A similarity measure to assess the stability of classification trees

Bénédicte Briand a,*, Gilles R. Ducharme b, Vanessa Parache a, Catherine Mercat-Rommens a

a IRSN, Laboratoire d’Études Radioécologiques en milieux Continentaux et Marin, DEI/SÉSURE/LERCM, Cadarache, bât. 153, BP 3, 13115 Saint-Paul-lez-Durance, France
b Equipe de probabilités et statistique, I3M, Université Montpellier II, cc 051, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

A R T I C L E   I N F O

Article history:
Received 29 October 2007
Received in revised form 21 October 2008
Accepted 22 October 2008
Available online 6 November 2008

A B S T R A C T

It has been recognized that Classification trees (CART) are unstable; a small perturbation in the input variables or a fresh sample can lead to a very different classification tree. Some approaches exist that try to correct this instability. However, their benefits can, at present, be appreciated only qualitatively. A similarity measure between two classification trees is introduced that can measure their closeness. Its usefulness is illustrated with synthetic data on the impact of radioactivity deposit through the environment. In this context, a modified node level stabilizing technique, referred to as the NLS–REP method, is introduced and shown to be more stable than the classical CART method.

1. Introduction

Classification trees (CART) are useful in many applied areas. They have been recognized in various applications ([Bel et al., in press; Briand et al., 2007; Mishra et al., 2003; Mokhtari et al., 2006]) as important tools, due to their very good explanatory capability and their build-in ability to hierarchize the predictive variables in a natural way. This makes them easy to use and interpret by non-specialists. However, CART trees are known for their instability ([Breiman, 1996; Ghattas, 2000]); a small perturbation in the data, or a fresh sample, can lead to a very different tree and thus a different misclassification rate. This instability has an impact on their interpretability, since the hierarchies of predictive variables can greatly change. This may be a real problem for decision-makers that need to take actions in critical situations.

The radioecological sensitivity problem that motivated this work is one such case. It pertains to the radiological impact of an accidental radioactive discharge on populations and the environment. This impact depends on the significance and nature of the discharge, but also on the territory that receives it. Radioecological sensitivity analysis ([Mercat-Rommens and Renaud, 2005; Mercat-Rommens et al., 2007]) tries to identify the characteristics strongly influencing the fate of a radioactive contaminant released into the environment. It requires the identification of the explanatory variables, arranged in a specific order, that push the output radioactive contamination level beyond regulatory thresholds. For this task, classification trees are ideal tools for decision-makers that need to devise recommendations in a post-accidental context as the situation evolves. If very different trees are available, these persons may have to take decisions based on contradictory information — an unpleasant situation.

To reduce instability, Dannegger (2000) has proposed a bootstrap-based node-level stabilizing approach, hereafter referred to as the NLS approach. However the stability properties of this, and similar methods, have so far been assessed qualitatively, for example, by visual inspection of the resulting trees or by indirect methods. This makes it difficult to rigorously compare various methods intended to stabilize tree-growing procedures. Moreover, the pruning procedure associated with the NLS method can lead to severe overpruning, a problem that seems to have gone undetected.

* Corresponding author. Tel.: +33 4 42 19 96 11; fax: +33 4 42 19 91 42.
E-mail address: briand.benedicte@caramail.com (B. Briand).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.10.033
The goals of this paper are twofold. First, we introduce a similarity measure between two classification trees that can quantitatively assess their closeness. This measure can also be used to define a measure of dispersion in a set of trees and identify a “central” or medoid tree. Second, using this measure, we empirically establish that the NLS method, coupled with a different pruning strategy, which we refer to as the NLS–REP approach, is indeed much more stable than the classical CART, while avoiding the severe overpruning. This is done through an example from radioecological sensitivity analysis.

The paper is organized as follows. Section 2 presents a brief review of CART. Dannegger’s NLS method is described in Section 3, along with the adapted pruning procedure that leads to the NLS–REP approach. In Section 4, we introduce our similarity measure and discuss some of its uses and variants. The stability of the NLS–REP approach is illustrated in Section 5. Note in closing, that our similarity measure can also be useful with approaches that produce many different trees (a forest) that are then somehow summarized as, for example, in the popular random forest methodology of Breiman (2001). These approaches often improve the predictive capability, but at the expense of interpretability, since the structure of a single tree is lost. This interpretability can be restored if the forest has a small dispersion about the medoid tree that may then serve as a representative of the forest. One instance of this is outlined in Section 4 and 5, as a byproduct of our analysis of the NLS–REP method. In other cases, the similarity matrix of the set of trees can be computed and analyzed for clusters using standard methods. But this is beyond the scope of the present paper.

2. The CART method

CART (Breiman et al., 1984) is probably the most widely used method in the field of recursive partitioning procedures. Here, CART is viewed as a classification method. To build a classification tree, a learning sample of \( n \) observations, each containing the values of the input variables \( X_k (k = 1, \ldots, K) \) and of the output variable \( Y \in \{1, 2, \ldots, J\} \), is observed. The classification tree results from the application of the following steps:

**Step (a) Building the maximal tree.** Start with a root node \( t_0 \) that contains the whole learning sample. The tree is obtained by successive segmentation of this sample into subsets called nodes. Each node is obtained by a split on one of the input variables, say \( X_k \), that sends to a (left) child node \( t_l \) all the observations verifying \( X_k \leq \delta \) (a threshold) and to a (right) child node \( t_r \), all observations such that \( X_k > \delta \). The process is repeated on the nodes \( t_l \) and \( t_r \), which are, in turn, divided into pairs of child nodes. At each node, the split is chosen so as to give the most homogeneous child nodes with respect to the values of \( Y \). An impurity function \( I \) is defined to measure the heterogeneity in each node. The impurity used here is the entropy at node \( t \):

\[
I(t) = -\sum_{j=1}^{J} \hat{p}(j|t) \log \hat{p}(j|t),
\]

where \( \hat{p}(j|t) \) is the proportion of observations with \( Y = j \) that have been assigned to node \( t \). Note that \( I(t) = 0 \) when the node is pure, i.e. containing observations with a single value of \( Y \). The decrease in impurity (\( D \)) due to split \( \Delta \) at node \( t \) is

\[
D(\Delta, t) = I(t) - \hat{p}_l I(t_l) - \hat{p}_r I(t_r),
\]

where \( \hat{p}_l \) and \( \hat{p}_r \) are the proportions of individuals in node \( t \) sent to child nodes \( t_l \) and \( t_r \), respectively. Consequently, the best split at node \( t \) is

\[
\Delta^* = \arg \max_{\Delta \in D_t} \{D(\Delta, t); \Delta \in D_t\},
\]

where \( D_t \) is the set of all possible splits at node \( t \).

The splitting process at node \( t \) is stopped when \( I(t) = 0 \) or when the number of observations \( N_t \) in the node is less than a predefined threshold. The node is then declared a terminal node or leaf and is assigned to the value of \( Y \) which is most abundant. Iterating in this way, the learning sample is successively split so as to build a highly detailed tree \( A_{\text{max}} \) called the maximal tree.

**Step (b) Pruning the maximal tree.** \( A_{\text{max}} \) will most certainly overfit the observations, since it uses all quirks in the learning sample to optimize the purity at each node. To correct this, a pruning process is applied. Pruning is done by successively removing a node and all its descendants (a branch) and replacing this node by a leaf. This creates a sequence \( \{A_{\text{max}}, \ldots, A_{\text{min}} = t_0\} \) of trees.

**Step (c) Selection of the optimal tree.** Within this sequence of trees, the “optimal” subtree is selected based on its misclassification rate, as estimated by cross-validation or by a second sample on the same variables, called the validation sample. Various selection rules can be used. See Breiman et al. (1984).

3. The NLS–REP approach

As stated, decision trees are known for their instability (Breiman, 1996; Ghattas, 2000). To correct this problem, Dannegger (2000) has introduced the NLS method, which uses the bootstrap at the node-level to stabilize the selection of the best split. For convenience, we recall his algorithm.
for every node \( t \) containing \( N_t \) observations do
  for \( b = 1, \ldots, B \) do
    Generate a bootstrap sample \( N_t(b) \) from the observations in node \( t \).
    With this bootstrap sample and for each explanatory variable \( X_k \), find
    the best split according to (1).
  end for
Compute the number of times each variable is used to split the node: \( \{ \text{num}_t(X_k) \}_{k=1,\ldots,K} \)
Select the variable to be used to split node \( t \) as \( X_d = \arg \max \{ \text{num}_t(X_1), \ldots, \text{num}_t(X_K) \} \)
Determine the threshold for the chosen variable as the median of the thresholds for this variable.
end for

Dannegger (2000) implementation of the NSL method uses the pruning procedure of Breiman et al. (1984) which
is based on cross-validation. However, we have noted in conducting the simulations of Section 5, that this can lead to
severe overpruning, in some cases down to the root node. CART builds \( A_{\text{max}} \) using optimal splits according to (1), so that
the misclassification rate systematically decreases as the tree gets bigger. The associated pruning procedure of Breiman et al. (1984)
uses this. In the NLS method however, the splits are not optimal and, as the number of leaves increases, the
misclassification rate does not behave similarly. This can force the complexity parameter in Breiman et al. (1984),
which controls the pruning, to be negative, thus leading to undesired behavior.

To prune NLS trees, the “reduced error pruning” (REP) method of Quinlan (1987) seems more adapted. A thorough study
and an empirical comparison of this method with other pruning techniques can be found in Esposito et al. (1997), Mingers (1989) and Quinlan (1987). REP bypassed the above problem by using a validation sample to measure the performance of the various branches of the maximal tree \( A_{\text{max}} \) and works as follows. Starting from \( A_{\text{max}} \), let \( R(A') \) be the estimate of the misclassification rate in branch \( A' \) (the branch having \( t \) for root node) as evaluated by the validation sample. Let also \( R(t) \) be the estimate of the misclassification rate in node \( t \) if this node is replaced with a leaf. Define

\[
g(t) = R(A') - R(t).
\]

The branch that maximizes (2) is pruned unless it contains a sub-branch starting at some node \( t' \) such that \( R(A'') < R(A') \),
because that would result in pruning a branch with a root node which is not very informative, but with a sub-branch bringing
useful information. The process is iterated until the tree contains only nodes for which the value of criterion (2) is negative.
Note that a problem can appear with this NLS–REP method when the validation sample is much smaller than the learning
sample. Some rare configurations in the learning sample that have led to the development of some branches, may be absent
from the validation sample, and this can cause overpruning (Esposito et al., 1997; Quinlan, 1987). This problem largely
disappears when both the validation and learning samples have comparable sizes.

4. Similarity between trees

Let \( A_1, A_2 \) be classification trees pertaining to the same variables, but built by different methods. Suppose first, that all
variables have finite range, and that both trees have the same structure. By structure, we mean the set of nodes and their
respective location without the information on the split at each node. Note that the following also applies to ordinal, discrete
or continuous variables. Let \( t_0, t_1, \ldots, t_T \) be the various non terminal nodes of the trees, numbered in descending order from
the root node, and from left to right. At node \( t, A_1 \) splits according to, say, \( X_k \leq \delta_1 \) while \( A_2 \) splits when \( X_k' \leq \delta_2 \).
Set

\[
S_t = I[k = k'] \left( 1 - \frac{|\delta_1 - \delta_2|}{\text{range}(X_k)} \right),
\]

as the dissimilarity between \( A_1 \) and \( A_2 \) at node \( t \), where \( I[k = k'] \) is the indicator of event \( X_k = X_k' \). We have \( S_t = 0 \) when
the split at node \( t \) is made on different variables \( (X_k \neq X_k') \) or, if the splitting variables coincide, when the splitting points
are located at opposite extremities of the range of \( X_k \cdot S_t = 1 \) if the splits at node \( t \) are identical. Otherwise, \( S_t \in (0, 1) \).
Now let \( q_0, q_1, \ldots, q_T \) be user-supplied non negative weights summing to 1. The measure of similarity between \( A_1 \) and \( A_2 \)
is defined as

\[
d(A_1, A_2) = 1 - \sum_{t=0}^{T} q_t S_t.
\]

This quantity is symmetric in its arguments and satisfies \( 0 \leq d(A_1, A_2) \leq 1 \) with \( d(A_1, A_2) = 0 \) if both trees are identical.
Similarly, when none of the split variables coincide, \( d(A_1, A_2) = 1 \).

Several remarks are in order. First, if the range of \( X_k \) is infinite, it can be replaced by the observed range of that variable
in the (combined) samples. Another variant would replace the ratio in (3) by some other measure of relative dispersion, for example \( |F_k(\delta_1) - F_k(\delta_2)| \), where \( F_k \) is the (perhaps empirical) cumulative distribution function of \( X_k \). Secondly, the indicator in (3) can be replaced by a “soft” function. In many cases, optimal splits are often so, by a small margin in (1). A surrogate
split is a replacement for the optimal split based on a variable that leaves the rest of the branch nearly identical. Surrogate
splits play an important role when some observations contain missing data. When a surrogate split exists at some node, it can make sense to replace the indicator in (3) by a function that takes into account the near exchangeability. For example, if \( \Delta^* \) is the optimal split at node \( t \) based on variable \( X_k \) and \( \Delta^{**} \) is a surrogate split based on \( X_k \), then one could consider \( 1 - (\mathcal{D}(\Delta^*, t) - \mathcal{D}(\Delta^{**}, t))/\mathcal{D}(\Delta^*, t)) \). The ratio in (3) would require adjustments to take into account the possibly different ranges of \( X_k, X_k' \), for example by standardizing both variables.

In (4), the weights \( q_i \) are introduced to allow emphasis to be put on parts of the tree that appear more relevant to their comparison. In many practical applications, it would seem that the uppermost part of the tree should be given maximal weight, while the nodes close to the leaves may receive less importance. Otherwise, uniform weights \( (q_i = 1/(T + 1)) \) could be used. It will be seen empirically in Section 5 that an irresolute user can feel comfortable: this choice does not seem crucial.

Similarity measure (4) can be adapted to compare trees with different structures. Each time a node is encountered in one tree at the position where a node is subsequently split in the other, then a “ghost branch” of the same structure is added to replace the leaf in the first tree. In the ghost branch, the splits and thresholds can be left undefined with all terms \( S_i \) trivially set to 0. Please note that the above similarity measure does not take into account the leaves of the tree. This is because in many cases, as in our radioecological application, the interest lies in the hierarchy of splits, as it will be translated into a decision making process. However, the above method could be extended to the leaves by adding to \( d(A_1, A_2) \) a term measuring the discrepancies of their assignation. For example, if a leaf of \( A_1 \) is assigned to \( Y = j \), while the same leaf of \( A_2 \) assigns to \( j' \), then this term can be taken proportional to \( I(j = j') \).

The dissimilarity (3) can be adapted to nominal variables. Suppose variable \( X_k \) is nominal (with cardinal \( \#V \)). Suppose further that, at node \( t, A_1 \) assigns to the left child node the categories in the set \( V_1 \) (with cardinal \( \#V_1 \)) and to the right node the categories in the set \( V_1' \) (with cardinal \( \#V_1' \)). Then we can take

\[
S_i = I(k = k') \left(1 - \frac{\#(V_1 \cap V_2) + \#(V_1' \cap V_2')}{\#V}\right).
\]

Note that if \( A_1 \) splits on the nominal variable \( X_k \) while \( A_2 \) splits on some other variable, nominal or not, then \( S_i = 0 \) in view of the effect of the term \( I(k = k') \).

We indicate some other uses of this similarity measure. Let \( \{A_1, \ldots, A_M\} \) be a set (forest) of trees pertaining to the same input and output variables. This forest could result from various samples, the application of different tree building methods on the same sample, or from the application of a systematic procedure, as the random forest method of Breiman (2001). The “dispersion” of this forest can be measured via the formula

\[
\frac{1}{M(M - 1)} \sum_{1 \leq i < j \leq M} d(A_i, A_j). \tag{5}
\]

An heuristic motivation for this expression comes from its relationship with the alternate expression for the variance of a set of observations \( \{X_1, \ldots, X_n\} \), given, for example, in Serfling (1980):

\[
n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n(n - 1)} \sum_{1 \leq i < j \leq n} (X_i - X_j)^2.
\]

The similarity measure can also be used to define a “central” or medioid tree in a forest. Recalling that the sample median can be defined as the element of \( \{X_1, \ldots, X_n\} \) that minimizes in \( a, \sum_{i=1}^{n} |X_i - a| \) leads to define this central tree as the element of \( \{A_1, \ldots, A_M\} \) that minimizes in \( A \) the quantity \( \sum_{i=1}^{M} d(A_i, A) \).

### 5. Stability of the NLS–REP method

To explore the stability of the NLS–REP method of Section 3 and show the usefulness of the similarity measure of Section 4, simulations were conducted using a radioecological model that has been recognized as useful in the context of the application that motivated this work. All synthetic samples were generated from this model, that involves 5 explanatory variables and one output variable that is subsequently categorized. Care has been taken so that the probability distributions of the explanatory variables be consonant with available subject-matter knowledge. This requires, at one point, a careful analysis of the relationship between the two variables. The details and results of the simulations are described below.

#### 5.1. Data simulation

The radioecological model considered here pertains to the radionuclide transfer for the contamination of lettuces by the accidental release of strontium-90 (\(^{90}\text{Sr}\)) in the atmosphere. It is adapted from the ASTRAL (Assistance Technique en Radioprotection Post-Accidentelle) model (Mourlon and Calmon, 2002) and takes the form

\[
C_{\text{veg}} = \frac{D \times R_c \times e^{-(\lambda_b + 0.000068) \Delta}}{Yld}, \tag{6}
\]
where the output variable \( C_{veg} \) (Bq kg\(^{-1}\), Becquerel per kilogram) is the plant activity (at harvest time) due to foliar transfer. This gives the radioactive contamination of the lettuces, and is coded into a binary variable \( Y \) according to radioprotection limits (as defined by the Codex Alimentarius Commission (1989)) in the following way:

\[
Y = 1 : C_{veg} \leq 100 \text{ Bq kg}^{-1} \quad \text{and} \quad Y = 2 : C_{veg} > 100 \text{ Bq kg}^{-1}.
\]

The following input variables are environmental or anthropogenic characteristics susceptible to worsening the impact of a pollution incident:

- \( D \) (Bq m\(^{-2}\), Becquerel per square meter) the radioactivity deposit,
- \( R_a \), the retention ratio on the day of the accident. This corresponds to the fraction of deposit (in Bq m\(^{-2}\)) intercepted by the foliar mass of the plant at the soil surface. It varies between 0 and 1, and is a dimensionless activity ratio,
- \( \lambda_b \) (day\(^{-1}\)), the biomechanical decay constant of the radionuclide for the plant, which corresponds to the dilution of activity during plant growth,
- \( \Delta \) (day), the time between deposit and harvesting,
- \( \text{Yld} \) (kg m\(^{-2}\)), the crop yield at harvest time.

Three of these variables may be generated independently from the others. Their probability distributions have been characterized from bibliographical research as follows.

- **Deposit** \((D)\): Uniform distribution with the upper and lower bounds determined as follows:
  - Based on the zoning criteria defined by the 1991 act (Law of the Republic of Belarus, 1991) (following the Chernobyl accident), a contamination level of 111 kBq m\(^{-2}\) in \(^{90}\)Sr is the pivotal value between an obligatory and immediate resettlement zone and a zone with less priority. We have chosen 100 kBq m\(^{-2}\) as the upper bound.
  - The lower bound is taken so that, below this value, the plant activity \( C_{veg} \) never achieves the threshold value of 10 Bq kg\(^{-1}\) of \(^{90}\)Sr, considered as negligible. This bound is determined by resolving Eq. (6), using the least favorable values of the various input variables and gives a lower bound of 20 Bq m\(^{-2}\).

- **Biomechanical decay constant** \((\lambda_b)\): Based on Renaud et al. (1999) and Rommens et al. (2001), we use an asymmetric triangular density with min = 0.03; mode = 0.046; max = 0.14.

- **Time between deposit and harvesting** \((\Delta)\): This quantity can also be expressed as the difference between the growing time of the plant \((T_c)\) and the date of the accident \((t_a)\). According to peer reviews reported in Rommens et al. (2001), \(T_c\) can be assumed to follow a triangular distribution with min = 30, mode = 45, max = 90. The accident can occur at any time during the development of the plant, thus we have set \(t_a\) as a uniform distribution over \([1, T_c]\).

The last two input variables \((R_a\) and \(\text{Yld}\)) are related and have been generated in the following way. The STICS (Simulateur multiDiciplinaire pour des Cultures Standard) model described in Brisson et al. (2003) gives a daily follow-up on the production of fresh biomass and the rate of soil coverage by the lettuce, variables to which the retention ratio is linked via the leaf area index. To obtain ranges of values and possible relationships for these variables, two stations in France with opposite climates were selected (Orange, for the Mediterranean climate and Rennes for the oceanic climate). Based on these stations, various simulations were performed to study the influence of climate, soil and crop management plan (Briand et al., in press). The results obtained are as follows.

- **Yield** \((\text{Yld})\): STICS showed a linear relationship between crop yield at harvest time and growing time of the lettuce (Fig. 1). The prediction interval for the regression line gives a range of plausible yield values for each growing time studied. Thus, knowing the plant’s growing time, values of \(\text{Yld}\) are generated according to a uniform distribution over this range.

- **Retention ratio** \((R_a)\): When an accidental radioactivity deposit occurs, a plant can be contaminated at any stage of its development. It is thus important to know what fraction of the activity deposited, will be intercepted by its aerial parts. The analysis carried out in Briand et al. (in press) yields a range of probable values for variable \(R_a\) for each growing time studied. Knowing the plant’s growing time and the date of the accident, a value of \(R_a\) can be generated according to an uniform distribution within the range of possible associated values (Fig. 2).

5.2. Variables importance

Even with as little as 5 explanatory variables, CART and NLS–REP trees can be quite large. To speed up the simulations and allow a better understanding of the behavior of the methods of the paper without losing sight of our application, we have looked into the possibility of simplifying the tree building process by removing from consideration some unnecessary input variables. To this end, we have measured the importance of each of the variables in model (6) by the variable importance measure defined by Breiman et al. (1984). This measure is the sum over all nodes in the tree, of the decrease in impurity caused by the surrogate split on this variable (surrogate split can be seen as replacement division for the optimal split).

We generated a pair of samples (a learning and a validation sample, each of size 5000), built a classification tree by the CART methodology and computed the importance of each variable according to the method described in Breiman et al. (1984) and Ghattas (1999). This procedure was repeated 100 times independently. Fig. 3 presents the results in boxplot form for each of the five explanatory variables. For each of the 100 trials, an identical hierarchy of variables was obtained.
Fig. 1. Linear relationship between yield and growing time of the lettuce.

Fig. 2. Variability of the retention ratio for a lettuce with a 40 day growing time.

Fig. 3. Original variable importance defined by Breiman et al. (1984) of the five input variables.
Fig. 4. Results of the variable importance measures. The plots in the top row display the MDG (Mean Decrease Gini) and the MDA (Mean Decrease Accuracy) when the random forest method is used. The plots in the bottom row display the MDA when the cforest method is used for sampling with replacement and for subsampling without replacement.

\( R_c; \Delta; D; \lambda_b; Yld \). Since the last two variables have little importance, we exclude them from further consideration. From a subject-matter point of view, this removal can be justified by the fact that they are difficult to measure in the environment, and therefore of little use in the post-accident decision making context of our application.

Some bias exists in this importance measure. To obtain further confirmation that the removal of these variables is warranted, we have computed two of Breiman (2002) importance measures (Mean Decrease Gini and Mean Decrease Accuracy) associated with random forests and one (Mean Decrease Accuracy) associated with the cforest methodology of Hothorn et al. (2006). A description and comparison of these measures can be found in Archer and Kimes (2008) and Strobl et al. (2007). We generated 100 learning samples of size 5000, and applied these methods. Results are presented in boxplot form in Fig. 4. They confirm the ranking that was previously obtained.

5.3. Empirical study of the NLS–REP method

For the reasons stated in the opening sentences of the previous section, we have, in addition, forced \( A_{\text{max}} \) to have a fixed depth of 6, the last level pertaining to the leaves. The subject-matter justification for this is that, in our application, simple trees are required to make them easily usable in an emergency situation. First, we study the misclassification rates of the NLS–REP and CART methods. In implementing the former method, the number of bootstrap replicates was fixed at \( B = 100 \). As before, the size of the learning and validation samples was set at \( n = 5000 \). This procedure was repeated 10 times. An independent test sample, also of size 5000, was generated to evaluate the performances of each method. The misclassification rates (based on the test sample) are presented in boxplot form in Fig. 5. For each method, the value between brackets represents the average misclassification rate (%) over the 10 runs. For the NLS–REP method, this average is 5.55% while CART yields a slightly larger value of 5.85%. More importantly, inspection of this figure indirectly shows the stability of the NLS–REP trees: the misclassification rate is much less variable than with the CART trees. Fig. 6 adds to this evidence by showing again less variation in the number of leaves produced by each method.
It is known (Breiman, 1996, 2001; Freund and Schapire, 1999) that more complex tree building procedures can yield smaller misclassification rates. To see if this behavior holds here, two methods based on model aggregation, bagging and random forests, have been applied to the same data sets with the following parameter values: number of bootstrap replicates = 100 and random inputs selected at each node for random forests = 2. As expected, bagging and random forests are slightly better than the NLS–REP and CART, with an average misclassification rate in the range of 4.4%. They exhibit a dispersion somewhat similar to the NLS–REP method.

We now directly study the structural stability of the trees grown by the NLS–REP and CART methods. For this, we computed the similarity measure of Section 4 on the 10 pairs of classification trees \( (A_{\text{NLS-REP},i}, A_{\text{CART},i}) \), \( i = 1, \ldots, 10 \), first with the uniform weights \( q_i = 1/(T + 1) \) and then with the depth adjusted weights \( q_i \) proportional to \( 1/2^{\ell_i-1} \) so as to sum to one, and where \( \ell_i \) represents the depth: 1 for the root node, 2 for the second level and so on. Results are presented in Fig. 7. The \((10 \times 9)/2 = 45\) points on this figure have coordinates \( (d(A_{\text{NLS-REP},i}, A_{\text{NLS-REP},j}), d(A_{\text{CART},i}, A_{\text{CART},j})) \). The much narrower spread of the NLS–REP trees shows that this method is more stable than CART. Quantitatively, this is captured by the magnitude of dispersion measure \( S \). It is 0.12 (resp 0.04) for the NLS–REP trees and 0.34 (resp 0.23) for the CART method using uniform (resp. depth adjusted) weights. Note that the same conclusions derive from either set of weights, showing the inconsequential impact of this choice here.

Since the NLS–REP trees exhibit a small dispersion, it makes sense to use our similarity measure to define a mediod tree as described in Section 4. This mediod tree for the forest \( \{A_{\text{NLS-REP},i}, i = 1, \ldots, 10\} \) is presented on Fig. 8. The number of observations in each node, its misclassification rate and the percentage of misclassified observations are indicated in each of its leaves. For intermediate nodes, the figure gives details on the voting procedure, the number between brackets representing the vote associated with the variable. The misclassification rate of this mediod tree (5.54%) is comparable to the average from the 10 NLS–REP trees (5.55%).
Fig. 7. Comparison of the similarity measures for CART and NLS–REP methods. Uniform weights: \( q_i = 1/(T + 1) \), depth adjusted weights: \( q_i \) proportional to \( 1/2^{(L_i - 1)} \).

Fig. 8. Medioid classification tree obtained by the NLS–REP method.

In this medioid tree, as in all trees of the above forest, the best division for the root node is based on \( R_c \). By taking the left branch of the tree (from the root node), a first classification rule emerges:

**IF** \( R_c < 0.04 \) **THEN** \( Y (= ^{90}\text{Sr activity}) = 1 \).

This leaf contains 1245 observations (25% of the sample) and only 0.6% are misclassified. The subject-matter interpretation of this rule leads us to conclude that a low \( R_c \) is associated with lettuces with a weakly developed foliar system (i.e. early growth stage). This rule is instructive for decision-makers; regardless of the radioactive deposit, lettuces contaminated during their early growth stage almost always exhibit a radioactive contamination less than or equal to 100 Bq kg\(^{-1}\) at harvest time. Likewise, by interpreting the different rules in the central tree, the user can identify the variables responsible for the different levels of radioactive contamination in lettuce. This translates into a decision-making process that can be applied, with the added comfort that it possesses some robustness with respect to the information on which it is based.

6. Conclusion

In this paper, we have introduced a similarity measure between classification trees that can assess, quantitatively, how close they are from one another. This measure can be computed with any type of explanatory variables entered in the classification process. It was shown that it could also be used to define a measure of dispersion among a set, or forest, of trees, and identify a medioid tree that may serve as a representative of this set.

In the radioactive contamination application that motivated this work, as with many other types of problems where classification trees are useful, the decision-maker needs to have robust decision rules. To this end, we have introduced a new method, called NLS–REP, to build robust classification trees. By a node level stabilizing technique, this method allows the construction of a classification tree with more stable splitting rules. Empirical comparisons between this and the CART
method have shown that, regarding prediction accuracy, the NLS–REP method has a slight advantage over CART. However, by using our similarity measure, it was shown that the splits in NLS–REP are much more stable than those of CART. As a result, the idea of extracting a medoid tree as a representative of a forest of NLS–REP trees becomes an interesting addition to the toolbox of classification techniques. Note, however, that the NLS–REP approach is very computer intensive. This is perhaps its main drawback.

Acknowledgments

The SENSIB project has been supported financially by the ADEME (contract No. 0472C0035). The authors would like to thank the referees, who have made many useful suggestions, leading to a much-improved version of the manuscript.

References


