Journal of Classification (2020) 37:730–753
https://doi.org/10.1007/s00357-019-09350-4

C443: a Methodology to See a Forest for the Trees

Aniek Sies[1] · **Iven Van Mechelen**[1]

Published online: 7 January 2020
© The Classification Society 2020

Abstract
Often tree-based accounts of statistical learning problems yield multiple decision trees which together constitute a forest. Reasons for this include examining tree instability, improving prediction accuracy, accounting for missingness in the data, and taking into account multiple outcome variables. A key disadvantage of forests, unlike individual decision trees, is their lack of transparency. Hence, an obvious challenge is whether it is possible to recover some of the insightfulness of individual trees from a forest. In this paper, we will propose a conceptual framework and methodology to do so by reducing forests into one or a small number of summary trees, which may be used to gain insight into the central tendency as well as the heterogeneity of the forest. This is done by clustering the trees in the forest based on similarities between them. By means of simulated data, we will demonstrate how and why different similarity types in the proposed methodology may lead to markedly different conclusions, and explain when and why certain approaches may be recommended over other ones. We will finally illustrate the methodology with an empirical data set on the prediction of cocaine use on the basis of personality characteristics.

Keywords Classification trees · Statistical learning · Bagging · Ensemble methods · Clustering

1 Introduction

Decision trees are widely used to solve statistical learning problems in many areas of research. Such trees recursively partition the predictor space, resulting in a set of rectangular regions based on dichotomized predictor variables (Breiman et al. 1984; Quinlan 1986). In the case of a continuous response variable, each rectangular region is associated with a predicted response value, and the whole may be referred to as a regression tree; if the response variable is discrete, each rectangular region is associated with a class label, and the whole

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00357-019-09350-4) contains supplementary material, which is available to authorized users.

✉ Aniek Sies
aniek.sies@kuleuven.be

[1] Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102 - Box 3713, Leuven, 3000, Belgium

is called a classification tree. In the present paper, we will only focus on the latter type of trees. A key attractive feature of decision trees in general is that they represent the relations between the response and predictor variables in a highly transparent and insightful way.

Nowadays, however, to solve a statistical learning problem, researchers often rely on a set of trees (i.e., a forest) rather than on a single tree. They do so for a variety of reasons: First, it is well-known that single decision trees are unstable with respect to small changes in the learning data (Breiman 1996b; Strobl et al. 2009; Turney 1995). As such, it may be interesting to explore this instability, for example, by growing and comparing trees based on different parts of the data, or based on different tuning parameters or algorithms. Second, it is well-known that growing multiple trees based on different subsets of the experimental units and/or the predictor variables and subsequently combining their predictions (making use of tools such as random forests (Breiman 2001), bagging (Breiman 1996a), and boosting (Freund and Schapire 1997)) may result in more accurate predictions than using single trees to do so Dietterich (2000); Bauer and Kohavi (1999); Skurichina and Duin (2002); Hastie et al. (2009). Hence, if the main goal is to make accurate predictions, a forest of trees may be more suitable than a single tree. Third, missing data constitute a frequently encountered problem in many research areas. Various approaches are available to handle this Little et al. (2012); Schafer and Graham (2002), with, however, multiple imputation (Rubin 1987) generally being considered as one of the best ways to do so. This procedure results in a number of imputed data sets, on each of which the statistical method of choice is to be applied. In our case, this will result once more in a forest of trees. Fourth, often, multiple response variables are of interest. Currently, within a tree context the most straightforward way to account for this, is to grow a separate decision tree for each outcome variable. The result will again be a forest of trees.

Despite many good reasons for using a forest rather than a single tree, this goes with a large cost: By using a forest, the most attractive feature of single trees, namely their insightfulness, gets lost. Therefore, a pressing question reads whether it could be possible to enjoy the benefits of a forest, without having to pay this cost. This comes down to the challenge to recover at least part of the insightfulness of single trees from a forest.

Three aspects of gaining insight may be most relevant at this point: First, one may wish to find out whether a forest can be summarized in one or a few central decision structure(s), and, if yes, how these look like; we will refer to this as capturing the central tendency of the forest. Second, one may wish to describe the variability within the forest; this may include the question as to whether all trees in the forest are slight variations on one central decision structure, versus whether the trees under study are subject to sizeable qualitative differences; we will further refer to this as describing the heterogeneity of the forest. Third, in case of a heterogeneous forest that came about on the basis of some known source of variation (such as, e.g., different types of tuning parameters or different types of response variables), one may wish to find out whether qualitative within-forest differences relate to this source of variation. For example, in the case of multiple response variables, one may wish to learn in which respect different response variables do not versus do result in different decision structures.

Several strategies have been proposed in the literature to gain some form of insight into a forest (Chipman et al. 1998; Banerjee et al. 2012; Briand et al. 2009). All of them are based on calculating similarities between the trees in it, and subsequently selecting one or a small number of summary trees on the basis of these similarities. More specifically, both Briand et al. (2009) and Banerjee et al. (2012) select the tree(s) with the highest average similarity with all other trees as the summary tree(s). Chipman et al. (1998) use the similarities to represent the trees in a low dimensional space, next perform an informal clustering of the trees in this space, and finally select a summary tree within each cluster.

These previously proposed strategies partially address the pursuit for insight formulated above. Yet, they also leave quite a number of issues unsolved: First, the similarity measures they rely on were chosen ad hoc, in absence of a clear overview of the full range of measures that may be relevant within a tree context. Second, Banerjee et al. (2012) and Briand et al. (2009) select the most central tree (or a number of central trees) from the forest as a whole; consequently, their procedures address the aspect of the central tendency of the forest only. In contrast, Chipman's method implies a clustering and a subsequent selection of a summary tree within each cluster; yet, both the clustering and the selection of the summary trees are performed in an informal way only; moreover, Chipman's approach does not allow to address the question of how the heterogeneity within a forest may relate to known sources of variation.

In this paper, we propose a comprehensive methodology to gain the three aspects of insight looked for, while overcoming the limitations of the previously proposed strategies. This methodology has been implemented in the R-package C443, which can be downloaded from the Comphrehensive R Archives Network (CRAN) at https://CRAN.R-project.org/package=C443, or installed in R using `install.packages(C443)`. The methodology includes three components: (1) a calculation of similarities between all trees in the forest, relying on a thoughtful choice of a similarity measure based on a novel conceptual framework to get hold of the vast number of relevant possible measures within a tree context; (2) a clustering of the trees in the forest based on similarities, where, importantly, given our critical concern about insightfulness, clusters are obtained that are well interpretable in terms of the underlying decision structure and of the relation between predictors and response variable; for this purpose, we made an appeal to an existing clustering method, that allows for this kind of interpretation; we further applied it within the specific context of forests of decision trees; (3) a post-processing of the clustering result to arrive at the essential insights looked for, in which quite a number of challenging choices need to be made by the user; we propose to rely on a set of well-chosen tools that can be helpful in this regard.

The structure of the remainder of this paper is as follows. In Section 2, we will successively explain the three components of our newly proposed methodology to see a forest for the trees. In Section 3, we will we will demonstrate by means of simulated data how and why different similarity types in the proposed methodology may lead to markedly different conclusions, and we will explain when and why certain approaches may be recommended over other ones. Finally, in Section 4, we will illustrate the methodology by means of a dataset on drug use. The paper will end with some concluding remarks in Section 5.

2 Methodology

In this Section, we will introduce the concepts underlying our newly proposed methodology, by elaborating upon each of its three components in successive order.

2.1 Similarities

Similarity between trees can be regarded from different respects or points of view, which translates into different types of similarities and different associated similarity measures. In Section 2.1.1, we will elaborate upon different respects of similarity that are relevant within the context of decision trees, and we will present a well-structured overview of the most important conceptual distinctions between the different types of similarity. In Section 2.1.2, we will introduce a number of factors that may complicate the calculation of similarities.

Finally, in Section 2.1.3. we will introduce some existing similarity measures and propose a number of new ones, and we will characterize these in terms of the conceptual distinctions outlined in Section 2.1.1.

2.1.1 Respects and Specifications of Similarity

We may distinguish between two respects of similarity within the context of decision trees, namely similarity in terms of (1) the content of the trees, focusing on commonalities on the level of predictors included in them, and (2) the classifications of the objects or experimental units that the trees imply, focusing on the degree of agreement between the classifications in question. Additionally, the two respects can be combined, in that similarities may rely on a comparison of both the predictor-related content and the classifications implied by the trees under study. The two respects of similarity reflect the fundamental difference between the two perspectives from which categories or concepts can be looked upon, namely (1) the perspective of the characteristics or attributes of a category and (2) the perspective of the membership, that is, the objects that belong to the category.

Furthermore, within each respect of similarity, a number of conceptual distinctions can by made, resulting in different types of similarities. A researcher should make a choice for each of these to obtain a type of similarity that matches his/her interests. Below, we will discuss the conceptual distinctions and associated types of similarity within each respect of similarity.

Similarity of the Predictor-Related Content of a Tree With regard to the similarity of the predictor-related content of a tree, a first consideration is whether to focus on predictor variables only, versus on the predictor-split point combinations that are included in the trees. The first case implies types of similarity based on the number of common predictors in two trees, whereas the second case includes types based on common predictor-split point combinations. In the latter case, however, focusing on the number of identical predictor-split point combinations would, especially in the case of continuous variables, lead to very low similarities. For this reason, it may be better either to take into account the difference between two split points of identical predictors in a gradual way, or to specify a tolerance zone for each predictor, within which split points are assumed to be identical. As an example, trees (a) and (b) in Fig. 1 have predictors X_2 and X_5 in common. The differences in their split points are 2 and 0.2, respectively; these numbers could be taken into account in a gradual way. In the case that a tolerance zone is specified, these predictor-split point combinations are identical if the tolerance zones are larger than 2 and 0.2, for X_2 and X_5, respectively.

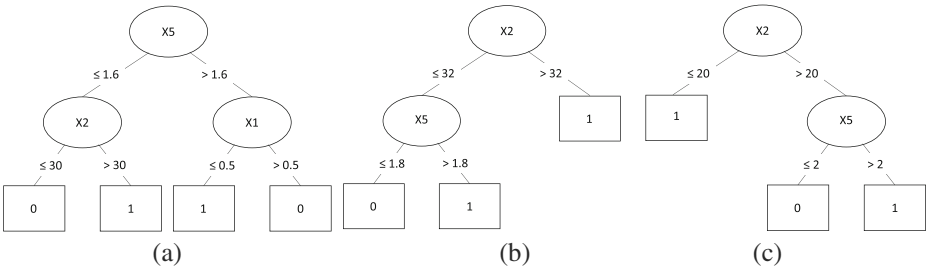


Fig. 1 Three examples of two-valued classification trees, with splits based on predictors X_1, X_2, and X_5. The squares are the leaves of the tree; each leaf contains the class label to which observations that end up in the leaf are assigned

A second consideration is whether to focus on individual predictors (or predictor-split point combinations) as we did in the previous paragraph, versus on *sets* of predictors (or predictor-split point combinations) implied by the leaves of the trees under study. As an example, the set of predictor-split point combinations implied by the leftmost leaf in tree (a) of Fig. 1 is $\{(X_5, 1.6), (X_2, 30)\}$, whereas for tree (b) it is $\{(X_2, 32), (X_5, 1.8)\}$.

In the case of comparing sets of predictors (or predictor-split point combinations) associated with leaves, one may further wonder whether or not to take the order of the predictors (splits) in a set into account as implied by the path from the root node to the leaf associated with that set. As an example, if the order is taken into account, the sets of predictor-split point combinations associated with the leftmost leaves in trees (a) and (b) in Fig. 1 are not identical.

A final consideration when comparing sets of predictor variables (or predictor-split point combinations), is whether or not the directions of the splits in the set, that is, $\leq$ or $>$, are taken into account. For example, if direction is taken into account, the set associated with the leftmost leaf of tree (b) in Fig. 1 is $\{(X_2, \leq), (X_5, \leq)\}$ in case of predictors, and $\{(X_2, 32, \leq), (X_5, 1.8, \leq)\}$ in case of predictor-split point combinations. Furthermore, the sets of predictors associated with this leaf and the middle leaf of tree (c) will be considered identical only if direction is not taken into account.

All distinctions with regard to the content of a tree are summarized in Fig. 2. The first distinction (in the top of the figure) is whether or not split points are to be taken into account, and, if so, whether this should be done in a gradual way versus by using tolerance zones. The second distinction is whether to focus on individual predictors versus on sets of predictors (or predictor-split point combinations). If the focus is on sets of predictors (or predictor-split point combinations), final considerations pertain to choices whether or not the order and/or directions of the splits should be taken into account. As further discussed below, different "leaves" in Fig. 2 will go with different types of similarity measures. In the remainder of this paper, we will refer to these similarity types by means of their corresponding number

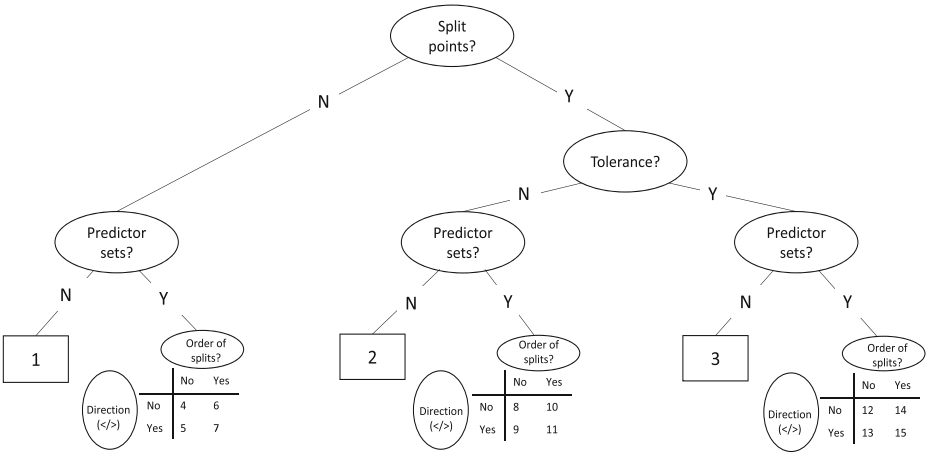


Fig. 2 Overview of types of similarity related to the predictive content of trees. The nodes on each level from top to bottom represent respectively. Whether split points are taken into account (versus predictors only); whether split points are taken into account using a tolerance zone (versus in a gradual way); whether predictor sets from root node to leaf are evaluated (versus individual predictors); whether the order of the splits is taken into account (versus not); whether the directions of the splits are taken into account (versus not). Each leaf represents the combination of distinctions from the root node to that leaf

in this figure. For example, similarity type 6 is based on ordered sets of predictors, while ignoring directionality of the splits and split points.

Similarity of the Partitions and Labellings Implied by a Tree Instead of evaluating the predictive content of trees, this respect of similarity is based on evaluating the extent to which the trees in question imply the same classifications of the experimental units or objects. Each classification tree implies two types of partitions of the experimental units, namely one in terms of the leaves, and one in terms of the class labels. When evaluating the similarity between two trees, one then may do so by examining whether pairs of objects that belong to a same partition class in the first tree also belong to a same class in the second tree (and vice versa). Additionally, while examining pairs of objects in this way, one may or may not take into account whether the class labels associated with the objects' partition classes are the same across the two trees. This leads to four possible types of similarity, as shown in Fig. 3. In the remainder of this paper, we will refer to these similarity types by means of their corresponding letter in this figure.

Similarity type A means that the partitions defined by the leaves of the trees under study are being compared while ignoring the specific labels associated with those leaves. This can be done by looking whether pairs of experimental units that are in a same leaf in one tree are also in a same leaf in the other tree and vice versa.

Similarity type B is based on a comparison of partitions in terms of the class labels implied by the trees while ignoring whether the labels in question are the same across the trees. This means that a pair of experimental units will contribute to a higher similarity if both units obtain the same class label in the two trees involved, irrespective of whether these labels are identical across the trees in question.

Similarity type C implies a comparison of the partitions in terms of the leaves of the trees, while also taking into account the class label associated with each leaf. This implies

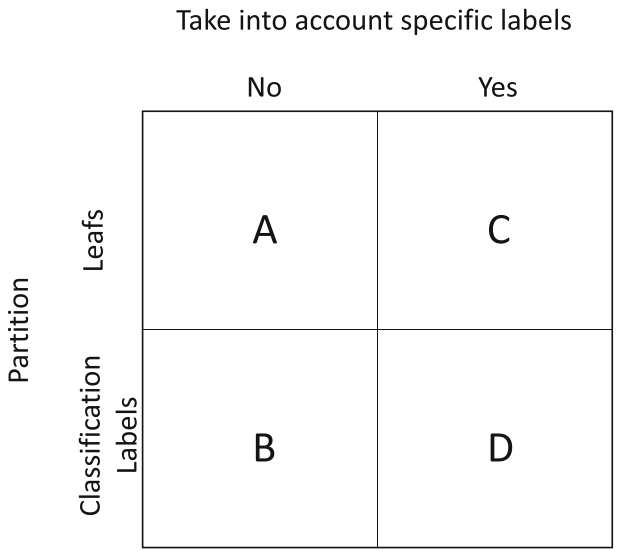


Fig. 3 Overview of types of similarity related to the partitions and labelings implied by a tree. The types of similarity result from a factorial combination of **a** looking at the partitions of two trees in terms of the leaves versus in terms of the class labels and **b** not taking into account versus taking into account the specific labels attached to each partition class

that a pair of experimental units will contribute to a higher similarity if both units are in the same leaf in the two trees under study, *and* if the two leaves are associated with the same class label.

At last, similarity type D is based on a comparison of the partitions defined by the class labels implied by the trees, while taking the specific value of the class labels into account. This means that a pair of experimental units will contribute to a higher similarity if both units have the same class label in the two trees under study, *and* if these class labels are the same in the two trees involved. This comes down to evaluating similarity on the level of the individual objects instead of the pairs of objects, that is, to evaluating the degree of agreement between the class labels assigned to the individual objects by the trees under comparison.

Similarity of the Content and Classifications of a Tree Finally, the predictive content of a tree and the implied classifications can be taken into account simultaneously. This can be achieved by evaluating similarity between trees in terms of leaves and their associated predictor set (see similarity types 4 through 15 in Fig. 2) only if they have the same associated class label. In case of two trees with exactly the same predictive content yet with different labels attached to the leaves, this type of similarity will be low, unlike in similarity types that take only the trees' predictive content into account.

2.1.2 Complicating Factors

When quantifying similarity between classification trees, three complicating factors play an important role. The first two are related to the fact that equivalence relations can be defined on families of trees, which raises the question of whether these should be taken into account or not in the calculation of similarities. The third complicating factor is that a predictor can be included in a tree multiple times, which implies the question of how to deal with this.

Logical Equivalence For simplicity's sake, in this subsection, we will focus on two-valued classification trees (with class labels 0 and 1), but a generalization to more than two classes is straightforward. A two-valued classification tree can be translated into a logical rule that indicates which experimental units will be assigned to class label 1. For example, tree (a) in Fig. 4 can be translated into the following logical rule: Class label 1 $\iff$ $\{(X_1 \leq$

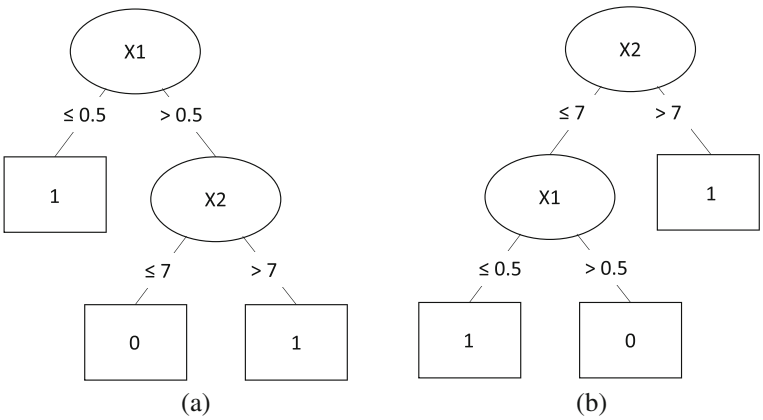


Fig. 4 Two trees that are different in terms of their topology, but that can be translated into logical rules that are logically equivalent

$0.5) \vee [(X_1 > 0.5) \wedge (X_2 > 7)]\}$. A complicating factor at this point is that two trees with a different topology can imply logically equivalent rules. For example, tree (b) in Fig. 4 implies a rule that is logically equivalent to that of tree (a). A special case of logical equivalence is the fact that trees have "rotational freedom," by which we mean that in each node of a tree an arbitrary decision has to be made as to which branch goes right and which one left. When calculating similarities between trees, an important question is whether two trees that imply logically equivalent rules, should be considered highly similar, despite possible differences in topology.

Depending on the type of similarity that is chosen, logical equivalence is taken into account or not. For example, if the logically equivalent trees (a) and (b) in Fig. 4 would be compared in terms of individual predictor variables (similarity type 1), they would be maximally similar. However, if they would be compared in terms of sets of predictors associated with the leaves, while taking into account order and direction, they would be highly dissimilar. Only similarity types 1, 2, 3, and D take into account logical equivalence.[1]

Empirical Equivalence In this subsection, we focus again, without loss of generality, on two-valued classification trees. Besides the possible logical equivalence of the rules into which trees can be translated, the rules may also be empirically equivalent, meaning that they assign exactly the same experimental units to each class label. For example, suppose that one tree can be translated into the rule: Class label 1 $\iff$ $(X_2 > 7)$, and another tree into: Class label 1 $\iff$ $(X_3 > 9)$, and that $X_3 = X_2 + 2$; in that case, the two rules under study will assign all experimental units to exactly the same class labels, and, hence, are empirically equivalent. Although exact empirical equivalence as in the example above is unlikely to occur often in practice, near empirical equivalence may be common. Near empirical equivalence means that the sets of experimental units that are assigned to each class label are nearly identical. An important question when quantifying similarities between trees is whether or not two trees that are different in terms of predictive content, yet that are (nearly) empirically equivalent, should be considered (highly) similar.

Similar to logical equivalence, some types of similarity take into account empirical equivalence, whereas others do not. For example, the two trees referred to above that can be translated into the rules Class label 1 $\iff$ $(X_2 > 7)$ and Class label 1 $\iff$ $(X_3 > 9)$, with $X_3 = X_2+2$, will be considered maximally similar if they are compared in terms of the class labels that they assign to individual experimental units (similarity type D). However, if the trees are compared in terms of the individual predictors included in them (similarity type 1), they will be considered highly dissimilar. None of the similarity types in Fig. 2 takes into account empirical equivalence, whereas all of the types in Fig. 3 do take it into account.

Multiple Splits on the Same Predictor The following consideration only applies to similarity types related to the predictive content of trees. If multiple splits in one tree are based on the same predictor (with identical or different split points), one may choose to take this into account or not. Taking multiple splits into account may be done in more or less involved ways. As an example, take similarity type 1. Not taking multiple splits on the same predictor into account, would imply binary decisions, where each predictor can either be included in both trees or not. Taking multiple splits into account, could, for example, be done by

[1]For some types of similarity that do not take logical equivalence into account, it is possible to rectify this by performing a manipulation on the trees under study. In the Supplementary Materials, we will explain how this can be done exactly, and introduce a similarity measure that can be used for this purpose (see Equation 15 of Supplementary Materials Section 2.2).

counting how often predictor k is included in tree (a) ($\#X_{k_{T_a}}$), and in tree (b) ($\#X_{k_{T_b}}$), and by subsequently defining a similarity as a function of the minimum of these two numbers ($\min(\#X_{k_{T_a}}, \#X_{k_{T_b}})$). For similarity type 3, it is more complicated to take multiple splits on the same predictor into account, because of the split points and tolerance zones involved. As a way out, some kind of optimal matching procedure could be used to find, for each predictor X_k, the largest possible bijection between a subset of the splits of X_k in tree (a) and a subset of compatible splits (in terms of tolerance zones) of X_k in tree (b), which may finally result in similarities based on the sum (across all predictors) of the cardinality of the domain of the bijection associated with each predictor.

2.1.3 Measures of Similarity

For each respect of similarity, we will now discuss a number of measures that have been proposed in the literature, followed by a conceptual introduction of some new measures (the mathematical formalization of which can be found in the Supplementary Materials). We will characterize the measures in terms of similarity types 1–15 and A–D, according to the principles on which they are based. Note that this overview is non-exhaustive. Also note that some of the measures discussed below are dissimilarities instead of similarities. All measures range between 0 and 1.

Similarity Measures Related to the Content of Trees

Existing Measures

Banerjee et al. (2012)
Banerjee et al. (2012) proposed a dissimilarity measure $d(T_1, T_2)$ based on a representation of the trees T_1 and T_2 under comparison as K-dimensional binary vectors (denoted by H^{T_1} and H^{T_2}), where K is the total number of predictors in the study, and where the kth element of vector H^T, denoted by H_k^T, indicates whether predictor X_k is included in tree T (with 0 = no and 1 = yes):

$$d(T_1, T_2) = \frac{\sum_{k=1}^{K} |H_k^{T_1} - H_k^{T_2}|}{K}. \tag{1}$$

Multiple splits on the same predictor are not taken into account by this measure, which is an instance of similarity type 1.

Shannon and Banks (1999)
The following dissimilarity measure was proposed by Shannon and Banks (1999):

$$d(T_1, T_2) = \sum_r \alpha_r |D^r_{T_1\ T_2}|, \tag{2}$$

where $|D^r_{T_1\ T_2}|$ is the number of distinctive paths of length r from the root node to a leaf that show up in only one of the trees T_1 and T_2, and α_r is a weight function that depends on r (e.g., when $\alpha_r = \frac{1}{r}$, shorter distinctive paths are penalized more heavily than longer ones). The paths are the sets of predictors implied by the leaves, taking into account the order of the predictors and the direction of the splits. As such, this measure is an instance of similarity type 7. This measure is not normalized (and, because of that, not comparable across different pairs of trees). Miglio and Soffritti (2004) proposed to normalize it

by dividing it by the maximum possible dissimilarity between the two trees under comparison (i.e., the dissimilarity between the trees if they would have different root nodes).

Briand et al. (2009)
Briand et al. (2009), proposed a dissimilarity measure $d(T_1, T_2)$ that takes into account the distance between split points of identical predictors in a gradual way. Suppose that two trees T_1 and T_2 have the same topology, then

$$F_v = I\{X_v^{T_1} = X_v^{T_2}\}\left(1 - \frac{|\delta_v^{T_1} - \delta_v^{T_2}|}{\text{range}(X_v)}\right), \tag{3}$$

where $v_0, v_1, ...v_V$ are the non-terminal nodes of the trees, numbered in descending order from root node to leaf, and from left to right, $I\{X_v^{T_1} = X_v^{T_2}\}$ is an indicator of the event that the split variables of the vth node are identical in both trees, and $\delta_v^{T_1}$ and $\delta_v^{T_2}$ are the split points of node v in T_1 and T_2, respectively. Then,

$$d(T_1, T_2) = 1 - \sum_{v=0}^{V} q_v F_v, \tag{4}$$

where $q_0, ..., q_V$ are user-supplied non-negative weights summing to 1. In case of two trees with different topologies, each time a leaf is encountered in a first tree at a position where the corresponding node is split in the other tree, a ghost branch of the same structure is added to replace the leaf in the first tree. In such a case, the terms in F_v pertaining to nodes in the ghost branch will be set to 0. The measure is an instance of similarity type 11.

New Measures

Common Predictors
In the measure of Banerjee et al. (2012), a predictor does not only contribute to a higher similarity if it is included in the two trees under comparison, but also if it is not included in either of them. As an alternative, the Jaccard index (1901) can be used as a similarity measure, which implies that only predictors that are included in both trees contribute to a higher similarity. This index is an instance of Similarity Type 1 (see Equation 1 of the Supplementary Materials).

Neither the measure of Banerjee et al. (2012), nor the Jaccard index take into account multiple splits on the same predictor. To do so, one could generalize both measures by looking at the frequencies with which a predictor is included in the two trees under study, rather than looking whether or not the predictor in question is included in the two of them. Formulas for the resulting generalized Banerjee and Jaccard measures can be found in Equations 2 and 3 of the Supplementary Materials.

Common Predictor-Split Point Combinations
If one wants to involve split points in the comparison of two trees (similarity type 3), an optimal matching procedure is needed to take into account the possibility of multiple splits on the same predictor X_k. The goal of this optimal matching procedure is to find the largest possible bijection between a subset of the splits of X_k in T_1 and a subset of compatible splits of that predictor in T_2, where splits are compatible if they lie within

the pre-specified tolerance zones of each other. The similarity between T_1 and T_2 then is the sum (over all predictors) of the cardinality of the domain of the bijection associated with each predictor (see Equation 6 of The Supplementary Materials).

Common Predictor Sets
A similarity measure that is an instance of type 4 or 5 focuses on the sets of predictors implied by a leaf, while or while not taking the directions of the splits into account. For every leaf in T_1, the similarity with every leaf of T_2 may be computed, for example, using the (generalized) Jaccard measure introduced above. Subsequently, a matching procedure may be performed to find a bijection between a subset of the leaves in T_1 and a subset of the leaves in T_2 that maximizes the sum of the similarities between the pairs of leaves that are linked to one another by it (see Equation 7 of the Supplementary materials).

Common Sets of Predictor-Split Point Combinations
To compare sets of predictor-split point combinations while or while not taking directions of the splits into account (similarity type 12 or 13), a double-matching procedure is needed. The first matching resembles that described in subsection *"Common predictor-split point combinations."* However, here the goal is to find the largest possible bijection between a subset of the splits of X_k in leaf l of T_1 and a subset of compatible splits of that predictor in leaf l' of T_2. The similarity between leaf l of T_1 and leaf l' of T_2, then is the sum (across all predictors) of the cardinality of the domain of the bijection associated with each predictor. Once the similarities between all pairs of leaves of T_1 and T_2 have been found, a matching that resembles that described in the previous section could be performed on the level of the leaves, where the goal is to find a bijection between a subset of the leaves in T_1 and a subset of the leaves in T_2 that maximizes the sum of the similarities computed earlier between the pairs of leaves that are linked to one another by it (see Equation 9 of the Supplementary Materials).

Similarity Measures Related to the Classifications of the Trees

Existing Measures

Chipman (1998)—Partitions
Chipman et al. (1998), proposed a dissimilarity measure that is closely related to the RAND index, as it is based on whether pairs of observations that belong to the same leaf (resp. to different leaves) in T_1 behave similarly in T_2:

$$d(T_1, T_2) = \frac{\sum_{o>o'} |I^{T_1}(o, o') - I^{T_2}(o, o')|}{\binom{n}{2}}, \tag{5}$$

where o and o' denote a pair of observations ($o = 1, ..., n$ and $o' = 1, ..., n$), and where $I^{T_j}(o, o') = 1$ if o and o' belong to the same leaf in T_j, where $j = 1, 2$, and $I^{T_j}(o, o') = 0$ otherwise. This measure compares the partitions induced by the trees in terms of the leaves, without taking into account the specific class labels attached to these, and is thus an instance of similarity type A. It can easily be adapted to compare the partitions induced by the trees in terms of the class labels (similarity type B), by defining $I^{T_j}(o, o') = 1$ if the observations are assigned to the same class label in T_j, where $j = 1, 2$, and $I^{T_j}(o, o') = 0$ otherwise.

Fowlkes and Mallows (1983)–Partitions
Fowlkes and Mallows (1983) proposed a similarity measure that is based on a count, for each pair of a leaf of T_1 and a leaf of T_2, of the number of experimental units that end up in those two leaves:

$$s(T_1, T_2) = \frac{\sum_{l=1}^{L}\sum_{l'=1}^{L} w_{ll'}^2 - n}{\sqrt{(\sum_{l=1}^{L} w_{l0}^2 - n)(\sum_{l'=1}^{L'} w_{0l'}^2 - n)}}, \tag{6}$$

where $w_{ll'}$ is the number of objects that belong to both the lth leaf of T_1 and the l'th leaf of T_2, $w_{l0} = \sum_{l'=1}^{L'} w_{ll'}$ and $w_{0l'} = \sum_{l=1}^{L} w_{ll'}$. Fowlkes and Mallows (1983) show that this measure boils down to the number of pairs of objects that belong to the same leaf in both trees divided by a scaling factor. This measure is closely related to the RAND index, with the differences between the two being that (1) in the RAND index pairs of objects that belong to different leaves in both trees also contribute to a higher similarity and (2) the two measures make use of different scaling factors in the denominator. This measure is an instance of similarity type A as well.

Chipman (1998)—Classifications
A straightforward measure to compare the class labels assigned to the experimental units by two trees (similarity type D) is the following:

$$S(T_1, T_2) = \frac{1}{n}\sum_{o=1}^{n} I(\hat{c}_o^{T_1} = \hat{c}_o^{T_2}), \tag{7}$$

where $\hat{c}_o^{T_j}$ is the predicted class label for observation o by T_j.

Miglio and Soffritti (2004)
Miglio and Soffritti (2004) proposed an extension of the similarity measure of Fowlkes and Mallows (1983), that is also based on a count, for each pair of a leaf l of T_1 and a leaf l' of T_2, of the number of experimental units that end up in those two leaves, yet only for those leaves that are associated with the same class label in T_1 and T_2:

$$s(T_1, T_2) = \frac{\sum_{l=1}^{L}\sum_{l'=1}^{L} w_{ll'}^2 z_{ll'} - \sum_{l=1}^{L}\sum_{l'=1}^{L'} w_{ll'} z_{ll'}}{\sqrt{(\sum_{l=1}^{L} w_{l0}^2 - n)(\sum_{l'=1}^{L'} w_{0l'}^2 - n)}}, \tag{8}$$

where $z_{ll'} = 1$ if leaves l and l' are associated with the same class label, and 0 otherwise. As this measure is an extension of the measure of Fowlkes and Mallows (1983), it can be shown to boil down to the number of pairs of objects that belong to the same leaf in both trees, provided that the leaves in question are associated with the same class label. Hence, it is an instance of similarity type C.

Similarity Measures Related to the Content and Classifications of Trees

New Measures

Predictor (Split Point) Sets—Adjusted for Class Label of Leaf
If one would be interested in comparing sets of predictors or predictor-split point combinations associated with a leaf, while also taking into account the class label of that leaf, a restriction should be added to the measures proposed in the sections "*Common predictor sets*" and "*Common sets of predictor-split point combinations*." This restriction reads that

if the leaves associated with the two sets under comparison have different class labels, the similarity between these leaves is always 0. For possible specifications, we may refer to Equation 13 (resp. 14) of the Supplementary materials, which yield an instance of similarity type 4 (resp. 12), albeit restricted by taking into account specific classification labels (as in similarity type D).

2.2 Clustering

When a similarity matrix has been obtained, it will subsequently be used to partition the trees in the forest into k clusters. Within each cluster, the similarity between the trees must be as high as possible. Many clustering algorithms can be used for this purpose. However, in our case, the interpretation of the resulting clusters is of critical importance. Standard clustering algorithms for object by variable data (like k-means), yield such an interpretation in terms of cluster centroids, which comes down to the mean of the objects belonging to a cluster of interest on all variables under study. However, that is not applicable in our case. As a way out, we may revert to a method that does result in a partitioning of the objects (i.e., trees) under study into k clusters, and that represents each cluster by a central object or medoid (i.e., in our case a medoid tree), namely the partitioning around medoids (PAM) algorithm (Kaufman and Rousseeuw 2009). This algorithm partitions all objects (i.e., trees) into clusters, by assigning each object to the cluster with the closest medoid.

2.3 Post-Processing

2.3.1 Selecting the Number of Clusters

The number of clusters, k, that the PAM solution should contain, must be pre-specified by the user. This immediately relates to the question of how many clusters are needed to summarize the forest. Many different measures to decide upon the number of clusters have been proposed in the literature (for an early overview, see, e.g., Milligan and Cooper (1985)). They can be divided into measures that evaluate the homogeneity of the clusters, versus measures that evaluate other aspects of the clustering.

The first type of measures may be based either on object by variable data or on object by object similarities. In our context, only similarity-based measures can be used. Similarity-based measures can further rely on within-cluster similarity only (like the average within-cluster similarity), or on both within- and between-cluster similarity (like the average silhouette width proposed by Rousseeuw (1987)). Note that to compare the one-cluster solution to solutions with more than one cluster (which is of direct importance in our case), only measures that are solely based on within-cluster similarity can be used.

The second type includes all measures that evaluate some aspect of the clustering other than the homogeneity of the clusters. As an example, within the context of a random forest of classification trees, one may compare an assignment of class labels based on all trees in the forest with an assignment based on a weighted voting in terms of the class label assignments by the medoids (with the weights reflecting the corresponding cluster sizes). This newly proposed measure indicates how much predictive power is lost by reducing the forest to a certain number of medoids.

From the viewpoint of consistency, selecting the number of clusters with a similarity-based measure with use of the same type of similarities that was used as the input of PAM, seems a most straightforward choice, with average within-cluster similarity being a suitable choice if a one-cluster solution deserves to be considered. The most important exception

to this rule of thumb is that of prediction problems dealt with by means of random forests. To deal with such problems, one may wish to cluster the random forest on the basis of a classification-related (or a combined classification- and content-related) similarity measure, and subsequently decide on the number of clusters by means of a weighted medoid-based voting measure as outlined above.

2.3.2 Central Tendency and Heterogeneity

Recall that one may wish to recover three types of information from a forest of classification trees: (1) the central tendency of the forest, (2) the within-forest heterogeneity, and, in the case of a known source of variation, (3) whether, and if so how, within-forest differences are related to this source. We will now address how each of these three types of information can be derived from the clustering result.

Central Tendency Information regarding the central tendency of the forest can be obtained from the medoid of the one-cluster solution. In principle, this medoid contains two pieces of information, namely information on the medoid's predictive content and information on its class label assignments. Importantly, however, which of these pieces of information may be used for interpretation, is limited by the similarity measure at the basis of the clustering.

Indeed, if on the one hand the clustering was based on a similarity measure that captures the predictive content of the trees only, the trees in each cluster will be similar in terms of the predictors (and possibly the split points) they include, yet they may be very different in terms of the assigned class labels or implied partitions. As a consequence, medoids will be central in their clusters in terms of predictive content, whereas they may be not central at all in terms of assigned class labels or implied partitions. This further implies that primarily only an interpretation of the medoid tree in terms of its predictive content is justifiable in this case.

On the other hand, if the clustering relied on a similarity measure based on class label assignments or implied partitions of the experimental units, trees in each cluster will be similar in terms of those and may not be similar at all in terms of their predictive content. Hence, in this case, primarily an interpretation of the medoid tree in terms of the assigned class labels or partitions is justifiable. Otherwise, if one nevertheless feels the urge to look at the predictive content of the medoid, one should be cautious with interpretations based only on the specific included predictors (and split points). For example, if the data would include two nearly empirically equivalent predictors, including the one or the other in a tree will lead to virtually identical assignments. As such, if any of those predictors would show up in the medoid, instead of going for an interpretation in terms of this specific predictor only, one may wish to go for an interpretation in terms of the class of predictors that are nearly empirically equivalent with it. To enable the user to arrive at this type of interpretation, charting the relations of nearly empirical equivalence between predictors may be most helpful.

Finally, if a clustering was based on a measure that takes both the predictive content and the class labels into account, the medoid can be interpreted in terms of both aspects as well

Heterogeneity Regarding the heterogeneity of the forest, three pieces of information may be of interest, namely the number of clusters (along with their sizes), the predictive content of the cluster medoids, and the class label assignments or partitions implied by them.

Firstly, with respect to the number of clusters and their size, in general, the fewer clusters are needed to summarize the forest, the less heterogeneous the forest is. The presence of a single or a small number of large clusters also indicates a lower level of heterogeneity.

Secondly, the predictive content of the medoid trees may again primarily be used for interpretation when the clustering was based on a similarity measure capturing predictive content. Medoids then may be visually inspected in terms of their predictors, split points, or sets of predictors (resp. sets of predictor-split point combinations) associated with a leaf. Medoids that are only slight variations on a same topology (e.g., only one extra split in one medoid compared to another one), point at a low heterogeneity.

Thirdly, the assigned class labels or implied partitions of the medoids may again primarily be used when the clustering was based on a similarity measure that captures the class labels or partitions. At this point, the medoids can be compared in terms of the marginal totals of each class label, and in terms of a contingency table with the numbers of experimental units assigned to each combination of class labels by the different medoid trees. This type of comparison may reveal particular types of heterogeneity such as, for example, the fact that a first medoid tree is a refinement of a second one, with almost all observations that are assigned to class label 1 by the first medoid also being assigned to that class label by the second one (but not the other way around).

Relation Between Within-Forest Differences and a Known Source of Variation With regard to the relation between within-forest differences and some known source of variation, two types of relation with this source may be of interest, namely the relation with (1) the contents of the cluster medoids, and (2) cluster membership. As regards the first type of relation, the type of similarity measure at the basis of the clustering again determines the kind of interpretation that is primarily justifiable for this relation. Suppose that the forest is based on multiple outcome variables (which can be considered a known source of variation). If a similarity measure related to predictive content was used for the clustering, one may look at the similarities and differences in predictive content between the different medoids, and try to relate this to similarities and differences between the outcome variables that are primarily associated with the corresponding clusters. Alternatively, if the similarity measure was based on the class labels, one may again look at the marginal means of the class labels for each medoid, or at a contingency table and relate these to the outcome variables. The latter may result in statements such as: "If a person is predicted to have class label 1 for outcome variables 1 and 2, then he/she will usually not be predicted to have class label 1 for outcome variable 3."

To illustrate the second type of relation (i.e., the relation with cluster membership), we consider again the example of multiple outcome variables. A covariation between the nature of the outcome variable and cluster membership then could, for instance, imply that trees based on one outcome variable mainly end up in one cluster, whereas trees based on another outcome variable primarily end up in one or more other clusters.

3 Simulated Data Examples

Given the broad range of possible similarity measures and the typology and methodology we proposed to deal with it, one may wonder whether the results of using different similarity types would lead to markedly different conclusions. If not, there would be no real need for the typology in question. If yes, however, one may wish to have a better understanding of how and why different conclusions would come about. In the present section, we will address these issues by means of three examples, all three of which are based on analyses of simulated data. On each data set, we will grow a forest using bootstrap samples. Subsequently, we will calculate similarities making use of several different similarity measures, and use these to cluster the forests under study.

3.1 Example 1: Logical Equivalence

First, we consider cases that naturally give rise to multiple, equally complex, and classification of trees that represent logically equivalent classification rules. In such cases, similarity measures that capture predictive content in a refined way (e.g., by taking the order of the splits into account) may lead to a clustering with multiple clusters, each of which represents one of the logically equivalent rules, whereas similarity measures that capture predictive content in a less subtle way (as well as similarity measures that capture the classifications of the experimental units) may lead to a single cluster only. We illustrate with simulated data with a binary criterion variable that relates through a conjunction of two predictor-split point combinations to the predictor space, with the two different orders of the splits leading to two different trees that represent logically equivalent rules.

In particular, we generate a sample of $n = 500$ data points from $X_1, ..., X_5$ multivariate normal, with mean 0 and variance 1 for all variables, and with all correlations equal to .2, and $Y \sim \text{Bern}(\theta)$, with $\theta = .9$ if $(X_1 > 0 \ \& \ X_2 \leq 0)$, and $\theta = .1$ otherwise. The data generation structure for the criterion variable Y is graphically represented by the leftmost tree of Fig. 5.

We calculated similarities on the trees of the forest derived from this data set based on the following three measures: (a) the measure of Shannon and Banks (1999) which measures predictive content while taking split order into account, (b) the generalized Jaccard measure as a less subtle measure of predictive content, and (c) the proportion of common classifications proposed by Chipman et al. (1998) as a measure that taps classifications. We subjected the resulting similarities to PAM. Average within-cluster similarity plots to decide on the number of clusters are represented in Fig. 6. As predicted, the plot for the first similarity measure points at a two-cluster solution, whereas the plots for the other two measures clearly suggest a one-cluster solution to be in place. The medoids of the corresponding clusters are shown in Fig. 7. For the first measure, they represent trees that formalize the same conjunctive rule, yet with two different orders of the splits, whereas for the other two measures trees with the two different orders of the splits are grouped into a single cluster as summarized by the corresponding medoid.

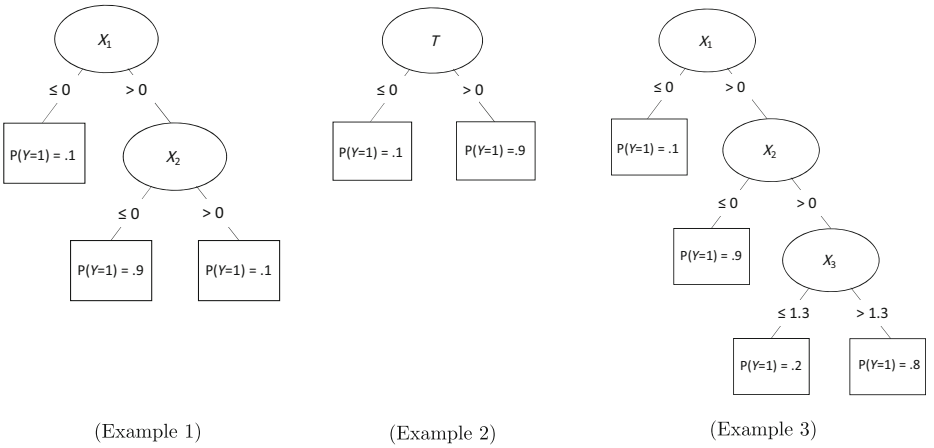


Fig. 5 Data generation structure for the criterion variable Y in the three simulated examples

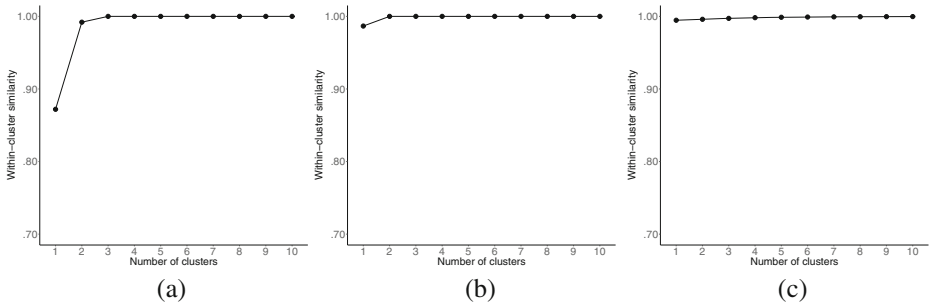


Fig. 6 Simulated example 1: Within-cluster similarity plots for the clusterings based on the **a** (Shannon and Banks 1999), **b** generalized Jaccard, and **c** (Chipman et al. 1998) similarity measures

3.2 Example 2: Empirical Equivalence

Second, we consider cases that give rise to multiple classification trees that represent empirically equivalent classification rules. In such cases, similarity measures that capture predictive content will lead to a clustering with multiple clusters, each of which represents one of the empirically equivalent rules, whereas similarity measures that capture the classifications of the experimental units may lead to a single cluster only. We illustrate with simulated data with a binary criterion variable that relates to the predictor space via a latent true variable T, two slightly perturbed variants of which are included in the predictor space; the two variants in question can be considered imperfect measures of T that are (almost) empirically equivalent.

In particular, we generate a sample of $n = 500$ data points from $X_1, ..., X_3, T$ multivariate normal, with mean 0 and variance 1 for all variables, and with all correlations equal to .2. Further, $X_4 = T + E$ and $X_5 = T + E'$. For the error variables E and E', it holds that $E, E' \sim \text{N}(0, 0.2)$, with statistical independence of each other and of all other predictor variables. Finally, for the criterion variable Y: $Y \sim \text{Bern}(\theta)$, with $\theta = .9$ if $T > 0$, and $\theta = .1$ otherwise. The data generation structure for the criterion variable Y is graphically represented by the middle tree of Fig. 5.

We calculated similarities on the trees of the forest derived from this data set based on (a) the generalized Jaccard as a measure of predictive content and (b) the proportion of common classifications proposed by Chipman et al. (1998) as a measure that taps classifications. We subjected the resulting similarities to PAM. Average within-cluster similarity plots to

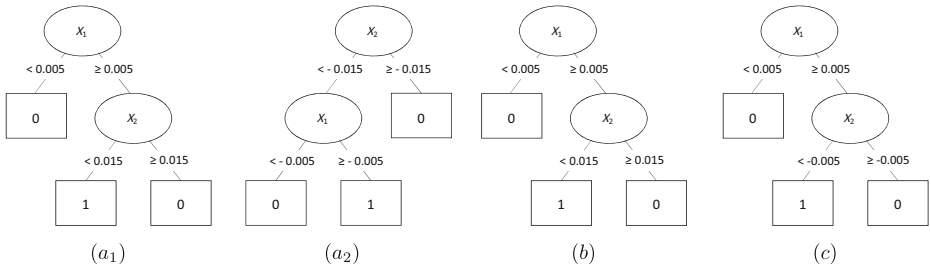


Fig. 7 Simulated example 1: Medoids of **a** the two-cluster solution based on the Shannon and Banks (1999) similarity measure, **b** the one-cluster solution based on the generalized Jaccard measure, and **c** the one-cluster solution based on the Chipman et al. (1998) measure

decide on the number of clusters are shown in Figure 2 of the Supplementary Materials. As predicted, the plot for the first similarity measure points at a two-cluster solution, whereas the plots for the second measure suggests a one-cluster solution. The medoids of the corresponding clusters are shown in Figure 3 of the Supplementary Materials. For the first measure, the two medoids represent the two empirically equivalent trees focused on, whereas for the second measure all empirically equivalent trees are grouped into a single cluster as summarized by the corresponding medoid.

3.3 Example 3: Small Subgroup

Third, we consider data stemming from a classification tree that includes a small leaf with a deviant classification. In such cases, one may expect in the resulting forest some trees that include the split that induces the leaf in question, in addition to trees that do not include it, since the split involved is rather hard to identify. Furthermore, one may expect that in such cases similarity measures that capture predictive content will lead to a clustering with two clusters, one with trees that do and one with trees that do not include the split in question. On the contrary, similarity measures that capture the classifications of the experimental units may lead to a single cluster only as the small additional group contributes very little predictive value.

To illustrate, we simulate a sample of $n = 500$ data points from $X_1, ..., X_5$ multivariate normal, with mean 0 and variance 1 for all variables, and with all correlations equal to .2. Furthermore, for the criterion variable Y: $Y \sim \text{Bern}(\theta)$, with θ depending on the values of the first three predictors according to the scheme represented by the rightmost tree in Fig. 5. Importantly, the rightmost leaf of this tree will have a very low cardinality.

We calculated similarities on the trees of the forest derived from this data set based on (a) the generalized Jaccard as a measure of predictive content and (b) the proportion of common classifications proposed by Chipman et al. (1998) as a measure that taps classifications. We subjected the resulting similarities to PAM. Average within-cluster similarity plots to decide on the number of clusters are shown in Figure 4 of the Supplementary Materials. As predicted, the plot for the first similarity measure points at a two-cluster solution, whereas the plot for the second measure suggests a one-cluster solution. The medoids of the corresponding clusters are shown in Figure 5 of the Supplementary Materials. For the first measure, the two medoids represent trees without and with the split on X_3, whereas for the second measure all trees are grouped into a single cluster as summarized by a medoid without the split in question.

3.4 Recommendations

The examples above clearly demonstrate how and why different similarity types in the proposed methodology may lead to markedly different conclusions. One may further wonder when and why certain approaches may be recommended over other ones.

First, if one cares about the order of the splits in the trees under study, the use of similarity measures that do take into account this order is obvious. However, one then should also beware of the possibility that an obtained order may be arbitrary. To check this, one may go for an additional cluster analysis based on a similarity measure that disregards order, and subsequently compare the results of the original and additional analyses. If the additional analysis leads to a markedly lower number of clusters, one may further wish to understand this result more thoroughly by subjecting the medoids of the original clustering to a logical analysis, in order to trace possible logical equivalences between them.

Second, if one cares about an interpretation of the trees under study in terms of the specific predictors included in them, the use of a similarity measure that captures predictive content would be appropriate. However, in such a situation one should also be on the alert for possible empirical equivalences that would undermine an interpretation in terms of specific predictors. To trace such equivalences, one may wish to go for an additional cluster analysis based on a similarity measure that captures the classifications implied by the trees. If such an analysis would once again suggest a markedly lower number of clusters, one may further try to gain a deeper understanding of this result by looking for empirical equivalences between predictors (or predictive combinations), for example, by inspecting the network of empirical correlations between predictors. If such an analysis would unveil empirical equivalences, indeed, one should take care to interpret the originally obtained trees in terms of empirical equivalence classes of predictors rather than in terms of individual predictors only.

Finally, if the researcher's key interest is in classifications, the choice of a similarity measure that taps such classifications is obvious. Nevertheless, if one would like to consider going beyond a global prediction and inspect classification subtleties with possible relevance for the identification of the substantive mechanisms underlying the classifications in question, an auxiliary cluster analysis based on a similarity measure that captures predictive content may be a useful addition.

4 Real Data Example

In this section, we will illustrate the proposed methodology with a real data example, viz., the Drug Consumption data set (Fehrman et al. 2017), which is freely available from the UCI Machine Learning Repository (Dheeru and Karra Taniskidou 2017): https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29. This set consists of anonymous survey data from 1885 respondents, who reported on their use of 18 types of drugs. We will focuse on the use of cocaine only in terms of the same binary response variable as Fehrman et al. (2017), with categories non-user (never used or used over a decade ago) versus user (used in the last decade, year, month, week, or day). We will further make use of all predictor variables in the data set: age, gender, level of education, the Big Five personality traits (i.e., extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience) measured by the NEO-FFI-R (McCrae and Costa 2004), impulsivity measured by the BIS-11 (Patton et al. 1995), and sensation seeking measured by the ImpSS (Zuckerman et al. 1993). Note that all predictors were initially categorical (ordinal or nominal) and were quantified by Fehrman et al. (2017).

Our initial research question is which type of persons are at risk for cocaine use. This question can be addressed by growing a single classification tree on the full data set. The result (obtained via the R-package `rpart`, with pruning using the complexity parameter associated with a cross-validated prediction error that is maximally one standard error higher than the lowest one) is presented in panel (a) of Fig. 9. This tree, which includes a single node only, implies that more sensation seeking persons (with a score > -0.07) are at risk for using cocaine. This makes sense from an intuitive point of view.

However, because of the well-known instability of trees, we will investigate how stable the obtained result is by means of the proposed methodology. Specifically, we draw 100 bootstrap samples and grow a classification tree on each sample. As we are interested in the variability across the trees in the resulting forest in terms of both their predictive content and their implied class labels, we use the similarity measure based on predictor-split point combinations that takes into account the class label of the leaves (Equation 14 in

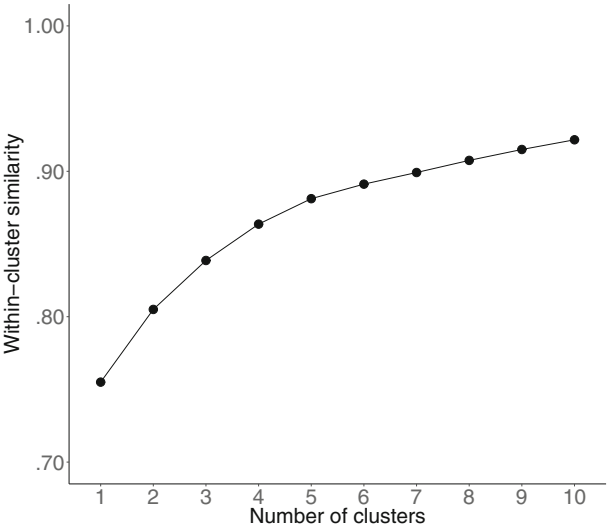


Fig. 8 Real data example on cocaine use: within-cluster similarity plot for the clusterings based on the mixed content- and classification-related measure

Supplementary Materials). Regarding the split points, we use a tolerance of 0.2 for all predictors in the data set.

The thus obtained similarities are subjected to PAM, with the number of clusters ranging from 1 through 10. To decide on the number of clusters, we inspect the within-cluster similarity plot in Fig. 8. This figure does not display a clear elbow, yet is already rather high from the beginning, and thus points at solutions with a small number of clusters.

We therefore decided to inspect more closely the solutions with one to four clusters. This inspection is facilitated by the fact that the corresponding sets of medoids are fully nested, in that, for k ranging from 1 to 3, the medoid set of the k-cluster solution is a subset of the medoid set of the $(k+1)$-cluster solution. Importantly, one should note that this result is not a mathematical necessity. The full set of the medoids under study is shown in Fig. 9. For each clustering solution, the percentages of trees that belong to each of the clusters in it are given in Table 1.

From an inspection of the medoids in Fig. 9, it appears that all of them make sense from a substantive viewpoint. More specifically, we learn the following: The first medoid coincides with the classification tree grown on the full data set as already discussed above. All three other medoids start with a primary split on sensation seeking. In the second medoid, there is a secondary split on age, with the medoid as a whole indicating that the group of people at risk for cocaine use scores higher on sensation seeking AND is older than 24 years.

Table 1 Real data set: percentage of trees that belong to each cluster

# clusters in solution	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	100			
2	79	21		
3	79	14	7	
4	72	14	7	7

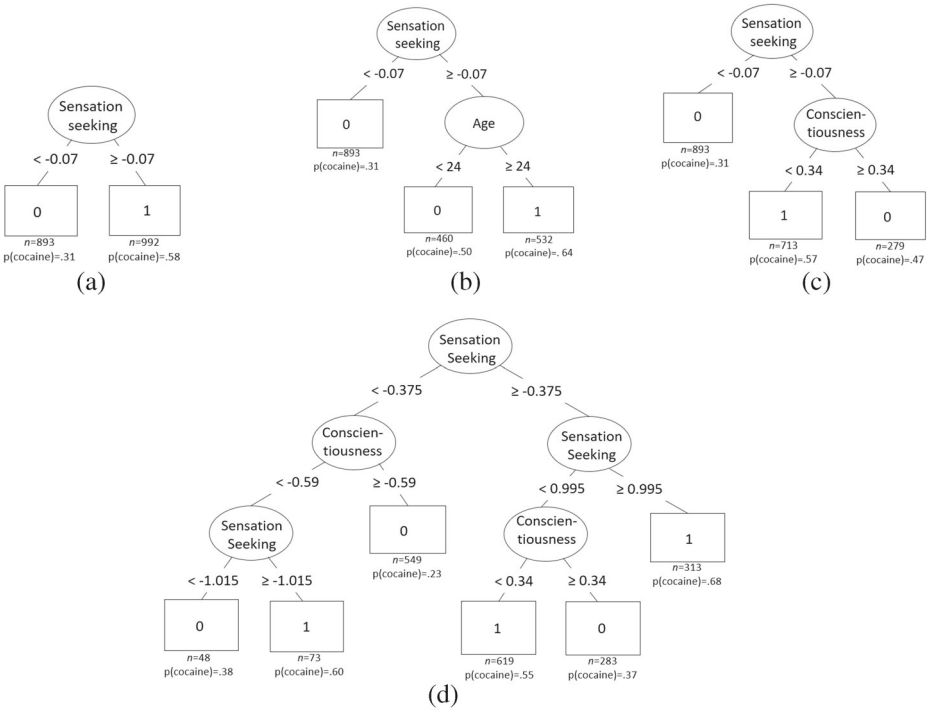


Fig. 9 Real data example on cocaine use: Classification tree grown on full data set (**a**); medoid of 1-cluster solution **a**; medoids of 2-cluster solution (**a**), **b**; medoids of three-cluster solution (**a**), (**b**), (**c**); medoids of 4-cluster solution (**a**), (**b**), (**c**), (**d**). Below each leaf the corresponding cardinality and proportion of cocaine users in the full data set is given

This means that, according to this medoid, the people with a higher score on sensation seeking and younger than 24 (middle leaf of medoid (b)) have been split off from the group at risk. This can be considered a subsidiary modification of medoid (a), not because of a small cardinality of the leaf in question (on the contrary, for this leaf $n = 460$) but because the proportion of cocaine users in the leaf is .50. Medoids (c) and (d) imply secondary splits on conscientiousness. In particular, medoid (c) implies that within the group with higher sensation seeking scores the subgroup with relatively higher conscientiousness scores (rightmost leaf) is split off from the group at risk for cocaine use. Again, this can be considered a subsidiary modification of medoid (a), not because of the cardinality of the leaf in question ($n = 279$) but because of the relatively high proportion of cocaine users in it, $p(Y = 1) = .47$. Medoid (d) appears to be the most involved one, as it implies a subdivision of the range of sensation seeking in four parts: $]-\infty, -1.015[$, $[-1.015, -0.375[$, $[-0.375, 0.995[$, and $[0.995, \infty[$. In the two most extreme parts, cocaine use depends on sensation seeking only (with in $]-\infty, -1.015[$ a predicted 0 and in $[0.995, \infty[$ a predicted 1. In the two middle parts, however, conscientiousness shows up as a moderating variable, with the group in $[-1.015, -0.375[$ being at risk for cocaine use if conscientiousness is lower than -0.59, and with that in $[-0.375, 0.995[$ being at risk if conscientiousness falls below 0.34. All this clearly points at a meaningful compensatory relationship between on the one hand cocaine use and on the other hand sensation seeking and conscientiousness. The medoid as a whole can further again be considered a subsidiary refinement of medoid (a), because of

the relatively lower cardinality of the subgroup at risk within the range $[-1.015, -0.375[$, $n = 73$, and because of the proportion of cocaine users in $[-0.375, 0.995[$ being close to .50, $p(Y = 1) = .49$.

Based on all findings, it seems that the result obtained from the full data is fairly stable: First, the within-cluster similarity is already rather high for the 1-cluster solution, the medoid of which coincides with the classification tree for the full data. Second, in the solutions with 2 to 4 clusters, the first cluster (with the classification tree of the full data set as medoid) comprises the overwhelming majority of the trees in the forest (see Table 1). Beyond this overall stability, we may note that in all medoids under study, sensation seeking constitutes the primary splitting variable (with in three out of the four medoids even the same split point). In addition to the clear and stable role of sensation seeking, there are indications that age and/or conscientiousness might play a secondary role in the prediction of cocaine use. This role, however, seems to be much weaker, and the nature of the interplay of the possible additional predictors with sensation seeking is much less clear, especially for conscientiousness. As such, the findings on these possible additional predictors can be considered to be tentative at best, and may strongly benefit from more targeted follow-up research.

The above application to real data in its entirety clearly demonstrates the added value of our newly proposed methodology. This methodology may lead to fairly rich answers to questions such as stability issues, indeed, which obviously go beyond a yes/no, thanks to the insights it may yield in the contents and composition of a forest under study.

5 Concluding Remarks

In this paper, we addressed the problem that if, instead of a single classification tree, a forest of trees is used to solve a statistical learning problem, the most attractive feature of single trees, namely their insightfulness, gets lost. We proposed a first principles conceptual framework for getting insight into such a forest, along with a novel associated methodology, and proper guidance in the use of both. The proposed methodology to see a forest for the trees can be used to get insight into many different types of forests, and can address a broad range of research questions.

Funding Information The research reported in this paper was supported in part by the Research Foundation—Flanders (G080219N) and by the Research Fund of KU Leuven (C14/19/054). The data used in the real data examples were obtained from the UCI Machine Learning Repository (Dheeru and Karra Taniskidou 2017; Fehrman et al. 2017).

References

Banerjee, M., Ding, Y., Noone, A.-M. (2012). Identifying representative trees from ensembles. *Statistics in Medicine*, *31*(15), 1601–1616. https://doi.org/10.1002/sim.4492.

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Machine Learning*, *36*(1), 105–139. https://doi.org/10.1023/A:100751542.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655.

Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, *24*(6), 2350–2383. https://doi.org/10.1214/aos/1032181158.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). *Classification and regression trees*. Belmont: Wadsworth. https://doi.org/10.1201/9781315139470.

Briand, B., Ducharme, G.R., Parache, V., Mercat-Rommens, C. (2009). A similarity measure to assess the stability of classification trees. *Computational Statistics & Data Analysis*, *53*(4), 1208–1217. https://doi.org/10.1016/j.csda.2008.10.033.

Chipman, H., George, E., McCulloch, R. (1998). Making sense of a forest of trees. In Weisberg, S. (Ed.) *Proceedings of the 30th symposium on the interface* (pp. 84–92). Fairfax: Interface Foundation of North America.

Dheeru, D., & Karra Taniskidou, E. (2017). UCI machine learning repository. Retrieved from http://archive.ics.uci.edu/ml.

Dietterich, T.G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139–157. https://doi.org/10.1023/A:1007607513941.

Fehrman, E., Muhammad, A.K., Mirkes, E.M., Egan, V., Gorban, A.N. (2017). The five factor model of personality and evaluation of drug consumption risk. In Palumbo, F., Montanari, A., Vichi, M. (Eds.) *Data science: innovative developments in data analysis and clustering* (pp. 231–242): Springer. https://doi.org/10.1037/10140-001.

Fowlkes, E.B., & Mallows, C.L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*(383), 553–569. https://doi.org/10.2307/2288117.

Freund, Y., & Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504.

Hastie, T.J., Tibshirani, R.J., Friedman, J.H. (2009). *The elements of statistical learning: data mining, inference and prediction*. New York: Springer. https://doi.org/10.1007/978-0-387-84858-7.

Jaccard, P. (1901). ÉTude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Socié,té Vaudoise des Sciences Naturelles*, *37*, 547–579.

Kaufman, L., & Rousseeuw, P.J. (2009). *Finding groups in data: an introduction to cluster analysis*. New York: Wiley. https://doi.org/10.1002/9780470316801.

Little, R.J., D'Agostino, R., Cohen, M.L., Dickersin, K., Emerson, S.S., Farrar, J.T., et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, *367*(14), 1355–1360. https://doi.org/10.1056/NEJMsr1203730.

McCrae, R.R., & Costa, P.T. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences*, *36*(3), 587–596. https://doi.org/10.1016/s0191-8869(03)00118-1.

Miglio, R., & Soffritti, G. (2004). The comparison between classification trees through proximity measures. *Computational Statistics & Data Analysis*, *45*(3), 577–593. https://doi.org/10.1016/s0167-9473(03)00063-x.

Milligan, G.W., & Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179. https://doi.org/10.1007/bf02294245.

Patton, J.H., Stanford, M.S., Baratt, E.S. (1995). Factor structure of the barratt impulsiveness scale. *Journal of Cinical Psychology*, *51*(6), 768–774. https://doi.org/10.1002/1097-4679(199511)51:6⟨768::aid-jclp2270510607⟩3.0.co;2-1.

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/bf00116251.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley. https://doi.org/10.1002/9780470316696.

Schafer, J.L., & Graham, J.W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147. https://doi.org/10.1037//1082-989x.7.2.147.

Shannon, W.D., & Banks, D. (1999). Combining classification trees using MLE. *Statistics in Medicine*, *18*(6), 727–740. https://doi.org/10.1002/(sici)1097-0258(19990330)18:6⟨727::aid-sim61⟩3.3.co;2-u.

Skurichina, M., & Duin, R.P. (2002). Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, *5*(2), 121–135. https://doi.org/10.1007/s100440200011.

Strobl, C., Malley, J., Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323. https://doi.org/10.1037/a0016973.

Turney, P. (1995). Technical note: bias and the quantification of stability. *Machine Learning*, *20*(1), 23–33. https://doi.org/10.1007/bf00993473.

Zuckerman, M., Kuhlman, D.M., Joireman, J., Teta, P., Kraft, M. (1993). A comparison of three structural models for personality: the big three, the big five, and the alternative five. *Journal of Paediatric Personality and Social Psychology*, *65*(4), 757. https://doi.org/10.1037//0022-3514.65.4.757.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.