Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance

Carolin Strobl[1] and Achim Zeileis[2]

[1] Department of Statistics, Ludwig-Maximilians-Universität München
Ludwigstraße 33, D-80539 München, Germany,
Carolin.Strobl@stat.uni-muenchen.de
[2] Department of Statistics and Mathematics, Wirtschaftsuniversität Wien
Augasse 2–6, A-1090 Wien, Austria, *Achim.Zeileis@wu-wien.ac.at*

Abstract. Random forests have become a widely-used predictive model in many scientific disciplines within the past few years. Additionally, they are increasingly popular for assessing variable importance, e.g., in genetics and bioinformatics. We highlight both advantages and limitations of different variable importance scores and associated testing procedures. For the test of Breiman and Cutler (2008), we investigate the statistical properties and find that the power of the test depends both on the sample size and the number of trees in an undesirable way that nullifies any significance judgments. Moreover, the specification of the null hypothesis of this test is discussed in the context of correlated predictor variables.

Keywords: feature selection, variable importance, permutation tests

1 Introduction

Within the past few years, random forests (Breiman (2001)) have become a popular and widely-used tool for non-parametric regression in many scientific areas such as genetics, bioinformatics, clinical medicine and psychology. Random forests are typically found to have high predictive accuracy and are applicable even in high dimensional problems, as well as problems involving correlated predictor variables and high-order interactions. Recently, their variable importance measures have also been suggested for the selection of relevant predictor variables in the analysis of microarray data, DNA sequencing and many other applications (cf. e.g., Lunetta et al. (2004), Arun and Langmead (2005), Bureau et al. (2005), Huang et al. (2005), Diaz-Uriarte and Alvarez de Andrés (2006), Qi et al. (2006) ,Ward et al. (2006)). Most random forest implementations offer two different variable importance measures (plus class-wise versions of the latter): the Gini importance, based on the Gini gain split selection criterion, and the permutation accuracy importance. However, Strobl et al. (2007) show that, when predictor variables vary in their scale of measurement or their number of categories, the Gini importance is biased

in favor of, e.g., predictor variables with many categories. As opposed to that, the permutation importance is reliable when the ensembles of trees are built on subsamples drawn without replacement instead of bootstrap samples drawn with replacement (Strobl et al. (2007)). Therefore, in the following we will only consider the permutation importance.

A key advantage of the random forest permutation importance, as compared to univariate screening methods, is that it covers the impact of each predictor variable individually as well as in multivariate interactions with other predictor variables. For example, Lunetta et al. (2004) find that genetic markers relevant in interactions with other markers or environmental variables can be detected more efficiently by means of random forests than by means of univariate screening methods like Fisher's exact test. Random forests can also be applied when predictor variables are highly correlated.

Currently, most applications of the random forest permutation importance rely on a merely descriptive ranking of the potential predictor variables with respect to their importance: The few top-ranked predictors are selected for further exploration, where the number of selected variables is chosen arbitrarily or with respect to subject matter. A different approach for variable selection with random forests is introduced by Diaz-Uriarte and Alvarez de Andrés (2006), who suggest a backward elimination strategy based on the variable importance scores that takes under consideration the prediction accuracy: The underlying rationale is that the prediction accuracy will remain almost constant when irrelevant predictor variables are excluded, while it drops when relevant ones are excluded.

While in statistical modelling the aim may often be to select a model as sparse as possible, it is of equal interest in many applied sciences to be able to identify *all* predictor variables that are associated with the response, even if some of them are correlated. The question of interest here is to decide for each variable whether or not its importance is significantly greater than zero. A statistical test for this question is suggested by Breiman and Cutler (2008). At first sight it looks like this test could aid the decision which or how many of the top-ranked variables have significant importance and can be considered relevant. However, in the following we will present statistical reasoning and simulation results illustrating that the suggested test is not appropriate for statements of significance. Moreover, we will explore the unclear null hypothesis of the suggested test and give an outlook on a new permutation scheme for variable importance in random forests that better represents the null hypothesis of zero importance of a given variable.

2 Testing random forest variable importance

The rationale of the random forest permutation accuracy importance is the following: By randomly permuting the predictor variable X_j, its original association with the response Y is broken. When the permuted variable X_j,

together with the remaining non-permuted predictor variables, is used to predict the response for the out-of-bag observations, the prediction accuracy (i.e. the number of observations classified correctly) decreases substantially if the original variable X_j was associated with the response. Thus, a reasonable measure for variable importance is the difference in prediction accuracy before and after permuting X_j, averaged over all trees:

Let $\overline{\mathfrak{B}}^{(t)}$ be the out-of-bag sample for a tree t, with $t \in \{1, \ldots, ntree\}$. Then the variable importance for one tree is

$$VI^{(t)}(\mathbf{x}_j) = \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_i^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|} - \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_{i,\pi_j}^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|}$$

where $\hat{y}_i^{(t)} = f^{(t)}(\mathbf{x}_i)$ is the predicted classes for observation i before and $\hat{y}_{i,\pi_j}^{(t)} = f^{(t)}(\mathbf{x}_{i,\pi_j})$ is the predicted classes for observation i after permuting its value of variable j, i.e. with $\mathbf{x}_{i,\pi_j} = (x_{i,1}, \ldots, x_{i,j-1},\ x_{\pi_j(i),j}, x_{i,j+1}, \ldots, x_{i,p})$. (Note that $VI^{(t)}(\mathbf{x}_j) = 0$ by definition, if variable X_j is not in tree t.) The raw variable importance score for each variable is then computed as the mean importance over all trees:

$$VI(\mathbf{x}_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(\mathbf{x}_j)}{ntree}$$

Because the individual importance scores $VI^{(t)}(\mathbf{x}_j)$ are computed from *ntree* independent bootstrap samples, a simple test for the relevance of variable X_j can be constructed based on the central limit theorem for the mean importance $VI(\mathbf{x}_j)$. If each individual variable importance $VI^{(t)}$ has standard deviation σ, the mean importance from *ntree* replications has standard error $\sigma/\sqrt{ntree}$. Therefore, under the null hypothesis of zero variable importance, the z-score

$$\widetilde{VI}(\mathbf{x}_j) = \frac{VI(\mathbf{x}_j)}{\frac{\hat{\sigma}}{\sqrt{ntree}}}$$

is asymptotically standard normal. Hence, when the z-score $\widetilde{VI}(\mathbf{x}_j)$ exceeds the α-quantile of the standard normal distribution, the null hypothesis of zero importance for variable X_j is rejected. This approach has been suggested by Breiman and Cutler (2008) for testing the variable importance. Note, however, that in the computation of the z-score averaging and scaling is not conducted with respect to the sample size n but to the number of trees in the ensemble *ntree* (cf. also Lunetta et al. (2004)).

2.1 Investigating the power of the current test

To investigate the power of the test suggested by Breiman and Cutler (2008), that is outlined in the previous section, a simulation study was conducted.

The experimental parameters that were varied are (a) the relevance of the predictor variable, (b) the sample size, and (c) the number of trees in the forest. For each combination of experimental parameters 1000 replications were run. In each replication a data set with the respective relevance and sample size was generated, a random forest with the respective number of trees was fit to the data, and the z-score was computed as described in the previous section. The test decision, i.e. whether or not the null hypothesis was rejected, was stored in every replication. The relative frequency of rejections of the null hypothesis (out of the 1000 replications) serves as an estimator for the power of the test in each combination of experimental parameters. In Figure 1 the empirical power is displayed as a function of the experimental parameters.

For a deeper understanding of the underlying mechanism we also display the curves for the unstandardized mean importance VI, the standard error of the mean and the z-score $\widetilde{VI}$ (all averaged over 1000 replications). In each iteration, a data set of sample size $n = 100, 200$ or 500 is generated that includes five predictor variables of which only one binary variable is relevant. Within the categories of this variable the binary response class is sampled from a binomial distribution with class probability $0.5 \pm \rho$, where ρ is the relevance parameter ($\rho = 0, 0.05, \ldots, 0.5$) as indicated on the abscissas of Figure 1. The parameter settings for the random forests were given by the varying number of trees ($ntree = 100, 200$ or 500) and a fixed number of two preselected variables per split. The simulation was conducted with the function *randomForest* (from the package of the same name by Breiman et al. (2007), Liaw and Wiener (2002) give an introduction), which is the reference implementation of random forests in the R system for statistical computing (R Development Core Team (2007)).

As depicted in the bottom row of Figure 1 the power of the test against the null hypothesis of zero importance shows the following irritating behavior: The power does increase with the relevance of the predictor variable as expected for any reasonable power curve. However, the power also does increase with the number of trees in the forest (the curves are shifted to the left, resulting in higher power for low relevance values), meaning that the power here depends on a tuning parameter that can be arbitrarily chosen by the user. This effect is due to the construction of the test statistic where, unlike in the standard test for the mean under normality, averaging and scaling is not with respect to a given sample size n but to the number of trees as outlined above. Even more dramatically, we find that the power does depend on the sample size—however not as expected for any reasonable test, where the power is supposed to increase with increasing sample size, but to the contrary: For large sample sizes (as compared to the number of trees) the power is zero.

To explore in more detail the mechanism responsible for this odd behavior we will follow the construction of the z-score, that is derived from the mean

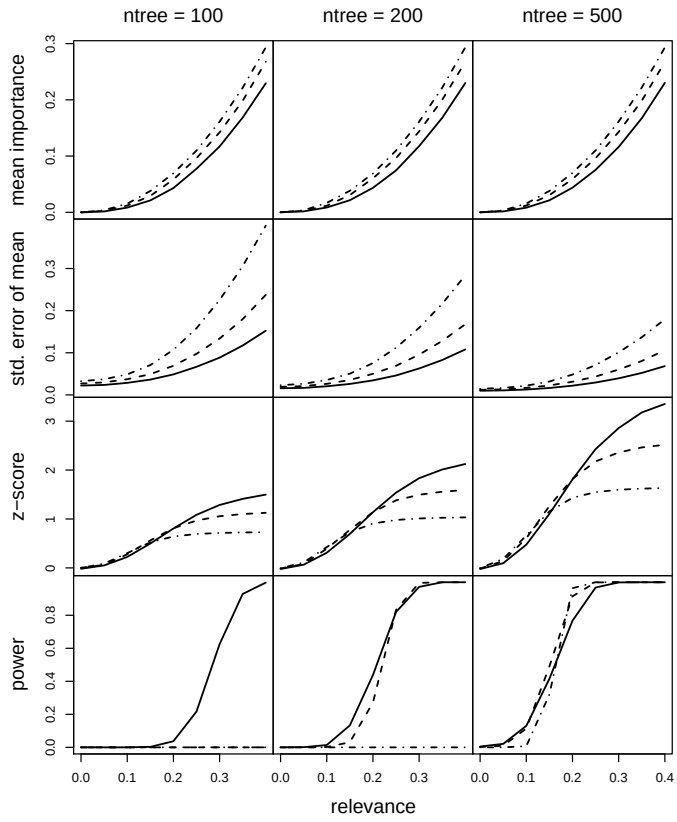

Fig. 1. Mean variable importance, standard error of mean, z-score and power as functions of relevance for sample size 100 (solid), 200 (dashed), and 500 (dash-dotted) and different numbers of trees.

importance by division through the standard error of the mean. The top row of Figure 1 shows that the unstandardized mean importance VI for one predictor variable increases with the relevance of the predictor variable and with the sample size as expected. There is no effect of the number of trees on the average importance—at least not when the number of trees is chosen sufficiently large to guarantee a stable estimate of the mean importance. This increase in the relevance and the sample size is desirable and exactly what we would have expected for any statistic to be employed in a test against the null hypothesis of zero importance. Therefore, the standard error of the mean, which is used for scaling, must be responsible for the odd behavior of the z-scores: The numerator of the fraction for the standard error of the

mean, the standard deviation, also increases with the relevance and with the sample size, and does not depend on the number of trees either. (The increase in the sample size is due to the resulting increase in the out-of-bag sample size that again extends the range of possible changes in the prediction accuracy induced by permuting the predictor variable. The dependence on the relevance is caused by a mechanism in the tree-building process: In many trees of the ensemble a variable with a low relevance may not be included at all, and produce an importance score of exactly zero, which diminishes the variation of the importance.) As a result of the division by the square root of the number of trees, however, an additional dependence on the number of trees is induced in the standard error of the mean, such that it decreases in the number of trees as depicted in the second row of Figure 1. Note also that the curves for the different sample sizes vary more strongly for the standard error of the mean than for the mean importance.

When finally the z-score is computed by means of standardizing the mean importance with the standard error of the mean, the rationale of this standardization is to account for the fact that the mean importance is an average over all trees in the ensemble—it does, however, not account for the effect of the sample size. The fact that the dependence of the mean importance on the sample size is less pronounced than that of its standard error causes an inversion of the importance pattern with respect to the sample size in the z-scores: We find in the third row of Figure 1 that the z-score decreases in the sample size but increases with the number of trees. This finally leads to the pattern for the power curves that we found in the bottom row of Figure 1: Only for high numbers of trees the overall level of the scaled importance is high enough for all sample sizes to ever reject the null hypothesis, while for lower numbers of trees the curves for the high sample sizes never exceed the threshold for rejecting the null hypothesis and result in a power of zero. This behavior is undesired and is an artefact of the scaling, that induces a dependence on the number of trees but at the same time inverts the dependence on the sample size. We therefore summarize the results of our simulation study that the mean variable importance VI shows the increase in the relevance and sample size that would be desired for a test for the null hypothesis of zero importance, while the scaled variable importance and the resulting test behave oddly.

2.2 Specifying the null hypothesis

Another issue when considering the test for the random forest permutation importance suggested by Breiman and Cutler (2008) is the very fundamental question: Exactly what null hypothesis is being tested? In the previous sections for simplicity we referred to the null hypothesis as "importance equal to zero". This implies some kind of independence between the predictor variable whose importance is being tested and the response. However, it is unclear

what kind of independence is being tested. The currently employed permutation scheme, where only the values of the variable of interest are permuted while the values of the response variable and the other predictors are held constant, does mimic the elimination of the predictor variable when predicting the response—however, at the same time it destroys all correlations between the variable of interest and the other covariates. Unlike standard permutation test of the global null hypothesis that the response is not correlated with any of the predictor variables, where the response is permuted against the complete predictor matrix and all associations within the predictor matrix are retained, the current random forest approach tests the rather unintuitive null hypothesis that the predictor of interest is not correlated with either one of the response or covariates. In cases where predictor variables may be correlated this permutation scheme might not reflect the actual null hypothesis of interest. This topic is investigated in more detail and a new, conditional permutation importance measure is suggested in Strobl et al. (2008).

3 Conclusion and outlook

We conclude that, in principle, a test for the random forest permutation importance could help identify relevant predictor variables. However, the results of our simulation studies also show that, in its current form, the test of Breiman and Cutler (2008) has prohibitively undesirable properties: The power of the test does not increase with the sample size, as would be expected for any reasonable statistical test, but rather remains zero for large sample sizes as compared to the number of trees. On the other hand the power does increase with the number of trees, which is a parameter that can be arbitrarily chosen by the user. This means that any statement of significance made with the current test for random forest variable importance is nullified.

Another issue, that is relevant in the context of correlated predictor variables, is the question whether the null hypothesis that is being tested in the current test is the one that reflects our understanding of the impact of a predictor variable on the response. A conditional permutation scheme that better reflects the null hypothesis of interest is suggested in Strobl et al. (2008).

Further research will address the issue of an adequate test statistic and rejection area for this null hypothesis. For high numbers of variables multiple testing issues will also have to be taken into consideration.

References

ARUN, K. and LANGMEAD, C. J. (2006): Structure based chemical shift prediction using random forests non-linear regression. In: T. Jiang, U. C. Yang, Y.-P. P. Chen and L. Wong (Eds.), *Proceedings of the Fourth Asia-Pacific Bioinformatics Conference*, Taipei, Taiwan, 317–326.

BREIMAN, L. (2001): Random forests. *Machine Learning 45 (1), 5–32.*

BREIMAN, L. and CUTLER, A. (2008): Random forests – Classification manual. URL `http://www.math.usu.edu/~adele/forests/`.

BREIMAN, L., CUTLER, A., LIAW, A. and WIENER, M. (2007): Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.5-22. URL `http://CRAN.R-project.org/`.

BUREAU, A., DUPUIS, J., FALLS, K., LUNETTA, K. L., HAYWARD, B., KEITH, T. P. and EERDEWEGH, P. V. (2005): Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology 28 (2), 171–182.*

DIAZ-URIARTE, R. and ALVAREZ DE ANDRES, S.. (2006): Gene selection and classification of microarray data using random forest. *BMC Bioinformatics 7:3.*

HUANG, X., PAN, W., GRINDLE, S., HAN, X., CHEN, Y., PARK, S. J., MILLER, L. W. and HALL, J. (2005): A comparative study of discriminating human heart failure etiology using gene expression profiles. *BMC Bioinformatics 6:205.*

LIAW, A. and WIENER, M. (2002): Classification and regression by randomForest. *R News 2 (3), 18–22.* URL `http://CRAN.R-project.org/ doc/Rnews/`.

LUNETTA, K. L., HAYWARD, L. B., SEGAL and J., EERDEWEGH, P. V. (2004): Screening large-scale association study data: Exploiting interactions using random forests. *BMC Genetics 5:32.*

QI, Y., BAR-JOSEPH, Z. and KLEIN-SEETHARAMAN, J. (2006): Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins 63 (3), 490–500.*

R DEVELOPMENT CORE TEAM (2007): *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

STROBL, C., BOULESTEIX, A.-L., ZEILEIS, A. and HOTHORN, T. (2007): Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics 8:25.*

STROBL, C., BOULESTEIX, A.-L., KNEIB, T., AUGUSTIN, T. and ZEILEIS, A. (2008): Conditional variable importance for random forests. *Technical Report 23*, Department of Statistics, Ludwig-Maximilians-Universität München, URL `http://epub.ub.uni-muenchen.de/2821/`.

WARD, M. M., PAJEVIC, S., DREYFUSS, J. and MALLEY, J. D. (2006): Short-term prediction of mortality in patients with systemic lupus erythematosus: Classification of outcomes using random forests. *Arthritis and Rheumatism 55 (1), 74–80.*

ZOU, H. and HASTIE, T. (2005): Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B 67 (2), 301–320.*