Artificial Representative Trees as Interpretable Surrogates for Random Forests

Lea Louisa Kronziel

Institute of Medical Biometry and Statistics, University of Luebeck, University Hospital Schleswig-Holstein - Campus Luebeck, Ratzeburger Allee 160, V24, 23562 Lübeck, Germany

Abstract

Random forests (RFs) are a popular machine learning method with good prediction performance, but difficult to interpret due to their structure as an ensemble method. To interpret RFs, decision trees can be used as surrogate models to preserve the tree structure. Alternatively, a tree of the RF can be selected as a surrogate model, whereby a direct part of the RF is used and not a similar model. The most representative tree (MRT) of the RF is used for this, which is most similar to all other trees of the RF. However, MRTs have the potential for misinterpretation due to non-informative early splits. To overcome this, the research in my PhD thesis will focus on generating an algorithm for artificial representative trees (ARTs) and comparing them with MRTs and decision trees as surrogate models using simulation studies as well as benchmark data. The first results show a promising improvement in terms of predictive quality and interpretability when comparing ARTs and MRTs.

Keywords

Random Forest, Surrogate Model, Machine Learning, Interpretability, Most Representative Tree

1. Context and motivation

Random forests (RFs) are a well-known and efficient machine learning (ML) algorithm for creating predictive models, especially for tabular data [1]. For example, RFs can be used to analyze high-dimensional molecular or genetic data [2, 3] as well as to enable individualized treatment options for patients in the context of precision medicine [4]. They consist of an ensemble of decision trees [5], whereby the decisions of a single tree are understandable. However, it is difficult for a human to understand the decisions of the RF in detail, which is why RFs are often referred to as black box models. Despite good prediction performance, this can be a barrier to the use of these methods in practice [6].

There are various approaches to make the decisions of such complex models understandable to enable an interpretation of the model. This includes understanding the individual predictions and which variables influence these predictions. For example, post hoc approaches such as partial dependence plots can be used to determine which variables influence the prediction of a model. In addition, variable importance can measure how important the variables are for the prediction performance [7]. Alternatively a surrogate model that is easier to interpret such as a decision tree can be developed instead. To ensure that the surrogate model is as similar as

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

l.kronziel@uni-luebeck.de (L. L. Kronziel)
0000-0002-2126-6831 (L. L. Kronziel)

possible to the original model, it is usually trained to make the same predictions as the black box model as it was done in [8]. However, there are also other approaches for using decision trees as surrogates. For example, a locally adapted decision tree is created in [9], which can be used to explain the prediction of a specific observation.

If a RF is to be interpreted, a decision tree as a surrogate model has the advantage that the tree structure is preserved. However, surrogate models that are trained for high predictive similarity cannot be guaranteed to actually use the same decisions as the original model. Instead of training a surrogate model it is suggested to select one decision tree from the ensemble of the RF to be interpreted as a surrogate model [10]. The decision tree that represents the RF best should be used and is therefore referred to as the most representative tree (MRT). This method has the particular advantage that the structure of the surrogate model and its decisions correspond directly to a part of the RF and are not just similar to it. MRTs therefore combine the predictive performance and interpretability of decision trees with the stability of RF. They also offer the advantage that they can be used more easily for external validation. A single MRT can be printed in a publication, while an RF can only be made available as an object of the programming language used or via a website interface.

2. Key related work

The idea of using a decision tree of the RF ensemble as a surrogate model was first reported in [10]. In their approach, the MRT is selected as the tree that is most similar on average to the other trees in the ensemble. Three distance metrics are proposed to calculate the pairwise similarities between the decision trees. For example, the difference in the predictions for a test data set is calculated for each pair of trees. Alternatively, the similarity can be measured by whether the observations of a test data set are assigned to the same terminal nodes of two trees. The third one measures pairwise similarity via the proportion of split variables used in both trees. However, it only focuses on the proportion of split variables used in both trees, but not where they are used in the tree. Whether a variable was used as the first split variable in one tree and the last in the other is not taken into account. In addition, it ignores if a split variable is used more than once in the trees.

However, these two aspects are are assigned to in the further development of this measure in [11]. Depending on the position of the split variable, its influence on the similarity is weighted, thus it is called weighted splitting variables (WSV). In [11], the three measures from [10] and the WSV measure were used for the selection of an MRT and their performance was compared. The results showed that the structure of the RF can be best represented with the WSV measure, as the predictions on a validation data set were most similar to the ones from the RF.

In [12], MRTs are also selected, but the authors suggest that in some cases it is better to use more than one MRT as a representation for an RF. To obtain this small ensemble of representative trees, the pairwise distances of the decision trees are clustered using the partitioning around medoids (PAM) algorithm. To do this, it is necessary to specify in advance how many clusters and thus MRTs are to be found. In addition, the aspect of interpretability was not investigated and no analyses were performed to determine whether the prediction quality actually changes depending on the number of MRTs.

However, MRTs have a disadvantage. When creating the RF, not all variables from the training data set are available to the trees at each split [5]. A random subset of the variables is drawn in each node, which means that potentially at some splits only noise variables are available for splitting. This can result in uninformative splits that do not improve the prediction quality of the RF but lead to deeper trees than necessary. Such uninformative splits can also occur in the selected MRTs. In addition, important variables are not necessarily used as top splits at the root. Decision trees and thus MRTs are easier for a human to interpret if they only consist of a few splits. To overcome this problem, artificial representative trees (ART) should be created.

3. Specific research questions, hypothesis and objectives

Resulting from the open topics from the previous section, my research will focus on the following four objectives:

1. To develop and evaluate an algorithm that generates an ART based on an existing RF. Analogous to the MRT, a surrogate model is to be created which can be interpreted instead of the RF.

   To evaluate if ARTs provide better interpretability compared to MRTs. The hypothesis is that ARTs are not as deep as MRTs, use a larger proportion of effect variables and a smaller proportion of null variables while having comparable prediction performance. (WP 1 & WP 2)
2. To compare ARTs with other surrogate models regarding to the same criteria as for the first objective, to assess if the higher effort for generating ARTs compared to classical surrogate models is worthwhile. I will also investigate which of the methods should be preferred for which type of data set. (WP 3)
3. To compare ensembles of ARTs and MRTs, based on the clustering approach from [12] with regards to the same criteria as the previous objectives. (WP 4)
4. To investigate whether other tree algorithms than the classic CART from [13] can improve prediction performance and interpretability of ARTs. Binary splits are often used in RF, as splits can be concatenated to any depth. However, trees with more than two splits in each layer are easier to understand than a deep concatenation of several splits. (WP 5)

4. Research approach, methods, and rationale for testing the research hypothesis

The following five work packages (WP) are defined to achieve the four objectives above.

WP 1 To create an ART, a new decision tree is grown iteratively using a greedy algorithm. First, all stumps that are possible with the available training data are created. The similarity to the RF is calculated for all stumps and the one with the greatest similarity to the RF is used. Analogous to MRT, various measures can be used to define similarity, such as the split variables used or the prediction errors. Then, in each additional iteration, all trees are created that are possible with exactly one more split. If one or more trees fulfill these criteria, the one with the greatest improvement in similarity or prediction is selected and a new iteration is started. If none of these trees improves the similarity or does not improve the prediction if the similarity remains the same, the algorithm stops.

WP 2 To compare the use of a single ART with a single MRT, I will first perform simulation studies. The advantage of simulation studies is that the relationships between the predictor variables and with the target variable are fully known. For the performance comparison of an ART and an MRT, I will initially consider only regression problems using the same structure as in [11]. In the first scenario, the data set consists exclusively of binary variables, with a small number of effect variables with large effects. The other scenarios represent variations of this, for example by using many effect variables with lower effect sizes, correlated variables, and interaction effects. The last scenario finally uses continuous variables. As quality measures, I will compare the prediction performance and consider the split variables used as well as the tree depth. The deviation of the predictive performance of the ARTs and MRTs from the RF is calculated using the MSE. In addition, it is measured how many of the splits use noise variables, which is called the false discovery rate (FDR). It is also measured how many effect and noise variables are used as split variables, as well as the runtime. Afterward, ARTs and MRTs will be compared with a benchmark data set from OpenML (https://openml.org/). Analogous to the simulation study, the deviation of the MSE and the tree depth are measured. In addition, the R2 and the Akaike information criterion (AIC) are estimated.

WP 3 I will perform extensive simulation studies analogous to WP 2 to compare ARTs with a decision tree as a surrogate model. Then the ARTs and decision trees will be applied to clinical or benchmark data sets. I will enlarge the simulation studies to cover more complex designs (e.g. classification problems and high-dimensional data) so that they are more similar to clinical use cases. For the performance comparison, I will additionally focus on the stability of the results by ARTs and decision trees. For this, I will compare the similarity of several ARTs and decision trees that were created on the same data. The various measures from [10] and WSV from [11] will be used as similarity measures so that the similarity is evaluated concerning various aspects.

WP 4 For the comparison of ensembles of MRTs and ARTs based on a clustering of the trees of the RF, I will first extend the approach of [12]. For example, I will integrate an automatic selection of the number of clusters using the improvement in prediction quality. As long as the prediction quality increases by adding a cluster and thus a representative tree, the number of clusters will be further increased. For the simulation study and benchmark data application, data sets containing latent subgroups will be used, as in these cases it is assumed that an ensemble of representative trees is more suitable than a single one. The remaining procedure for the simulation will be done in the same way as for WP 2. This will also increase the focus on predictive quality and interpretability as quality measures compared to [12] for MRTs.

WP 5 To obtain ARTs that do not only split binary, the ART algorithm from WP 1 should be extended. For example, splitting with the same variable several times in succession could be favored by a higher weighting. As soon as the split variable of a node is used a second time for splitting in the child node, the two nodes can be combined into a single node with more than two child nodes. I will again investigate the performance of this approach using simulations. In addition, I will vary the hyperparameter for weighting the repeated splits with the same variable to examine its influence on the quality criteria mentioned in WP 2. The tree depth and the number of terminal nodes are also compared.

5. Results and contributions to date

The aim of developing an ART algorithm was successfully realized, which is shown in algorithm 1. We integrated the implementation of the algorithm into the R package *timbR* (https://github.com/imbs-hl/timbR), which is based on trees built with the R package *ranger* [14]. ARTs can now be used for global interpretation of the RF so that, for example, a physician can understand the model's predictions. ARTs also enable local interpretability, so that the individual decisions of the ART can be compared with known knowledge from the literature or can be discussed with a physician as is medically plausible. ARTs can also be used as a prediction model with the option of interpreting individual predictions.

I have performed the simulation study mentioned in WP 2 and the comparison of an ART and MRT using the benchmark data. The ART was superior in terms of interpretability and the use of fewer noise variables. In fig 1, it can be seen that the predictions of the ARTs were more simular to the RF than those of the MRTs. In addition, ARTs used almost no noise variables. However, ARTs are somewhat more conservative in the use of effect variables than MRTs (results not shown, but displayed in [15]).

The manuscript was accepted as a conference paper at the XAI-2024 conference under the title "Construction of artificial most representative trees by minimizing tree-based distance measures"[15]. This study was funded by the Medical Section of the University of Lübeck (J01–2024 to BL).

For WP 3, I have compared ARTs and decision trees as surrogate models in a few simple structured simulated scenarios. The MSE of the predictions and the FDR of the ARTs were smaller than those of the decision trees. In addition, the ARTs again consist mainly of effect variables, whereas the decision trees use a higher proportion of noise variables. Nevertheless, the predictions of the decision trees were more similar to those of the RFs than ones of the ARTs. However, the process is not finished yet. For the final simulation study, I will extend the simulated scenarios to several different ones and will focus on the structure of more complex clinical data. For example, I will use gene expression data as a possible application example. In addition, I will investigate both classification and regression problems.

For the ensembles of representative trees in WP 4, we compared various clustering methods such as k-means or hierarchical clustering using the ward method with simulations using different numbers of MRTs. This was done as part of a master's thesis that I co-supervised. The most stable results in terms of prediction quality were provided by k-means. In addition, we used various similarity measures, of which WSV from [11] provided the best prediction quality.

6. Expected next steps and final contribution to knowledge

The comparison between single ARTs and MRTs has been completed. ARTs were found to be a better alternative to MRTs, as their results are better and easier to interpret. The code for the

Algorithm 1 Generate ART

```
Require: random forest *RF*, similarity metric *metric*
  Extract all *split_points* from *RF*
  Reduce *split_points* using only important variables
  Build all possible stumps using *split_points*
  Estimate similarities of all stumps to *RF* using *metric*
  Select stump with maximum similarity $\rightarrow$ *ART_candidate*
  repeat
    *ART* $\leftarrow$ *ART_candidate*
    Build all possible trees with one additional split using *split_points*
    Estimate similarities of all new trees to *RF* using *metric*
    Select new tree with maximum similarity $\rightarrow$ *ART_candidate*
  until similarity(*ART_candidate*) $<$ similarity(*ART*)
  return *ART*
```

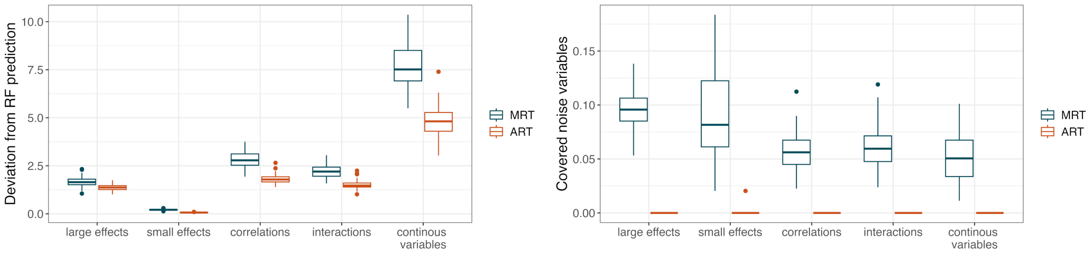

Figure 1: Comparison of the performance of ARTs (orange) and MRTs (blue) for all five simulation scenarios in 100 repetitions. The deviation from the prediction performance of the RF was measured using the MSE. The fractions of covered noise variables were calculated by the number of noisevariables that occur at least once in the surrogate model divided by the number of all simulated noise variables.

simulation as well as for the ART algorithm is freely available to enable other scientists to use it easily (https://github.com/imbs-hl/ART_paper; https://github.com/imbs-hl/timbR).

Next, I will extend the performance and interpretability comparison of ARTs with decision trees as surrogate models to identify the advantages and disadvantages of both methods in different scenarios.

Additionally, I will integrate the use of ensembles of ARTs into the R package *timbR* and carry out the planned comparison with MRTs. We assume that prediction performance and interpretability can be further improved through the use of ARTs.

Furthermore, I will further improve the performance by extending the CART based ART algorithm so that a very easy-to-interpret model is available for a wide variety of data structures, which also has a good prediction quality.

Finally, I will apply ARTs in ongoing collaborative research projects in Neuro- and Cardiogenetics to provide interpretable models for the clinical context. For example, to use an ART to investigate the influence and interaction of genetic variants in prediction modeling of age at onset in X-linked dystonia-parkinsonism.

In summary, the use of ARTs offers promising opportunities to develop interpretable models for the clinical context, and further research will lead to a surrogate model that is easy to use and interpret.

Acknowledgments

Thanks to Prof. Silke Szymczak and Dr. Björn-Hergen Laabs for their supervision and support with my doctoral thesis.

References

[1] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on tabular data?, 2022. URL: http://arxiv.org/abs/2207.08815. doi:10.48550/arXiv.2207.08815, arXiv:2207.08815 [cs, stat].

[2] X. Chen, H. Ishwaran, Random forests for genomic data analysis, Genomics 99 (2012) 323–329. URL: https://www.sciencedirect.com/science/article/pii/S0888754312000626. doi:10.1016/j.ygeno.2012.04.003.

[3] M. S. O. Brieuc, C. D. Waters, D. P. Drinan, K. A. Naish, A practical introduction to Random Forest for genetic association studies in ecology and evolution, Molecular Ecology Resources 18 (2018) 755–766. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12773. doi:10.1111/1755-0998.12773, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12773.

[4] K. Doubleday, J. Zhou, H. Zhou, H. Fu, Risk controlled decision trees and random forests for precision Medicine, Statistics in Medicine 41 (2022) 719–735. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9253. doi:10.1002/sim.9253, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.9253.

[5] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[6] G. Heinze, C. Wallisch, D. Dunkler, Variable selection – A review and recommendations for the practicing statistician, Biometrical Journal 60 (2018) 431–449. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201700067. doi:10.1002/bimj.201700067, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.201700067.

[7] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2 ed., 2022. URL: https://christophm.github.io/interpretable-ml-book.

[8] U. Johansson, C. Sönströd, T. Löfström, One tree to explain them all, in: 2011 IEEE Congress of Evolutionary Computation (CEC), 2011, pp. 1444–1451. URL: https://ieeexplore.ieee.org/abstract/document/5949785. doi:10.1109/CEC.2011.5949785, iSSN: 1941-0026.

[9] E. Parimbelli, T. Buonocore, G. Nicora, W. Michalowski, S. Wilk, R. Bellazzi, Why did AI get this one wrong? - Tree-based explanations of machine learning model predictions, Artificial Intelligence in Medicine 135 (2022) 102471. doi:10.1016/j.artmed.2022.102471.

[10] M. Banerjee, Y. Ding, A.-M. Noone, Identifying representative trees from ensembles, Statistics in medicine 31 (2012) 1601–16. doi:10.1002/sim.4492.

[11] B.-H. Laabs, A. Westenberger, I. R. König, Identification of representative trees in random forests based on a new tree-based distance measure, Advances in Data Analysis and Classification (2023). URL: https://doi.org/10.1007/s11634-023-00537-7. doi:10.1007/s11634-023-00537-7.

[12] A. Sies, I. Van Mechelen, C443: a Methodology to See a Forest for the Trees, Journal of Classification 37 (2020) 730–753. URL: https://doi.org/10.1007/s00357-019-09350-4. doi:10.1007/s00357-019-09350-4.

[13] L. Breiman, J. Friedman, C. Stone, R. Olshen, Classification and Regression Trees, Taylor & Francis, 1984. URL: https://books.google.de/books?id=JwQx-WOmSyQC.

[14] M. N. Wright, A. Ziegler, ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, Journal of Statistical Software 77 (2017) 1–17. URL: https://www.jstatsoft.org/index.php/jss/article/view/v077i01. doi:10.18637/jss.v077.i01.

[15] B.-H. Laabs, L. L. Kronziel, I. R. König, S. Szymczak, Construction of Artificial Most Representative Trees by Minimizing Tree-Based Distance Measures, in: L. Longo, S. Lapuschkin, C. Seifert (Eds.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2024, pp. 290–310. doi:10.1007/978-3-031-63797-1_15.