman (2001). In practice, data with ordinal responses have often been handled using classification or regression trees. However, the former fully ignores the ordering and the latter assumes the response to be measured on a metric scale and yields metric values instead of class predictions. The RF implementation of Hothorn et al. (2006b) in contrast, implements so-called ordinal regression trees. The tree construction of ordinal regression trees corresponds to that in regression trees, while predictions and the importance of variables are obtained in the same way as for classification trees. This RF version is thus promising for applications in which the response has an inherent ordering. Moreover, this version is based on a conditional inference framework and, in contrast to the classical RF version of Breiman (2001), implements unbiased split selection. For these reasons we based our studies on the RF version of Hothorn et al. (2006b).

In this paper we investigated whether prediction accuracy improves when making use of the ordering of the levels of the response variable. For this purpose, using simulated and real data, we compared the prediction accuracy of RF composed

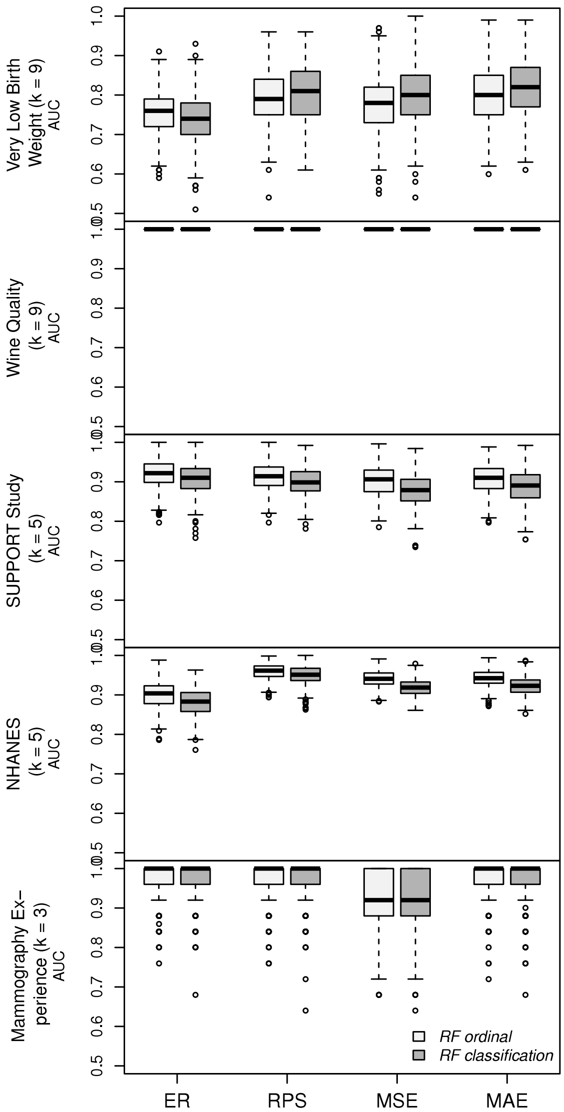

Fig. 8. Performance of different VIMs for five real datasets when computed on *RF ordinal* and *RF classification*. VIMs are computed using the error rate (ER), the ranked probability score (RPS), the mean squared error (MSE) and the mean absolute error (MAE). The performance of VIMs is measured in terms of the area under the curve (AUC), which corresponds to the probability that a randomly drawn potentially important predictor has a higher importance value than a randomly drawn noise predictor.

of classification trees to that of RF composed of ordinal regression trees. Our studies indicate that there are only small differences in prediction accuracy. For 16 of 18 studies based on simulated data and for 3 of 5 studies based on real data, more accurate class predictions were obtained for RF consisting of ordinal regression trees, suggesting that ordinal regression trees are a reasonable alternative to classification trees if the response is ordinal. However, the differences were only small and their practical relevance is questionable.

The choice of the scores (reflecting distances in response levels), which are required for constructing ordinal regression trees, did not impact the prediction accuracies of the ordinal regression trees. This indicates that our conclusions do not depend on the specific choice of the scores.

Prediction accuracy was primarily assessed by using the ranked probability score. We also investigated the results if prediction accuracy was evaluated based on the error rate and two alternative measures which are described in the supplementary material (see Appendix A). The results obtained with the error rate and the two alternative measures were consistent with the findings reported in this paper. Thus our conclusions do not depend on the choice of the accuracy measure.

Note that in this paper we investigated the incorporation of the ordering of the response levels when constructing trees and when computing the importance of variables. The ordering of the response levels in the context of another stage could also be considered in future studies, namely when aggregating tree predictions to obtain a final prediction of a class (see, e.g., Tutz, 2011, Section 15.9). In the context of k-nearest-neighbors it has for example been shown that such a procedure might give more accurate predictions (Hechenbichler and Schliep, 2004).

In addition to prediction accuracy, we also investigated if making use of the ordering for the computation of VIMs leads to more accurate predictor rankings. In the presence of an ordinal response the current RF implementation of Hothorn et al. (2006b) uses the error rate based permutation VIM. We introduced two novel permutation VIMs for RF that are promising in settings in which the response has an inherent ordering. Our results on simulated and on real data showed that a VIM which makes use of the ordering in the levels of the response yields in many cases an at least slightly more accurate predictor ranking than the classical error rate based VIM, and thus should be used when analyzing ordinal response data. Our studies suggest that by using ordinal regression trees a further improvement in the predictor rankings might be obtained. We discovered that this is most likely related to the fact that ordinal regression trees more often select relevant predictors for a split than classification trees since hypothesis tests used for split selection in conditional inference trees have higher statistical power for the detection of relevant effects if making use of the ordering in the response levels. In data settings where the response variable is ordinal we thus recommend using a permutation VIM which makes use of the ordering in combination with ordinal regression trees if the aim is to obtain a predictor ranking or to select important variables.

Among the VIMs that make use of the ordering, our two novel VIMs outperformed the well-known MSE-based VIM on real data. Note that the MSE-based VIM was developed for regression trees but had not been considered for ordinal responses to this point. While the RPS-based VIM relies only on the ordering of the levels, the MAE- and MSE-based VIMs require the specification of distances between the response levels. Though in our simulation studies different distances did not lead to different results, we cannot be sure that this also applies to other settings. Thus we recommend the use of the RPS-based VIM – which does not make any assumptions on the distance between response levels – over the MAE- and MSE-based VIMs.

All results reported in this paper are based on settings, in which the number of candidate predictors is smaller than the number of observations. However, results generalize to high-dimensional data settings, as shown by simulations in which the number of candidate predictors is larger than the number of observations (provided as supplementary material, see Appendix A).

The `R` code implementing our novel VIMs is provided at the website http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/070_drittmittel/janitza/rf_ordinal/index.html. The VIMs can be applied to forest objects fitted using the function `cforest` from the party package. Though in our studies on the performance of VIMs we exclusively used the RF version of Hothorn et al. (2006b), we expect that VIMs that make use of the ordering, like the RPS-based VIM, give more accurate rankings also when using the classical RF version of Breiman (2001).

Acknowledgments

SJ was supported by grants BO3139/6-1 and BO3139/2-2 from the German Science Foundation to ALB. The authors thank Bernd Bischl for valuable discussions on the computational details of the variable importance measures and Rory Wilson for linguistic improvements of the paper.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2015.10.005.

References

Agresti, A., 2002. Categorical Data Analysis. John Wiley, New York.
Archer, K., Mas, V., 2009. Ordinal response prediction using bootstrap aggregation, with application to a high-throughput methylation data set. Stat. Med. 28 (29), 3597–3610.
Boulesteix, A.L., Bender, A., Bermejo, J.L., Strobl, C., 2012a. Random forest Gini importance favours SNPs with large minor allele frequency: assessment, sources and recommendations. Brief. Bioinform. 13, 292–304.
Boulesteix, A.L., Janitza, S., Kruppa, J., König, I., 2012b. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2 (6), 493–507.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Briggs, F., Goldstein, B., McCauley, J., Zuvich, R., De Jager, P., Rioux, J., Ivinson, A., Compston, A., Hafler, D., Hauser, S., et al., 2010. Variation within DNA repair pathway genes and risk of multiple sclerosis. Am. J. Epidemiol. 172 (2), 217.
Chang, J., Yeh, R., Wiencke, J., Wiemels, J., Smirnov, I., Pico, A., Tihan, T., Patoka, J., Miike, R., Sison, J., et al., 2008. Pathway analysis of single-nucleotide polymorphisms potentially associated with glioblastoma multiforme susceptibility using random forests. Cancer Epidemiol. Biomarkers Prevent. 17 (6), 1368–1373.
Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009. Modeling wine preferences by data mining from physicochemical properties. Decis. Support Syst. 47 (4), 547–553.
Epstein, E.S., 1969. A scoring system for probability forecasts of ranked categories. J. Appl. Meteorol. 8 (6), 985–987.
Fürnkranz, J., Hüllermeier, E., 2010. Preference Learning. Springer, Berlin.
Harrington, D.L., Liu, D., Smith, M.M., Mills, J.A., Long, J.D., Aylward, E.H., Paulsen, J.S., 2014. Neuroanatomical correlates of cognitive functioning in prodromal Huntington disease. Brain Behav. 4 (1), 29–40.
Hechenbichler, K., Schliep, K., 2004. Weighted k-nearest-neighbor techniques and ordinal classification. Discussion Paper 399, University of Munich. https://epub.ub.uni-muenchen.de/1769/1/paper_399.pdf.
Hosmer Jr., D.W., Lemeshow, S., 2004. Applied Logistic Regression. John Wiley & Sons, New York.
Hothorn, T., Hornik, K., Van De Wiel, M., Zeileis, A., 2006a. A lego system for conditional inference. Amer. Statist. 60 (3), 257–263.
Hothorn, T., Hornik, K., Zeileis, A., 2006b. Unbiased recursive partitioning: A conditional inference framework. J. Comput. Graph. Statist. 15 (3), 651–674.

Hothorn, T., Hornik, K., Zeileis, A., 2012. Party: a laboratory for recursive partytioning. R package version 10-3, URL http://cranr-projectorg/package=party.
Janitza, S., Binder, H., Boulesteix, A.L., 2015. Pitfalls of hypothesis tests and model selection on bootstrap samples: Causes and consequences in biometrical applications. Biometrical J. http://dx.doi.org/10.1002/bimj.201400246.
Janitza, S., Strobl, C., Boulesteix, A.L., 2013. An AUC-based permutation variable importance measure for random forests. BMC Bioinformatics 14, 119.
Karamanian, V.A., Harhay, M., Grant, G.R., Palevsky, H.I., Grizzle, W.E., Zamanian, R.T., Ihida-Stansbury, K., Taichman, D.B., Kawut, S.M., Jones, P.L., 2014. Erythropoietin upregulation in pulmonary arterial hypertension. Pulmon. Circ. 4 (2).
Knaus, W.A., Harrell, F.E., Lynn, J., Goldman, L., Phillips, R.S., Connors, A.F., Dawson, N.V., Fulkerson, W.J., Califf, R.M., Desbiens, N., et al., 1995. The support prognostic model: objective estimates of survival for seriously ill hospitalized adults. Ann. Intern. Med. 122 (3), 191–203.
Liu, C., Ackerman, H., Carulli, J., 2011. A genome-wide screen of gene–gene interactions for rheumatoid arthritis susceptibility. Hum. Genet. 129 (5), 473–485.
Louppe, G., 2014. Understanding random forests: From theory to practice. arXiv preprint arXiv:1407.7502.
Murphy, A.H., 1970. The ranked probability score and the probability score: A comparison. Mon. Weather Rev. 98 (12), 917–924.
National Center for Health Statistics (2012). NHANES 2007 to 2008 public data general release file documentation. http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/generaldoc_e.htm.
Nicodemus, K., 2011. Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. Brief. Bioinform. 12 (4), 369–373.
Nicodemus, K., Callicott, J., Higier, R., Luna, A., Nixon, D., Lipska, B., Vakkalanka, R., Giegling, I., Rujescu, D., Clair, D., et al., 2010. Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: biological validation with functional neuroimaging. Hum. Genet. 127 (4), 441–452.
Nicodemus, K., Malley, J., 2009. Predictor correlation impacts machine learning algorithms: implications for genomic studies. Bioinformatics 25 (15), 1884–1890.
O'Shea, T.M., Kothadia, J.M., Roberts, D.D., Dillard, R.G., 1998. Perinatal events and the risk of intraparenchymal echodensity in very-low-birthweight neonates. Paediatr. Perinat. Epidemiol. 12, 408–421.
Pepe, M., 2004. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, USA.
Piccarreta, R., 2001. A new measure of nominal-ordinal association. J. Appl. Stat. 28 (1), 107–120.
Steidl, C., Lee, T., Shah, S.P., Farinha, P., Han, G., Nayar, T., Delaney, A., Jones, S.J., Iqbal, J., Weisenburger, D.D., et al., 2010. Tumor-associated macrophages and survival in classic Hodgkin's lymphoma. New Engl. J. Med. 362 (10), 875–885.
Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25.
Sun, Y., Cai, Z., Desai, K., Lawrance, R., Leff, R., Jawaid, A., Kardia, S., Yang, H., 2007. Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests. BMC Proc. 1 (Suppl 1), S62.
Tutz, G., 2011. Regression for Categorical Data. Cambridge University Press, New York.