Computational Statistics (2008) 23:407–427
DOI 10.1007/s00180-007-0077-5

ORIGINAL PAPER

# Classification trees for ordinal variables

**Raffaella Piccarreta**

Accepted: 16 June 2007 / Published online: 25 July 2007


**Abstract** We introduce new criteria to obtain classification trees for ordinal response variables. At this aim, Breiman et al. (Classification and regression trees. Wadsworth, Belmont, 1984), extended their *twoing* criterion to the ordinal case. Following CART procedure, we extend the well known Gini–Simpson criterion to the ordinal case. Referring to the exclusivity preference property (introduced by Taylor and Silverman in Stat Comput 3:147–161, 1993, for the nominal case), suitably modified for the ordinal case, a second criterion is introduced. The hereby proposed methods are compared with the ordered twoing criterion via simulations.

**Keywords** CART · Gini–Simpson criterion · Gini index of heterogeneity · Ordered categorical variables · Twoing criterion

## 1 Introduction

Suppose we are interested in predicting a categorical response variable, $Y$ via tree-structured classification. In order to predict $Y$, we therefore generate a tree on the basis of measurement on $Q$ predictors, $X_1, X_2, \ldots, X_Q$, available for a *training set* of $N$ cases.

There are different ways to grow a tree. One of the most famous is the CART procedure, introduced by Breiman et al. (1984) (from now onwards B & A). CART consists of two phases: *growing* and *pruning*. In the growing phase, the training set is recursively partitioned into subsets, called *nodes*, through a sequence of binary splits. At each step, a node is split into two sub-nodes according to the values of one of the explanatory variables. A sequence of nested trees is thus obtained, having an

R. Piccarreta (✉)
Istituto di Metodi Quantitativi, Università "L. Bocconi", viale Isonzo, 25, 20136 Milan, Italy
e-mail: raffaella.piccarreta@uni-bocconi.it

increasing number of *terminal nodes*, i.e., nodes which are no longer split. The growing phase ends when the *maximal tree*, $T_{Max}$, is obtained, having the maximum number of terminal nodes.

In the pruning phase, terminal nodes are merged to prevent over-fitting of the classification tree to the training set it is based upon. The tree selected after the pruning is the one characterized by the lowest risk (properly measured). In Sect. 2, we describe the Gini–Simpson and the *twoing* criterion, at the basis of CART methodology (B & A).

The aim of the paper is twofold. Our *first object* is to introduce new criteria to grow trees in the case when the response is an ordered categorical variable (from now onwards, ordinal variable). The rationale for this is to enrich the class of ordinal splitting criteria: we think that the availability of many criteria and the inspection of different trees can also prove very useful from an exploratory point of view, allowing insights into the data to be analyzed.

In Sect. 3, we describe the *ordered twoing criterion*, an extension of the twoing method proposed by B & A to deal with the ordinal case. Following this reasoning, we provide an extension of the Gini–Simpson criterion. Ordinal criteria are compared and their properties are discussed. We first introduce the distinction between "global" and "not global" criteria, and show that Gini–Simpson criterion is a global one, while the twoing criterion is not. We then compare the two criteria by referring to the "exclusivity preference property", studied by Taylor and Silverman (1993) for the nominal case, and extended in this paper to the ordinal case. We provide some insights about this property, show that the ordinally extended Gini–Simpson criterion does not possess it (as its nominal counterpart), and introduce a new (global) criterion possessing it.

Still following Taylor and Silverman (1993) we discuss the problem of the presence of the so-called anti-end-cut factor in all the considered ordinal criteria. In Sect. 3 the criteria are compared by referring to a real data set, while in Sect. 4 their performance is evaluated on the basis of simulations.

As is evident, the *second object* of the paper is to discuss in details properties and characteristics either of the newly introduced criteria and of the "standard one" (the ordered twoing criterion). We think this is an important point, especially because ordinal criteria have so far received less attention both from the substantial and from the methodological side.

As a final consideration, in Sect. 5 we evidence a further property of the ordinally extended Gini–Simpson criterion, permitting to fasten the growing phase. This property, which proves particularly relevant in the case when big data set are analyzed, is an extension of a similar property already illustrated for the nominal Gini–Simpson criterion by Mola and Siciliano (1997, 1998).

Section 6 concludes and draws directions for future research.

## 2 The Gini–Simpson and the twoing criteria

Let $\boldsymbol{X}$ be a vector of $Q$ explanatory variables, $X_1, \ldots, X_Q$. On the basis of $\boldsymbol{X}$, $Y$ is to be predicted, $Y$ being a nominal categorical variable assuming $G$ levels, $y_1, \ldots, y_G$. Suppose for a training set consisting of $N$ cases, measurements are available on $(\boldsymbol{X}, Y)$.

A classification tree permits to predict the (unknown) level of $Y$ for a new individual, characterized by a vector $\boldsymbol{x}_{(*)}$.

The first phase in building a tree is the *growing phase*. At the first step, all cases constitute a single *node*. At subsequent steps, each node splits into two subgroups (nodes), on the basis of the values of one of the explanatory variables, called the *splitting variable*. In CART, the best split of a node $t$ is selected by referring to the concept of *impurity*, a measure of the heterogeneity of cases in $t$, defined as:

$$I_N(t) = \sum_{g=1}^{G} \pi_t(g) \cdot [1 - \pi_t(g)] = 1 - \sum_{g=1}^{G} \pi_t^2(g), \tag{1}$$

$\pi_t(g)$ being the proportion of cases in $t$ characterized by the $g$-th level of $Y$. Suppose now $t$ splits into the two subgroups $t_L$ and $t_R$ according to a split $s$, and let $p_L$ and $p_R$ be the proportions of individuals placed in $t_L$ and $t_R$ ($p_L + p_R = 1$). Denote by $\pi_t(g|L)$ and $\pi_t(g|R)$ the proportion of cases characterized by the $g$-th level of $Y$ in $t_L$ and $t_R$ respectively. The Gini–Simpson criterion to evaluate the split $s$ is the *decrease in impurity* achieved when passing from $t$ to $t_L$ and $t_R$:

$$\begin{aligned} C_{GN}(s,t) &= \Delta I_N(s,t) = I_N(t) - p_L I_N(t_L) - p_R I_N(t_R) \\ &= p_L p_R \sum_{g=1}^{G} [\pi_t(g|L) - \pi_t(g|R)]^2. \end{aligned} \tag{2}$$

The best split of node $t$, $s^*(t)$, is selected so as to maximize the decrease in impurity, $s^*(t) = \arg\max_s C_{GN}(s,t)$.

B & A introduce the twoing criterion as an alternative to evaluate a split. Let $\mathcal{C}(t) = \{y_1, \ldots, y_G\}$ be the set of categories of $Y$ in a node $t$. Consider two subsets of $\mathcal{C}(t)$ , say $\mathcal{C}_1(t) = \{y_{g_1}, y_{g_2}, \ldots, y_{g_h}\}$ and $\overline{\mathcal{C}}_1(t) = \mathcal{C} \backslash \mathcal{C}_1$. A binary response variable can thus be defined: $Y^* = 1$ if $Y \in \mathcal{C}_1$, 0 otherwise. The impurity within node $t$, depending of course on the choice of $\mathcal{C}_1(t)$, is measured by applying (1) to $Y^*$. We have that

$$I_N(t, \mathcal{C}_1) = 2\pi_t(\mathcal{C}_1)\pi_t(\overline{\mathcal{C}}_1),$$

where $\pi_t(\mathcal{C}_1) = \sum_{g:y_g \in \mathcal{C}_1} \pi_t(g)$. Given $\mathcal{C}_1$ the decrease in impurity provided by split $s$ is, by (2):

$$\begin{aligned} C_{TN}(s,t|\mathcal{C}_1) &= I_N(t, \mathcal{C}_1) - p_L I_N(t_L, \mathcal{C}_1) - p_R I_N(t_R, \mathcal{C}_1) \\ &= 2 p_L p_R \left[\pi_t(\mathcal{C}_1|L) - \pi_t(\mathcal{C}_1|R)\right]^2. \end{aligned}$$

Now let $\mathcal{C}^*_{1|s} = \arg\max_{\mathcal{C}_1} C_{TN}(s,t|\mathcal{C}_1)$. Split $s$ is evaluated through

$$C_{TN}(s,t) = 2 p_L p_R \left[\pi_t(\mathcal{C}^*_{1|s}|L) - \pi_t(\mathcal{C}^*_{1|s}|R)\right]^2. \tag{3}$$

The best split is then $s^*(t) = \arg\max_s C_{TN}(s,t)$.

$C_{GN}$ and $C_{TN}$ both consist of two components. The term $p_L p_R$, called *anti-end-cut* factor, forces the criteria to encourage splits producing subsets of similar sizes. The second term synthesizes the difference between the two conditional distributions of $Y$ in the sub-nodes induced by the split.

Whatever the criterion selected to grow the tree, the growing phase ends when the maximal tree, $T_{Max}$, is obtained. The growing phase is followed by the *pruning phase*, whose aim is to prevent over-fitting of the classification tree to the training set it is based upon. At each step of the pruning phase two terminal nodes are merged. A sequence of nested sub-trees having a decreasing number of nodes is thus determined, $T_{Max} \succ T_1 \succ T_2 \cdots \succ \{t_1\}$, $\{t_1\}$ indicating the root node of $T_{Max}$. The "best" tree is then chosen, having the lowest risk [measured by the misclassification rate ($mr$) or by the misclassification cost ($mc$)], evaluated by referring to a validation set or by using a $v$-fold cross-validation (see B & A, for details). Throughout the paper, the optimum size tree, selected after the pruning, is characterized by the minimum tenfold cross-validation risk.

In the next section, we consider the problem of growing a maximal tree in the case when the response variable is ordinal.

## 3 The ordinal problem: old and new solutions

If the response, $Y$, is ordinal, it can be of interest to grow a classification tree taking into account the ordering of the levels of $Y$.

Before proceeding, it is worthwhile to evidence which kind of split has to be considered as a "good" one in an "ordinal sense". For a nominal response, a good split should lead to "pure" sub-nodes, possibly characterized by a strong mode (hence the modal category should be characterized by a high frequency). When considering an ordinal response, the "purity" of a node is not necessarily the primary criterion to evaluate the goodness of a split. Consider for example a node $t$ with the following composition: $\boldsymbol{n}_t = \{25, 25, 25, 25\}$. Let $s_1$ and $s_2$ be two splits of $t$, both with $p_L = p_R = 0.5$. Suppose that split $s_1$ leads to sub-nodes with compositions $\boldsymbol{n}_{t_L}(s_1) = \{25, 0, 25, 0\}$ and $\boldsymbol{n}_{t_R}(s_1) = \{0, 25, 0, 25\}$, while for split $s_2$ it is $\boldsymbol{n}_{t_L}(s_2) = \{25, 20, 5, 0\}$ and $\boldsymbol{n}_{t_R}(s_2) = \{0, 5, 20, 25\}$. Split $s_1$ is purer than split $s_2$: it sends *all* cases of classes 1 and 3 into $t_L$ and all cases of classes 2 and 4 into $t_R$, assuring thus a substantial simplification of node $t$. Nevertheless, in an ordinal sense split $s_1$ does not appear preferable to $s_2$, since it leads to sub-nodes with cases characterized by non adjacent levels of $Y$. Hence, an ordinal criterion could select the less pure split $s_2$ as the best.

Since nominal and ordinal criteria may select different splits, the inspection of trees obtained by referring to ordinal criteria may be of substantial interest even from an exploratory point of view. In fact, it allows insight into data and might possibly evidence relations among the explanatory structure and the response variable which are not revealed by "nominal trees".

In this section, we illustrate some old and new criteria for *growing the maximal tree* taking into account the ordering of the levels of $Y$.

The twoing criterion (3) can easily be extended to this case. While in the nominal case any subset of the categories of $Y$ can be considered, in the ordinal case attention is limited to partitions such as $\mathcal{C}_g = \{y_1, y_2, \ldots, y_g\}$ and $\overline{\mathcal{C}}_g = \{y_{g+1}, y_{g+2}, \ldots, y_G\}$. It is clear that $\pi_t(\mathcal{C}_g) = \sum_{j=1}^{g} \pi_t(j) = F_t(g)$, the cumulative distribution function (cdf) of $Y$ evaluated in $y_g$. Recalling (3), simple calculations yield that for given $\mathcal{C}_g$ and $s$ it is:

$$\begin{aligned} C_{TO}(s,t|\mathcal{C}_g) &= p_L p_R \left[\pi_t(\mathcal{C}_g|L) - \pi_t(\mathcal{C}_g|R)\right]^2 \\ &= p_L p_R \left[F_t(g|L) - F_t(g|R)\right]^2 . \end{aligned} \tag{4}$$

Heterogeneity between the two sub-nodes is now measured by referring to the dissimilarity between the two conditional cdf's. It can be easily shown that, *for a given split* $s$, the class maximizing (4) is

$$\mathcal{C}_{g^*|s} \quad \text{with } g^* = \arg\max_g [F_t(g|L) - F_t(g|R)]^2.$$

Hence the ordered twoing criterion to evaluate a split is

$$C_{TO}(s,t) = p_L p_R \max_g [F_t(g|L) - F_t(g|R)]^2. \tag{5}$$

The dissimilarity between the two cdf's is thus measured by their *maximal* distance. In this sense, the ordered twoing criterion is not based upon a "global" measure of dissimilarity between the two cdf's. We will discuss this issue in more detail later.

The twoing criterion is extended to the ordinal case by substituting the conditional distributions in (3) with the conditional cdf's in (4). Using a similar substitution in (2), the Gini–Simpson criterion can also be extended to the ordinal case:

$$C_{GO}(s,t) = p_L p_R \sum_{g=1}^{G-1} [F_t(g|L) - F_t(g|R)]^2. \tag{6}$$

It is simple to show that $C_{GO}$ represents the decrease in impurity when passing from $t$ to the sub-nodes induced by split $s$. The impurity measure considered in this case is obtained by substituting $F_t(g)$ for $\pi_t(g)$ in (1):

$$I_O(t) = \sum_{g=1}^{G} F_t(g) \cdot [1 - F_t(g)]. \tag{7}$$

$I_O(t)$ is a measure of the heterogeneity of an ordinal variable originally proposed by Gini (1954).

The main difference between $C_{GO}$ and $C_{TO}$ is the measure of dissimilarity between the two conditional—cumulative—distributions. In $C_{GO}$ the "global" distance between the two conditional cdf's is taken into account, while in $C_{TO}$ only their maximal distance is considered. To appreciate the difference between the two criteria, consider

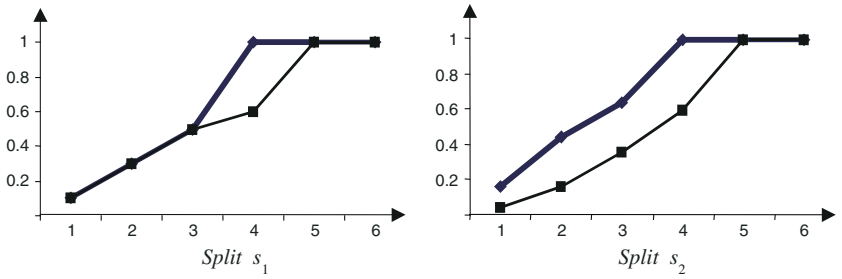


**Fig. 1** Conditional cdf's for different splits

the following example. Let $t$ be a node characterized by $\boldsymbol{n}_t = \{10, 20, 20, 30, 20\}$. Let $s_1$ and $s_2$ be two splits of $t$, both with $p_L = p_R = 0.5$. The sub-nodes generated by $s_1$ have compositions $\boldsymbol{n}_{t_L}(s_1) = \{5, 10, 10, 25, 0\}$ and $\boldsymbol{n}_{t_R}(s_1) = \{5, 10, 10, 5, 20\}$, while for split $s_2$ it is $\boldsymbol{n}_{t_L}(s_2) = \{8, 14, 10, 18, 0\}$ and $\boldsymbol{n}_{t_R}(s_2) = \{2, 6, 10, 12, 20\}$. The conditional cdf's corresponding to the splits are reported in Fig. 1.

In this case it is $C_{TO}(s_1) = C_{TO}(s_2)$, since the maximal distance between the cdf's is equal to 0.4 for both splits. Hence, criterion $C_{TO}$ does not take into account the fact that the two cdf's defined by $s_1$ coincide for $Y < 4$. On the contrary, criterion $C_{GO}$ prefers split $s_2$ to split $s_1$ in that, as is evident from Fig. 1, in this case the distance between the two cdf's is greater.

In the next section, $C_{GO}$ and $C_{TO}$ are compared by referring to a property first considered by Taylor and Silverman (1993) (from now onwards T & S).

### 3.1 The exclusivity preference property

As for criteria to build classification trees when the response is nominal, T & S introduced the concept of "exclusivity preference property", and showed that Gini–Simpson criterion *does not* have this property.

In this section, we introduce an extension of the exclusivity preference property to the ordinal case, we show that $C_{GO}$ does not possess this property, and we try to explain the meaning of "not possessing" this property.

A splitting criterion has the (nominal) exclusivity preference property if:

i) given $p_L$ and $p_R$, it gives the largest value to *any* splits that are *exclusive*, i.e., such that: $\sum_{g=1}^{G} \pi_t(g|L) \cdot \pi_t(g|R) = 0$;
ii) regardless of $p_L$ and $p_R$, it takes the smallest possible value if the two conditional distributions induced by a split are identical, i.e., when $\pi_t(g|L) = \pi_t(g|R)$ for all $g = 1, \ldots, G$.

To extend the exclusivity preference property to the ordinal case, we introduce the concept of *ordinal exclusive* splits. A split will be said *ordinal exclusive* if a level of the response variable exists, $g^*$, such that

$$\sum_{g=1}^{g^*} \pi_t(g|L) \cdot \pi_t(g|R) = 0 \quad \text{and} \quad \sum_{g=g^*+1}^{G} \pi_t(g|L) \cdot \pi_t(g|R) = 0.$$

**Table 1** Monotone exclusivity preference property

| | |
|---|---|
| Split $s_1$: $\boldsymbol{n}_{t_L}(s_1) = \{10, 30, 39, 1, 0\}$ | $\boldsymbol{n}_{t_R}(s_1) = \{0, 0, 1, 19, 20\}$ |
| Split $s_2$: $\boldsymbol{n}_{t_L}(s_2) = \{0, 0, 40, 20, 20\}$ | $\boldsymbol{n}_{t_R}(s_2) = \{10, 30, 0, 0, 0\}$ |
| Split $s_3$: $\boldsymbol{n}_{t_L}(s_3) = \{10, 30, 40, 0, 0\}$ | $\boldsymbol{n}_{t_R}(s_3) = \{0, 0, 0, 20, 20\}$ |

Hence a criterion has the "monotone exclusivity preference property" if, given $p_L$ and $p_R$, it favors ordinal exclusive splits, assigning the largest value to them, and if it satisfies condition ii). This happens if and only if one of the conditional distributions is entirely above or entirely below the other (see for example splits $s_2$ and $s_3$ in Table 1 below).

It can be easily shown that $C_{GO}$ always satisfies property ii), but (for given $p_L$ and $p_R$) it does not always favor ordinal exclusive splits (as shown by T & S for the Gini–Simpson criterion). To see this, consider splits $s_1$ and $s_2$ in Table 1.

Even if $s_2$ is ordinal exclusive, $C_{GO}$ selects $s_1$ as the best split (since $C_{GO}(s_1, t) = 0.3205 > C_{GO}(s_2, t) = 0.3056$). It is instead simple to show that the ordered twoing criterion *does* have the monotone exclusivity preference property.

Before proceeding, we want to deeper inspect the meaning of this property. In their work, T & S provided an example similar to the one in Table 1 for the nominal case, and argued "Split $A$ is an exclusive split and its Gini–Simpson score is ..., however the Gini–Simpson criterion for the non exclusive split $B$ is the higher value ... It is hard to imagine that split $B$ would be preferred from any intuitive point of view" (T & S, p. 155). Actually, *there is* a reason for which this happens (either in the nominal and in the ordinal case). A criterion *having* the exclusivity preference property considers, for given $p_L$ and $p_R$, all the exclusive splits as equivalent. On the contrary, $C_{GO}$ *discriminates* among ordinal exclusive splits. To verify this, consider split $s_3$ in Table 1. It is $C_{GO}(s_3, t) = 0.3368 > C_{GO}(s_1, t) > C_{GO}(s_2, t)$. So, $s_1$ is preferred to the ordinal exclusive split $s_2$ because it is "more similar" to the ordinal exclusive split $s_3$, which is preferred to the alternative ordinal exclusive split $s_2$ (a similar argument also holds for the Gini–Simpson criterion).

In conclusion, the point we wish to make is that not having the monotone exclusivity preference property should not be necessarily considered a drawback (as in T & S) for a given criterion. Nevertheless, we can suppose that in applications we will rarely need to compare exclusive splits, and we can then be interested in using a criterion characterized by the exclusivity preference property.

This is the reason why we now introduce another criterion *having* the monotone exclusivity preference property and being based on a "global" measure of the dissimilarity between the two conditional distributions induced by a split (we are in fact interested in evaluating the impact of both factors on the performances of the classifier: the kind of dissimilarity measure and the exclusivity preference property). To do this, we refer to an index introduced by Agresti (1981) to evaluate nominal–ordinal association. Consider a node $t$ and a split $s$. $s$ defines a nominal variable (which we will still indicate by $s$) assuming the two categories $L$ and $R$. A possible measure of the association between $s$ and $Y$ is (Agresti 1981):

$$\Delta_A(s,t) = \sum_{g=1}^{G}\sum_{j>g}\pi_t(g|L)\pi_t(j|R) - \sum_{g=1}^{G}\sum_{j>g}\pi_t(g|R)\pi_t(j|L)$$
$$= \sum_{g=1}^{G}\pi_t(g|L)[1-F_t(g|R)] - \sum_{g=1}^{G}\pi_t(g|R)[1-F_t(g|L)].$$

Our idea is to use $\Delta_A(\cdot,t)$ as a criterion for selecting the split. Notice that $-1 \leq \Delta_A(s,t) \leq 1$, with $|\Delta_A(s,t)| = 1$ if and only if $s$ induces an ordinal exclusive split. It is then clear that the maximum value is reached by *all* the ordinal exclusive splits, irrespective of $p_L$ and $p_R$.

As T & S point out, "under certain circumstances, we might be prepared to sacrifice exclusivity of the offspring if it can only be attained at the cost of creating one very small offspring node". This is the reason why exclusivity should be preferred only conditionally on $(p_L \cdot p_R)$, reaching its maximum value when $p_L = p_R = 0.5$. So, the criterion we will refer to is:

$$C_A(s,t) = p_L p_R |\Delta_A(s,t)|, \tag{8}$$

containing the *anti-end-cut factor*, as the other criteria described above.

### 3.2 Adaptive anti-end-cut factors

As pointed out above, all the considered ordinal criteria contain the so called *anti-end-cut* factor (from now onwards $aecf$), forcing a criterion to encourage splits producing sub-nodes of similar sizes. T & S illustrate that in some situations this characteristic may lead to unsatisfactory trees.

To evidence the effects of the $aecf$ consider splits $s_1$ and $s_2$ in Table 2. The (pure) split $s_1$, putting together adjacent classes of $Y$, should intuitively be preferred to $s_2$. Nevertheless, all of the ordinal criteria prefer $s_2$ to $s_1$, due to the $aecf$. In fact, $aecf_1 < aecf_2$, while the dissimilarity measures between conditional distributions all favor split $s_1$. Hence, the concentration upon equally-sized subsets leads the criteria to ignore the simplifying split $s_1$ in favor of $s_2$. To emphasize the different impact of the $aecf$ on $C_{GO}$, $C_{TO}$ and $C_A$, compare $s_1$ and $s_3$ in Table 2. $s_1$, appearing preferable to $s_3$, is selected as the best by $C_{GO}$ and $C_{TO}$. On the contrary, $C_A$ selects $s_3$ as the best, hence appearing more influenced by the $aecf$.

These considerations show that in some cases it may be sensible to remove the tendency of splitting criteria to favor splits producing sub-nodes of similar sizes, still avoiding the opposite tendency to obtain very small off-springs. To attenuate the effects of the $aecf$ without totally eliminating it, T & S suggest to *adapt* the $aecf$ according to the *complexity* of a node. The simplest measure of the complexity of a node $t$ is the *class number index*, $m(t)$, the number of levels of $Y$ in $t$. An alternative measure, taking into account not only the number of levels assumed by $Y$ but also their relevance within $t$, is the so called *reciprocal entropy index*:

**Table 2** The effects of the *aec* factor

| $s_1 \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | | |
|---|---|---|---|---|---|---|
| $t_L^1$ | 22 | | | | 22 | $C_{GO} = 0.1716 \cdot 1.692$ |
| $t_R^1$ | | 21 | 26 | 31 | 78 | $C_{TO} = 0.1716 \cdot 1$ |
| | 22 | 21 | 26 | 31 | 100 | $C_A = 0.1716 \cdot 1$ |
| $s_2 \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | | |
| $t_L^2$ | 22 | 18 | | | 40 | $C_{GO} = 0.2400 \cdot 1.4719$ |
| $t_R^2$ | | 3 | 26 | 31 | 60 | $C_{TO} = 0.2400 \cdot 0.9025$ |
| | 22 | 21 | 26 | 31 | 100 | $C_A = 0.2400 \cdot 0.9775$ |
| $s_3 \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | | |
| $t_L^3$ | 22 | 3 | | | 40 | $C_{GO} = 0.1875 \cdot 1.5228$ |
| $t_R^3$ | | 18 | 26 | 31 | 60 | $C_{TO} = 0.1875 \cdot 0.7744$ |
| | 22 | 21 | 26 | 31 | 100 | $C_A = 0.1875 \cdot 0.9712$ |

$$m^*(t) = \left[ \sum_{g=1}^{G} \pi_t^2(g) \right]^{-1} = [1 - I_N(t)]^{-1} .$$

The complexity of a node depends now upon its impurity. It is simple to see that when $Y$ is uniformly distributed $m(t) = m^*(t)$. On the basis of the complexity of a node a threshold is obtained, $p_{LOW} \leq 0.5$, and the same $aecf$ is associated with all the splits for which $p_L$ and $p_R$ are both greater than $p_{LOW}$. Thus, the penalization for asymmetrical splits is applied only to splits leading to sub-nodes with a proportion of cases lower than $p_{LOW}$. According to this criterion, an adaptive $aecf$ for a given $p_{LOW}$ is defined as:

$$AEC(p_L) = \min[p_L(1 - p_L), p_{LOW}(1 - p_{LOW})].$$

T & S introduce the *basic* adaptive $aecf$ (*ba-aecf*), based upon $m(t)$, and the *enhanced* adaptive $aecf$ (*ea-aecf*), based upon $m^*(t)$. The thresholds are respectively

$$p_{LOW} = [1/m(t)] \qquad p^*_{LOW} = \min\left[(1/2), 1/m^*(t)\right].$$

### 3.3 An application to real data

To compare the considered criteria we now refer to a data set consisting of $N = 1{,}002$ observations on Italian investors, customers of an Italian firm. The response variable is an ordinal qualitative variable having six levels and measuring the "importance" of an investor for the considered firm (as stated by the firm itself on the basis of the amount of capital invested in years subsequent to the entrance, to portfolio's diversifications and so on). The response has to be predicted on the basis of information about the first

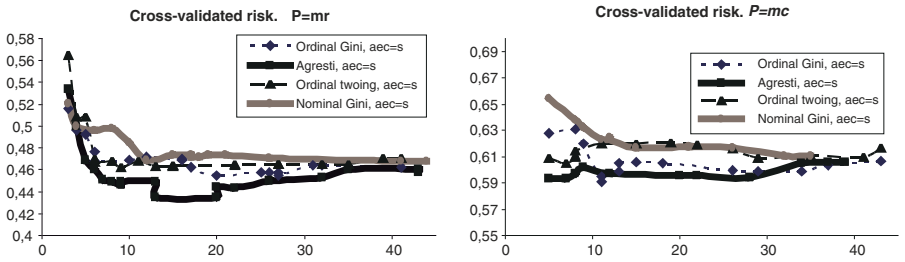


**Fig. 2** Cross-validated risk for different criteria

investment choices (the amount of invested capital, the first form of investment) and of social-demographic characteristics.

The trees were grown using the different criteria, each combined with the standard and the (two) adaptive $aecf$. The obtained trees were then pruned and cross-validated (tenfold cross-validation was considered). The first measure of risk considered to evaluate and to prune a tree was the misclassification rate, $mr$. To take into account also the ordinality of $Y$, a cost was then assigned to each misclassification error. The cost of misclassifying a class $j$ object as a class $g$ object was set equal to $c_{g|j} = |g - j|$: it depends only on the "distance" between $j$ and $g$. Actually, we are not attaching costs reflecting "avversion" to particular errors (this is done to avoid the comparison to be affected by subjective choices), and the matrix of misclassification costs is symmetric. Of course, in applications different structures of costs may be considered. Thus, each tree was pruned by referring to $mr$ and to $mc$ (in the following, $P = mr, mc$ will indicate the criterion used to measure the risk).

In Fig. 2, we report the values of the cross-validated risk as depending on the number of terminal nodes for trees pruned by referring either to $mr$ and to $mc$. Notice that ordinal criteria perform better than $C_{GN}$, even in the case when the tree is $mc$-pruned.

This is an interesting point, since one may be convinced that the ordinal problem may be taken into account only in the pruning phase (i.e., the tree is grown using a nominal criterion and then pruned using an ordinal criterion). In this example, instead, we observe that when the relationship between the ordinal response and the explanatory variables is monotone, a fully ordinal approach is preferable either w.r.t. the $mr$ and w.r.t. the $mc$.

Nevertheless in other applications, in particular in situations when the relationship between the predictors and the response variable is not so strongly monotone, we observed a good performance of the nominal Gini–Simpson criterion, at least w.r.t. $mr$. In this sense, taking into account either ordinal and nominal criteria does not only permit to build different trees—which is in itself relevant from an exploratory point of view—but it also permits to inspect the kind of relationship (monotone or not) between the response variable and the explanatory ones. Moreover, in many applications we observed that ordinal criteria often lead to trees with intermediate subnodes characterized by adjacent levels of the ordinal response. A better performance (as compared to the nominal criterion) was noticed either in terms of $mc$ and in terms of ordinal–ordinal measures of association (Spearman coefficient, Kendall tau-b coefficient) between the response variable and its prediction.

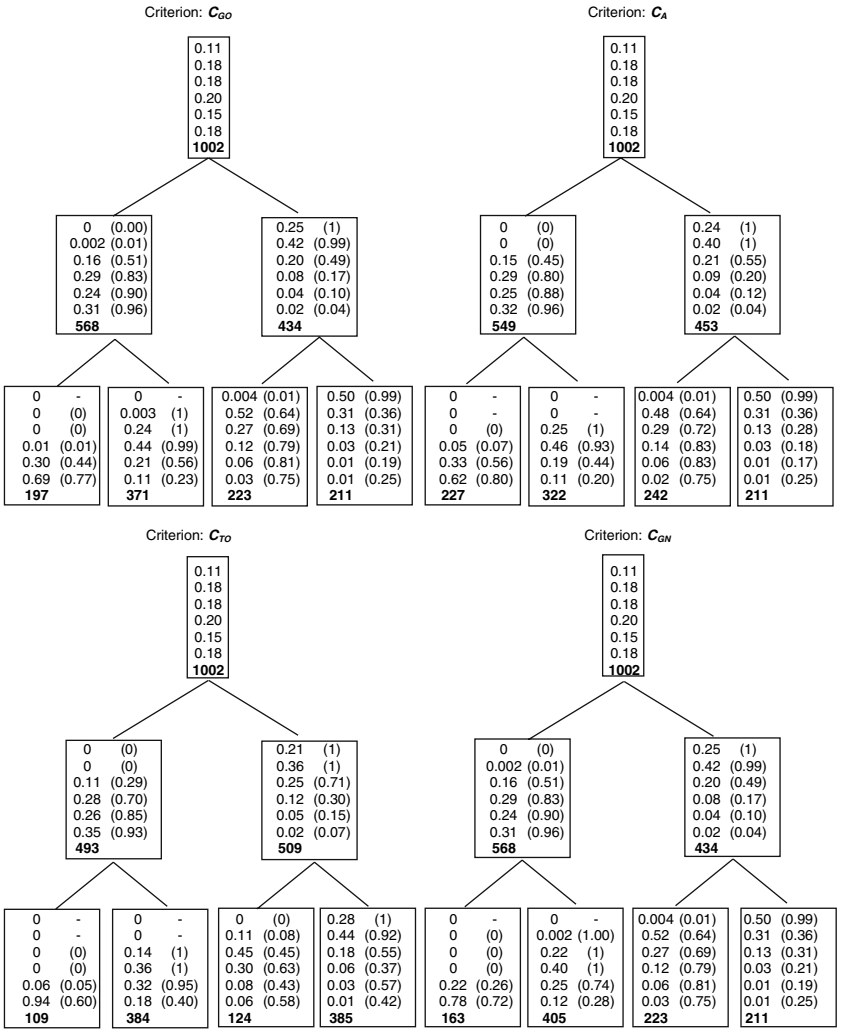


**Fig. 3** First splits for all criteria ($aec = s$)

Among the ordinal criteria, we observe that the best performance is attained by $C_A$, and that $C_{GO}$ performs slightly better than $C_{TO}$. Our aim is not to state conclusions about the performances of the criteria on the basis of these results. In fact, we are mainly interested in showing that different criteria lead to different trees, and that our criteria enrich the class of ordinal criteria.

Actually, performance of a classifier is not the only relevant aspect when comparing trees. A very important issue concerns the comparison of the "structure" of the intermediate nodes characterizing the trees. To better emphasize the "qualitative" differences in the splits characterizing the trees, in Fig. 3 we report the first splits of the trees grown using $C_{GO}$, $C_A$, $C_{TO}$, and $C_{GN}$.

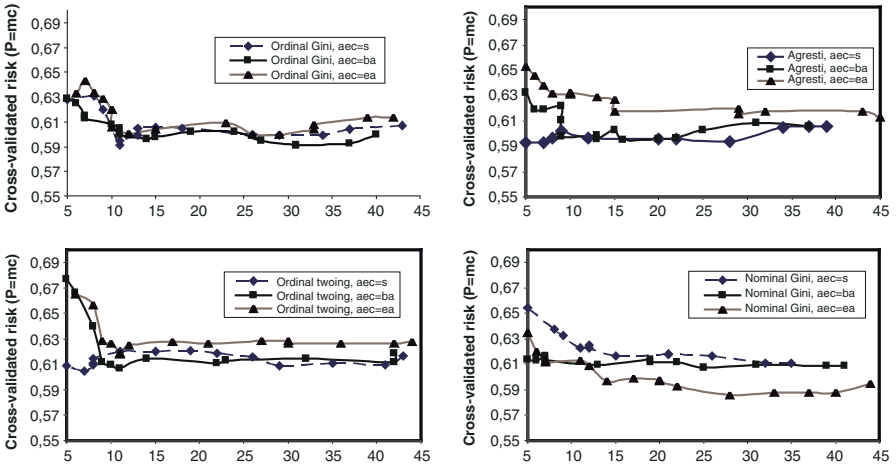


**Fig. 4** Cross-validated risk and $aec$ ($P = mc$)

For each node in the figure the conditional distribution of the response within the node itself is reported. To emphasize the degree of separation between two sub-nodes we also report, for a given class of the response, the proportions of cases sent to each sub-node (i.e., for class $j$ we report $p(t_L|j)$ and $p(t_R|j) = 1 - p(t_L|j)$).

We immediately observe the monotone exclusivity preference property "at work" for $C_A$ and $C_{TO}$. Both criteria attempt at individuating pure nodes, characterized by the absence of low classes. A comparison with $C_{GO}$ evidence that this criterion "sacrifice" purity to obtain nodes "absorbing" a high proportion of high-level classes. Moreover, notice that $C_A$ is more influenced than other criteria by the presence of $aecf$, since we observe sub-nodes with similar size.

As concerns $C_{GN}$, we observe a first split identical to that of the $C_{GO}$-tree. To appreciate the difference between the two criteria, compare the splits of the left-node in the first level. $C_{GN}$ leads to left sub-node characterized by a strong presence of class-6 cases; also class-5 cases are in the node, even if the highest proportion of class-5 cases is sent to the right sub-node. Notice that the proportion of classes 5 and 6 cases in the right sub-node is higher than its counterpart in the $C_{GO}$-tree, appearing more able to absorb a higher proportion of higher-level cases.

In Fig. 4, the cross-validated risk as depending on the number of terminal nodes is reported for trees ($mc$-pruned) obtained by combining the adaptive $aecf$ with considered criteria.

Notice that, at least in this application, the ordinal criteria do not suffer very much for the presence of the $aecf$: trees obtained by adapting the $aecf$ are not better than the "standard" ones. Different conclusions can be drawn for the $C_{GN}$: trees combined with adaptive $aecf$ perform better. To understand why, we recall that in their work T & S pointed out that the use of adaptive $aecf$ in conjunction with $C_{GN}$ leads to trees having similar performances in terms of $mr$. They moreover emphasized that when considering $C_{GN}$-trees for an ordinal response, one of the main consequences

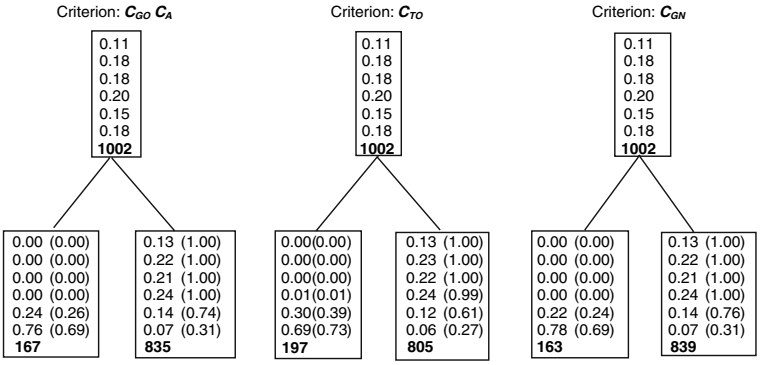


**Fig. 5** First splits for all criteria, $aec = ba$

of adapting the $aecf$ was that intermediate sub-nodes were constituted by adjacent levels of the response (so that a decrease in the $mc$ should be observed).

We can observe a stronger tendency (as compared with the "standard" tree) to separate cases characterized by the lowest levels of the response from the others. In this sense, the presence of the adaptive $aecf$ helps in identifying asymmetrical sub-nodes, isolating classes also in the first splits. Similar results can be observed (at least when considering the first splits) when the $aecf = ea$ is used. This difference between ordinal and nominal criteria is maybe due to the fact that our criteria still lead to sub-nodes possibly constituted by adjacent levels of the response, so that adapting the $aecf$ has not so significant consequences in this context. Nevertheless, as T & S point out, adaptive $aecf$ can prove useful in individuating small and well identified splits from the first phases of the segmentation process. To verify this, in Fig. 5 we report the first splits of the trees obtained with $aec = ba$.

We tested the impact of adaptive $aecf$ on ordinal trees generated for different data sets. The ordinal criterion appearing more influenced by the $aecf$ is $C_A$. As for $C_{GO}$ and $C_{TO}$, we often observed similar trees and similar performances (both in terms of $mr$ and $mc$) of the standard and of the adapted algorithms. We also considered measures of ordinal–ordinal association between the response and the predicted variable and conclusions were similar. As evidenced in this application, the criterion most influenced by the $aecf$ is $C_{GN}$.

## 4 Performance evaluation

In this section the performance of the splitting criteria will be compared via simulations. As it was evidenced in the previous section, ordinal criteria perform better than the nominal one (eventually combined with a $mc$-pruning phase) in cases when the relationship between the response variable and the explanatory variables is ordinal. This is the reason why we here consider simulations where the relationship between the response and the explanatory variables is structured but not strongly ordinal. Thus, we are interested in evaluating and comparing the capability of ordinal criteria to ordinally-predict the response variable in a not too favorable simulation.

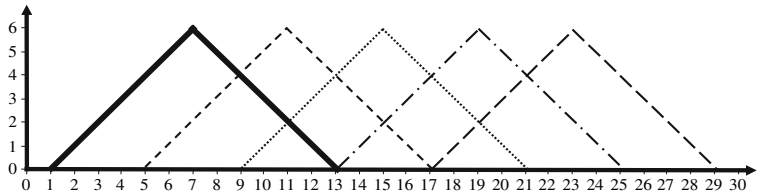


**Fig. 6** Waveforms referred to in simulations

At this aim, we adapted the waveform recognition model, described by B & A (p. 49), to our problem. This model is based upon the waveforms $h_1(t)$, $h_2(t)$, ... graphed in Fig. 6. The $i$-th function is centered on a value, indicated by $c_i$. The vertices of the $i$-th triangle will be denoted by $l_i$ and $u_i$: $l_i = c_i - 6$, $u_i = c_i + 6$.

Let $G$ denote the number of levels of the response variable, $Y$. $N = (50 \cdot G)$ explanatory vectors were generated using prior probabilities $(1/G, \ldots, 1/G)$. Hence, there are approximately 50 cases for each class of $Y$. For each case, the vector of explanatory variables is obtained as a random combination of two waveforms, with noise added. The number of waveforms taken into account and the number of explanatory variables generated, $n_x$, depend upon $G$. More precisely, if $g^*$ is the number of waveforms necessary to obtain $G$ classes, the number of explanatory variables generated is $n_x = u_{g^*}$. To create an explanatory vector, $\boldsymbol{x} = [x_1, \ldots, x_{n_x}]$, a random number $u$ is sampled from a uniform distribution, and $n_x$ random numbers $\epsilon_1, \ldots, \epsilon_{n_x}$ are independently sampled from a standardized normal distribution ($u$ is independent of $\epsilon_m$ for each $m = 1, \ldots, n_x$). Different explanatory vectors are then generated for each class:

$$\begin{aligned} \text{Class } \ j \qquad & x_m = u h_{\frac{(j+1)}{2}}(m) + (1-u) h_{\frac{(j+1)}{2}+1}(m) + \epsilon_m, \\ \text{Class } \ (j+1) \qquad & x_m = h_{\frac{(j+1)}{2}}(m) + (1-u) h_{\frac{(j+1)}{2}+2}(m) + \epsilon_m, \\ & \text{with } \ j = 1, 3, 5, 7 \ \text{ and } \ m = 1, \ldots, n_x. \end{aligned}$$

In what follows, a data set created as described above will be named a *waveform*-set. For each $G$, we generated $K = 40$ *training waveform-set*, $\mathcal{L}_{train,G,k}$, $k = 1, \ldots, 40$. For each $\mathcal{L}_{train,G,k}$, the *maximal tree* was built using $C_{GO}$, $C_A$, $C_{TO}$ and $C_{GN}$, each one combined with the standard *aecf* ($aec = s$), the *basic* adaptive *aecf* ($aec = ba$), and the *enhanced* adaptive *aecf* ($aec = ea$).

Summing up, we have:

$G$: number of levels of the response variables, $G = 3, 4, 5, 6, 7$;
$K$: number of simulations, $K = 40$;
$cr$: criterion to obtain the maximal tree, $cr = GO, A, TO, GN$;
$aec$: anti-end cut factor, $aec = s, ba, ea$.

Hence, we have 2,400 maximal trees, $T_{Max|G,k,cr,aec}$. Each maximal tree was *pruned* taking into account two measures of risk, namely $mr$ and $mc$[1]: thus, two se-

[1] Again, we set $c_{g|j} = |g - j|$.

quences of nested subtrees, $\{T_j\}_P$ were obtained, $P = mr, mc$. The tree characterized by the minimum risk (evaluated via tenfold cross-validation) is finally selected as the optimum-size tree, $T^*_{G,k,cr,aec,P}$: thus we have $2{,}400 \times 2 = 4{,}800$ classifiers. The trees are evaluated on the basis of a *test waveform-set*, $\mathcal{L}_{test,G,k}$, of $N = (50 \cdot G)$ observations. Each case in $\mathcal{L}_{test,G,k}$ a prediction is assigned.[2] The true category of $Y$ for $\mathcal{L}_{test,G,k}$-cases is known. Hence for each $T^*_{G,k,cr,aec,P}$ it is possible to evaluate the $\mathcal{L}_{test}$–$mr$ and the $\mathcal{L}_{test}$–$mc$, indicated by $M^{cr}_{G,k,aec,P}$ and $C^{cr}_{G,k,aec,P}$, respectively.

As concerns the "absolute" performance of the algorithms, the following measures of errors are considered:

$$\mathcal{M}^{cr}_{G,k,aec,P} = \frac{M^{cr}_{G,k,aec,P}}{M^{L}_{G,k}} \qquad \mathcal{C}^{cr}_{G,k,aec,P} = \frac{C^{cr}_{G,k,aec,P}}{C^{L}_{G,k}},$$

$M^L_{G,k}$ and $C^L_{G,k}$ being the $mr$ and the $mc$ of the *root* node of the *test set* (representing thus the risk incurred when predicting $Y$ ignoring information on the explanatory variables).

To further emphasize the differences between algorithms with respect to $mr$ and $mc$, we obtained the standardized measures of errors[3], $\mathcal{M}^{s,cr}_{G,k,aec,P}$ and $\mathcal{C}^{s,cr}_{G,k,aec,P}$. The syntheses of (standardized) errors for each $(G, aec, P)$, $\widehat{\mathcal{M}}^{s,cr}_{G,aec,P}$ and $\widehat{\mathcal{C}}^{s,cr}_{G,aec,P}$, are obtained by average.

We also considered $Min^{cr,\mathcal{M}}_{G,aec,P}$ and $Min^{cr,\mathcal{C}}_{G,aec,P}$, the number of times each criterion is a *strong* minimizer of the error (in terms of $\mathcal{M}$ or $\mathcal{C}$), reaching alone the minimum value of error. Besides these "marginal" measures of errors, also the number of times each criterion simultaneously minimizes ($\mathcal{M}$ and $\mathcal{C}$), $Min^{cr,\mathcal{MC}}_{G,aec,P}$ was calculated. Similarly, $Max^{cr,\mathcal{M}}_{G,aec,P}$, $Max^{cr,\mathcal{C}}_{G,aec,P}$, and $Max^{cr,\mathcal{MC}}_{G,aec,P}$ will denote the number of times a criterion is a maximizer of the errors.

For paired comparisons, the following quantities were taken into account

$$\Delta^{\mathcal{M}}_{G,k,aec,P}(cr_1, cr_2) = \mathcal{M}^{cr_1}_{G,k,aec,P} - \mathcal{M}^{cr_2}_{G,k,aec,P}$$
$$\Delta^{\mathcal{C}}_{G,k,aec,P}(cr_1, cr_2) = \mathcal{C}^{cr_1}_{G,k,aec,P} - \mathcal{C}^{cr_2}_{G,k,aec,P}.$$

All the measures used to compare the criteria were also adapted to evaluate the impact of the different kind of $aecf$ on the performance of the algorithms.

### 4.1 Comparison between splitting criteria

The performance of the criteria can be analyzed (1) conditionally to the $aecf$ (for a given $aecf$, we compare the relative performances of the criteria) and (2) conditionally to the criterion (for a given criterion, we analyze the impact of adapting the $aecf$).

[2] The prediction associated to each terminal node is the category of the response minimizing the risk in the node itself. The predictions only depend upon the observations in the training set.

[3] Of course, the standardization was carried out conditionally upon $G, aec, P$.

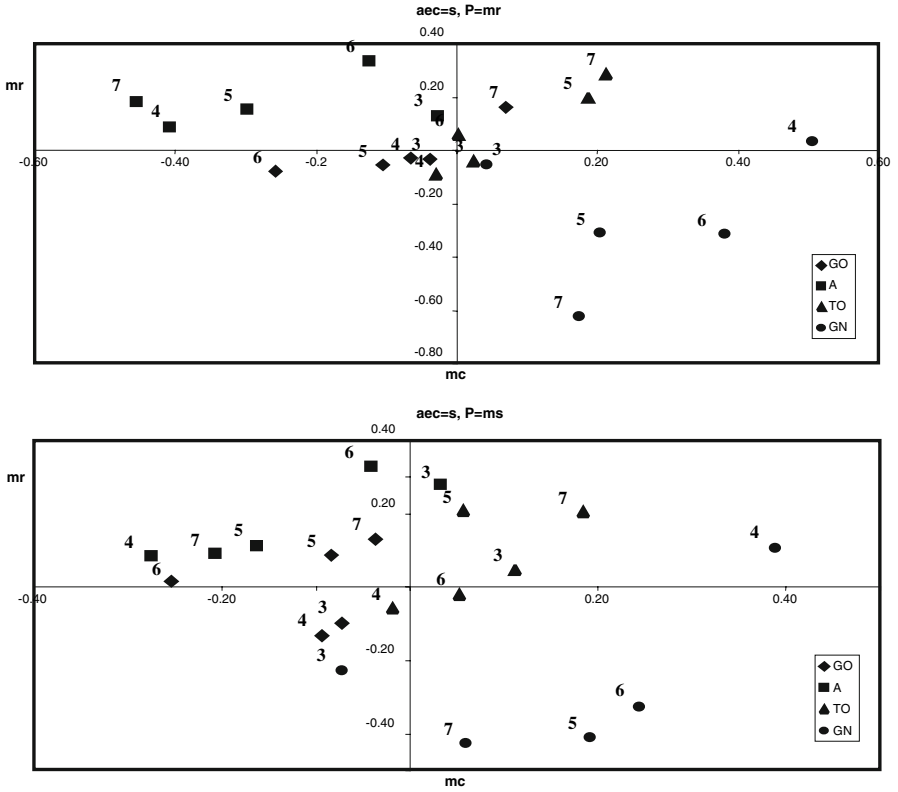


**Fig. 7** Scatter plots of $\widehat{\mathcal{M}}^s$ and $\widehat{\mathcal{C}}^s$ $(aec = s)$

**Comparison conditional to** $aecf$**.** We first consider the case when $aec = s$. In Fig. 7 the scatter plots of $\widehat{\mathcal{M}}^s$ and $\widehat{\mathcal{C}}^s$ are reported (for $P = mr$ and $P = mc$). Since the standardization was carried out conditionally to $G$, the measures of errors for different values of $G$ are not comparable.[4] Inspection of the scatter plots shows that we have a $mr$-oriented criterion, $C_{GN}$, and a $mc$-oriented criterion, $C_A$.

- $C_{GN}$ is often characterized by the lowest $mr$ (for a given $G$): ordinal criteria perform worse than $C_{GN}$ w.r.t. $mr$, even if a significant difference in performance is observed only between $C_A$ and $C_{GN}$ (the former criterion performs significantly worse than the latter). As for $C_{GO}$ and $C_{TO}$, the difference with $C_{GN}$ is significant (worse performance) only for the highest values of $G$.

Moreover, in many simulations the $C_{GN}$-tree is characterized by the minimum $mr$, but it rarely minimizes $mc$, and often maximizes it. As for $mc$, we observed that ordinal criteria perform significantly better than $C_{GN}$ also for low values of $G$. These results are somehow expected: $C_{GN}$ favors splits leading to pure nodes (which are not

[4] Actually, as $G$ increases a worse performance of all the considered algorithms can be noticed both in terms of $mr$ and in terms of $mc$.

necessarily ordinally pure), and may thus perform better w.r.t. $mr$. When using ordinal criteria, we are prepared to accept a (possibly not too high) increase of $mr$, to obtain a tree with terminal nodes characterized by (possibly) adjacent levels of the ordinal response, and hence by a lower $mc$.

These considerations also hold when the tree is pruned according to $mc$. Pruning nominal-trees taking into account the ordering of $Y$ is not "enough" to improve the performance of $C_{GN}$ w.r.t. $mc$ (as compared to ordinal criteria).

- Criterion $C_A$ is often characterized by a low $mc$ (and in many simulations it minimizes the $mc$). On the other hand, it often performs worse than the other algorithms w.r.t. $mr$ (it seldom minimizes $mr$ and often maximizes it). By analyzing the differences $\Delta^{\mathcal{M}}$ and $\Delta^{\mathcal{C}}$ between $C_A$ and the other criteria we noticed that $C_A$ has a significantly better performance w.r.t. $mc$ but has also a worse performance w.r.t. $mr$.

The two "oriented" criteria, $C_A$ and $C_{GN}$, while having a better "marginal" performance (resp. w.r.t. $mc$ and $mr$), have a poor "joint" performance, showing low values of $Min^{\mathcal{MC}}$ and high values of and $Max^{\mathcal{MC}}$.

- As concerns $C_{GO}$ and $C_{TO}$, they are more "balanced" w.r.t. the two kind of errors, showing a "medium" performance either w.r.t. $mr$ and w.r.t. $mc$. Analyzing the number of times their trees reach the minimum and the maximum errors, we observe a better performance, as compared to $C_A$ and $C_{GN}$). Nevertheless, we have to point out that $C_{GO}$ leads to classifiers often showing a (slightly) better $mc$-performance: for some values of $G$ the twoing ordinal criterion performs (significantly) worse than the other ordinal criteria. $C_{TO}$ maximizes $\mathcal{C}$ and $(\mathcal{M}, \mathcal{C})$ more often than $C_{GO}$ ($C_{GO}$ also has a slightly better "joint" performance).

Considerations above also hold when comparing criteria conditionally to adaptive $aecf$, even if a worsening of the relative $mc$-performance of $C_{GN}$ is observed. Moreover, when $aec = ba$ and (especially) when $aec = ea$ the relative $mr$-performance of $C_{GN}$ is not as good (with an exception for $G = 7$) as in the case when $aec = s$.

As a final consideration, we point out that the differences between the errors are seldom significant (especially between ordinal criteria). In our opinion this does not mean that the criteria are equivalent but, on the contrary, that they are "competitors". Actually, we observed even high differences in the performance relative to single simulations, meaning that classifiers are different. This is also confirmed when analyzing the correlation coefficients between the errors characterizing the different classifiers: they are low and not significant. In this sense, we think that our criteria can be useful alternatives to $C_{TO}$ for growing classification trees in the case when the response is ordinal. As another conclusion, we think that *ordinal criteria should be referred to* when growing trees for ordinal responses.

**Comparison conditional to criteria.** An analysis similar to that described above was conducted also to evaluate the impact of the $aecf$ (standard and adaptive) on the performances of the splitting criteria. We summarize very briefly the main results of simulations.

As for ordinal criteria, $C_A$ appears mostly influenced by the $aecf$. In particular, when $P = mr$ the (combination with the) *ea-aecf* leads to classifiers showing a

better performance either w.r.t. $mc$ and w.r.t. $mr$ (except when $G = 7$). The $s$-$aecf$ is instead characterized by the worse performance. This last consideration also holds when $P = mc$ (even if the relative performance of the two adaptive $aecf$ depends upon $G$).

Concerning $C_{GO}$ and $C_{TO}$ we can not draw conclusion about the effects of the adaptation of the $aecf$. Moreover, in some cases a worsening of the performances was observed when adapting the $aecf$, either in terms of $mr$ and in terms of $mc$.

With regard to the nominal criterion, the best performance is observed when $aec = s$.

The above mentioned results confirm the considerations of T & S. The adaptation of the $aecf$ is not necessarily expected to lead to trees characterized by a better performance in terms of $mr$ and $mc$ but, rather, by a better "characterization" of the intermediate sub-nodes.

## 5 A property of the ordinal Gini–Simpson criterion

At a first glance, it may seem criterion $C_{GO}$ to be more computationally expensive than its nominal counterpart. Actually, when growing a tree using the nominal criterion, it is necessary to obtain the conditional distributions of the response variable into the two sub-nodes induced by a given split. When considering the ordinal criterion, the cumulative conditional distributions have to be determined. It is immediate to notice that when the number of categories of the response variable is small, this further calculation is not excessively costly from a computational point of view.

Nevertheless, the evaluation of the conditional distributions for all the possible splits of a given node is computationally intensive in itself. On the basis of this consideration, Mola and Siciliano (1997, 1998) introduce an algorithm to fasten the growing phase in the case when the response variable is nominal, and the Gini–Simpson criterion is used to evaluate the splits.

We now extend the procedure to the ordinally extended Gini–Simpson criterion. We recall that the criterion to evaluate a split $s$ of a node $t$ is:

$$C_{GO}(s,t) = p_L p_R \sum_{g=1}^{G-1} [F_t(g|L) - F_t(g|R)]^2. \tag{9}$$

Now consider the following measure of the impurity within node $t$:

$$I_O(t) = \sum_{g=1}^{G} F_t(g) \cdot [1 - F_t(g)]. \tag{10}$$

As it was mentioned in Sect. 3, $I_O(t)$ is a measure of the heterogeneity of an ordinal variable originally proposed by Gini (1954).

It can be easily shown that:

$$C_{GO}(s,t) = I_O(t) - p_L I_O(t_L) - p_R I_O(t_R) \tag{11}$$

Now recall that a split $s$ is defined on the basis of the values assumed by one out of the explanatory variables, say $X_I$, the subscript indicating the number of categories of $X$. In particular, consider a split $s$ defined on the basis of $X_I$. $s$ will divide a node $t$ into two sub-nodes, $t_L$ constituted by cases having values of $X_I$ in a given set, say $A$, and $t_R$ constituted by cases having values of $X_I$ not in $A$.

Suppose now that a node $t$ is split, according to the levels of $X_I$, into $I$ sub-nodes. By extending (11) to this situation, we obtain:

$$C_{GO}(X_I, t) = I_O(t) - \sum_{i=1}^{I} p_i I_O(t_i), \tag{12}$$

$t_i$ and $p_i$ being, respectively, the node constituted by all case in $t$ characterized by the $i$-th value of $X_I$ and the proportion of cases in $t$ placed in $t_i$.

This quantity is the numerator of the index:

$$\tau_{O(t)}(Y|X_I) = \frac{I_O(t) - \sum_{i=1}^{I} I_O(i|t)p_t(i)}{I_O(t)}, \tag{13}$$

introduced by Piccarreta (2001) to measure the strength of the nominal–ordinal association between the (nominal) predictor $X$ and the (ordinal) response variable $Y$ (here this association is evaluated restricting attention only to cases in $t$).

This consideration emphasizes that when evaluating a split as in (11) we are substantially evaluating the strength of the association between the response variable $Y$ and the binary variable defined by grouping the categories of an explanatory variable. Hence, the ordinal Gini–Simpson criterion searches for the splitting variable (and the grouping of the categories of the splitting variable) predicting the response variable in the best possible way (according to the predictability index $\tau_O$).

Suppose now that two categories of $X_I$, $i_1$ and $i_2$, are grouped into a combined category $i(12)$, and let $X_{I-1}$ denote the resulting categorical variable. Of course, the number of sub-nodes defined on the basis of $X_{I-1}$ will be $(I-1)$. It will be:

$$C_{GO}(X_{I-1}, t) = I_O(t) + \sum_{i=i_1,i_2}^{I} p_i I_O(t_i) - p_{i(12)} I_O(t_{i(12)}), \tag{14}$$

and

$$C_{GO}(X_I, t) - C_{GO}(X_{I-1}, t) = p_{i(12)} I_O(t_{i(12)}) - p_{i_1} I_O(t_{i_1}) - p_{i_2} I_O(t_{i2}). \tag{15}$$

Recalling (10) and observing that:

$$\begin{aligned} p_{i(12)} &= p_{i_1} + p_{i_2} \\ F_{t_{i(12)}}(g) &= \frac{p_{i_1} F_{t_{i_1}}(g) + p_{i_2} F_{t_{i_2}}(g)}{p_{i(12)}}, \end{aligned}$$

it can be easily shown that:

$$C_{GO}(X_I, t) - C_{GO}(X_{I-1}, t) = \frac{p_{i_1} p_{i_2}}{p_{i(12)}} \sum_{g=1}^{G} \left( F_{t_{i_1}}(g) - F_{t_{i_2}}(g) \right)^2 \geq 0. \quad (16)$$

So, the main result here is that if two categories of $X_I$ are collapsed, the heterogeneity of the new partition is greater than the heterogeneity of the partition induced by the $I$ categories of $X_I$. This leads, by induction, to the following stating: if $s_X$ is a split defined on the basis of the explanatory variable $X$, it is:

$$C_{GO}(X, t) \geq C_{GO}(s_X, t) \quad (17)$$

We now recall that to select the best split for a node $t$ we have (1) to take into account all the explanatory variables, (2) for each explanatory variable we have to try all the possible splits defined on the basis of its categories. We have then to select the best split for each variable and then the best split among all these best splits. Property (17) of the ordinal Gini–Simpson criterion, can be used to find the best split of a node without trying out all possible splits.

Consider in fact a node $t$ and two explanatory variables, say $X$ and $Z$, and suppose it is $C_{GO}(X, t) \geq C_{GO}(Z, t)$. Let $s_X^*$ and $s_Z^*$ be the best splits for node $t$ which can be defined on the basis of $X$ and $Z$, respectively. If $C_{GO}(s_X^*, t) \geq C_{GO}(Z, t)$, then it will be $C_{GO}(s_X^*, t) \geq C_{GO}(s_Z^*, t)$, so that it is not necessary to evaluate the splits which can be defined on the basis of $Z$.

Hence the algorithm proceeds as follows. At the first step $C_{GO}(X_q, t)$ is calculated for all the explanatory variables. Let $(X_{(1)}, X_{(2)}, \ldots, X_{(q)})$ be the explanatory variables in an increasing order according to $C_{GO}(X, t)$. $X_{(1)}$ is the variable maximizing $C_{GO}(X, t)$): all the possible splits defined on the basis of $X_{(1)}$ are evaluated, and the best is selected, $s_{X_{(1)}}^*$. Now $C_{GO}(s_{X_{(1)}}^*, t)$ is compared with the criterion evaluated for the second explanatory variable, $C_{GO}(X_{(2)}, t)$. If $C_{GO}(s_{X_{(1)}}^*, t) > C_{GO}(X_{(2)}, t)$ then the algorithm can be stopped, since $s_{X_{(1)}}^*$ is the best split. Otherwise, all the splits of $X_{(2)}$ are evaluated, and the best one is selected, $s_{X_{(2)}}^*$. The best splits for $X_{(1)}$ and $X_{(2)}$ are compared, the best one is selected and the procedure goes on by comparing the best split with $X_{(3)}$. The algorithm iterates until there are no more explanatory variables characterized by a value of the criterion lower than the value characterizing the best (current) split.

Notice that in this way, attention can be limited only to the most promising explanatory variables, while explanatory variables with a low value of the criterion (in particular, lower than the value characterizing the best current split) can be disregarded.

The computational gain in terms of the reduction of the number of splits inspected and of CPU time is illustrated in Mola and Siciliano (1997, 1998).

## 6 Conclusions and directions of future research

The problem of building a classification tree in the nominal case has been given great attention in literature. The original proposal by B & A has been deeply analyzed and

criticized and many methods to improve it have been introduced. The case when the response is ordinal has not received the same attention.

In this paper, we illustrate why ordinal criteria should be used in the ordinal case. We analyze the twoing ordinal criterion, evidencing its characteristics and properties. We also introduce possible alternative ordinal criteria. The aim is to enrich the class of ordinal splitting criteria. Actually, we are convinced that the availability of more criteria and the inspection of different trees can also prove very useful from an exploratory point of view, allowing insights into the data to be analyzed.

The criteria are first theoretically compared and analyzed, mainly referring to some drawbacks illustrated by T & S for the Gini–Simpson criterion, namely the exclusivity preference property and the possible—excessive—influence of the $aecf$ on the tree-classifiers. As for this point, applications and simulations emphasize that in the ordinal case the problem is not so relevant as it is in the nominal case.

With regard to the comparison (via simulations) between ordinal criteria, we can conclude that none of them appears preferable to the others. $C_A$ often performs better in terms of $mc$, but is characterized by a trade-off between $mr$ and $mc$. As for the "global" performance, criteria $C_{GO}$ and $C_{TO}$ appear more promising. At least in simulations they often lead to trees with low $mc$ without sacrificing too much in terms of $mr$, as compared to results obtained by referring to $C_{GN}$. With respect to this consideration, it is interesting to point out that both $C_{GO}$ and $C_{TO}$ are based upon a measure of distance between the two cdf's in the sub-nodes induced by a split. As a direction of further research, we think it would be sensible to enrich the class of ordinal criteria by considering different measures of distance between the two cdf's (see e.g. Shih 1999, for this kind of analysis in the nominal case).

From another point of view, $C_{GO}$ is based upon a particular measure of the heterogeneity of an ordinal variable. It would be interesting to consider criteria based on different measures of heterogeneity.

## References

Agresti A (1990) Categorical data analysis. New York, Wiley
Agresti A (1981) Measures of nominal–ordinal association. J Am Stat Assoc 76:524–529
Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont
Gini C (1954) Variabilità e concentrazione. Veschi, Roma
Mola F, Siciliano R (1997) A fast splitting procedure for classification trees. Stat Comput 7:209–216
Mola F, Siciliano R (1998) A general splitting criterion for classification trees. Metron 56:156–171
Piccarreta R (2001) A new measure of nominal ordinal association. J Appl Stat 28:107–120
Shih Y-S (1999) Families of splitting criteria for classification trees. Stat Comput 9:309–315
Taylor PC, Silverman BW (1993) Block diagrams and splitting criteria for classification trees. Stat Comput 3:147–161