Predicting Ordinal Classes via Classification Trees

Radaelli, Paolo[1]

Borroni, Claudio G.
Department of Quantitative Methods for Economics and Business
Via B. degli Arcimboldi, 8
20126, Milano, Italy
E-mail: claudio.borroni@unimib.it

Zenga, Mariangela
Department of Quantitative Methods for Economics and Business
Via B. degli Arcimboldi, 8
20126, Milano, Italy
E-mail: mariangela.zenga@unimib.it

1 Introduction

CART (Classification and Regression Trees) is a non-parametric tree-structured recursive partitioning method, introduced by Breiman et al. [1984], to predict a response variable Y on the basis of p predictors: $X_1, \ldots, X_p$ observed on a learning sample of N units. The algorithm consists of two main stages: growing and pruning. In growing the tree is recursively partitioned into subsets (nodes); each partition is obtained by examining all the possible binary splits along the observed data of each predictor variable and selecting the split that most reduces some measure of node impurity. The result is a sequence of nested trees, with increasing number of leaves (terminal nodes), until no more splits are possible and the fully grown tree is reached. The pruning operation on the fully grown tree aims then to select the best subtree and consists in declaring an internal node as terminal and deleting all its descendants; this makes the tree more general and prevents an overfitting on the training set.

In this paper the case where the response Y is an ordered categorical variable with k levels $y_1 \prec \ldots \prec y_k$ is considered. The aim of the classification tree is thus to predict the level of Y on the basis of the vector $\mathbf{X}$ of the p explanatory variables. For example in a credit scoring application a bank is interested in classifying loan applicants into risk classes such as: "very low", "low", "medium", "high", "very high", according to their characteristics (monthly income, outstanding debt, financial assets, age,...). The tree is grown according to a training set of N cases whose measurements, both for the response and for the predictors, are available. The derived classification rule is then applied to predict the level of the response for a new unit with explanatory variables vector $\mathbf{x}$.

Predicting the ordinal classes can thought to be somewhat intermediate between classification and regression trees; however, while classification trees for unordered categorical variables and regression trees have been widely studied, the use of decision trees in ordinal regression is largely unexplored. By ignoring the ordering information in the class attribute, standard classification algorithms for nominal classes can be applied but some information is lost. This fact prejudices the predictive performance of the classification rule because, besides the accuracy, the severity of the error should be taken into account. On the other side the use of regression trees requires the ordinal response to be transformed into a numeric one. However, the obtained regression trees may largely depend on the mapping procedure adopted rather than on the ordinal relationship among classes.

Instead of accommodating existing algorithm to the ordinal classification task, some papers face directly the problem of identifying a suitable rule to grow ordinal classification trees. Notable works are

[1]Dr Paolo Radaelli passed away unespectedly at the age of 33 on October 19, 2009.

the ones by Xia et al. [2006] and Piccarreta [2008]. In the former the authors propose the use of a new impurity measure named as *ranking impurity* while in the latter new criteria to obtain classification trees for ordinal response are introduced and compared with other methods via simulations. We will consider these proposals into details in the next sections and propose some improvements based on a known decomposition result regarding Gini's mean difference.

2 Impurity measures for ordinal variables

Consider a generic node with N cases and the dependent variable having k levels. The number of cases and the relative proportion of class j are denoted by n_j and p_j $(j = 1, \ldots, k)$, respectively. Denotes with $\mathbf{p} = (p_1, \ldots, p_k)$ the proportion vector in the current node.

The most used function to measure node impurity when the response is nominal is the *Gini heterogeneity index*:

$$I_N(\mathbf{p}) = \sum_{i \neq j} p_i \; p_j = \sum_{j=1}^{k} p_j \;\; (1 - p_j) = 1 - \sum p_j^2 \tag{1}$$

which assumes that the cost of misclassifying a $j-$class case into a class i is equal to 1 for all $i \neq j$. I_N satisfies the properties usually required to an impurity function (see Breiman et al. [1984] p. 32), as it takes its minimum value 0 iff all cases of the node belong to the same class i.e. the node is as pure as possible; conversely, the maximum value is reached iff the proportion vector is $\mathbf{p} = (1/k, \ldots, 1/k)$ and the node is as impure as possible. Moreover $I_N(\mathbf{p}) = I_N(g(\mathbf{p}))$, $g(\mathbf{p})$ being a permutation of the elements of $\mathbf{p}$.

When the categories of the response Y are naturally ordered, (1) is inadequate to measure the impurity of the node given that the misclassification cost of a case depends also on the number of categories between the assigned and the actual class. Consider a loan applicant with actual risk class "high"; misclassifying this case as "low" is doubtless more serious than misclassifying it as "medium". For the same reason it seems improper to let the impurity measure be invariant to permutations of the proportion vector: (0.5, 0.3, 0, 0, 0.2) is more impure than (0.5, 0.3, 0.2, 0, 0). Even if the purity concept is strictly related to the concentration of cases on one or few classes (regardless of which), we thus need a measure accounting for the dispersion of cases among classes as well. A suitable *dispersion measure* has been introduced by Gini [1912]. After denoting as $F(y_j) = F_j = \sum_{i=1}^{j} p_i$ the cumulative distribution function (cdf) of Y evaluated at y_j, the impurity of the node is thus measured by:

$$D(\mathbf{p}) = 2 \sum_{j=1}^{k-1} F_j \;\; (1 - F_j)\,. \tag{2}$$

$D(\mathbf{p})$, as $I_N(\mathbf{p})$, takes its minimum value 0 iff all cases of the node belong to the same class; the maximum value $(k-1)/2$ is instead reached iff the proportion vector is $\mathbf{p} = (1/2, 0, \ldots, 0, 1/2)$ i.e. the cases are equally separated on the two extreme classes. $D(\mathbf{p})$ is equivalent to the measure:

$$I_O(\mathbf{p}) = \sum_{j=1}^{k} F_j \;\; (1 - F_j)\,. \tag{3}$$

considered in Piccarreta [2008]. Another impurity measure to be used in the ordinal case is the *ranking impurity*

$$I_{rank}(\mathbf{p}) = \sum_{j=1}^{k} \sum_{i=1}^{j} (j - i) \; n_j \; n_i \tag{4}$$

proposed by Xia et al. [2006].

Measures (3) and (4) are easily shown to be equivalent. Replace the k ordered categories of the response Y with the set of integers $1, 2, \ldots, k$ and consider the Gini mean difference with repetition (gmd), Gini [1912], of the node distribution:

$$\Delta(\mathbf{p}) = \frac{1}{N^2} \sum_j \sum_i |j - i| \; n_j \; n_i = \sum_j \sum_i |j - i| \; p_j \; p_i = 2 \sum_j \sum_{i<j} (j - i) \; p_j \; p_i. \tag{5}$$

It follows that $\Delta(\mathbf{p}) = D(\mathbf{p})$ (see for instance Leti [1983]). Hence

$$I_O(\mathbf{p}) = N^2 \; I_{rank}(\mathbf{p}) = \frac{1}{2} D(\mathbf{p}). \tag{6}$$

As the scale factor in the impurity function does not influence the splitting rule, using (2), (3) or (4) will then lead to the same tree.

3 Splitting rules for ordinal variables

The tree growing phase starts with all the N cases in a single node, the root. In the following stages, the algorithm performs an optimal search for a suitable division of each node into two disjoint subnodes according to one of the predictors. For every binary split s, denote with L and R the two subnodes obtained, and let $n_L, \; n_R$ and $\pi_L, \; \pi_R$ be the number and the proportions of cases (of the parent node) placed into L and R, respectively ($\pi_L + \pi_R = 1$). The number of cases and the relative proportion of class j in the subnodes are denoted, respectively, as n_{jt} and $p_{jt}, \; j = 1, \ldots, k, \; t \in \{L, R\}$ and the proportion vectors as $\mathbf{p}_t = (p_{1t}, \ldots, p_{kt}), \; t \in \{L, R\}$.

Given an impurity function I, descendant nodes should be less impure than their parents. Thus the algorithm searches the split s^* that maximizes the impurity reduction

$$I(\mathbf{p}) - \pi_L \; I(\mathbf{p}_L) - \pi_R \; I(\mathbf{p}_R). \tag{7}$$

As shown in section 2, the impurity functions (3), (4) and (5) are proportional; consequently maximizing (7) leads to the same split when such functions are used.

Besides splitting rules operating on an overall measure of node impurity to achieve the maximum impurity reduction, some different strategies can be adopted. An interesting proposal by Piccarreta [2008] makes use of a measure introduced by Agresti [1981] to evaluate the degree of association between a nominal and an ordinal variable. The idea is to consider the nominal variable induced by a split s, whose categories are L and R, and to evaluate its association with the response Y by the index:

$$A = \sum_{j=1}^{k} \sum_{i>j} p_{jL} \; p_{iR} - \sum_{j=1}^{k} \sum_{i>j} p_{jR} \; p_{iL} = \sum_{j=1}^{k} p_{jL} \, (1 - F_{jR}) - \sum_{j=1}^{k} p_{jR} \, (1 - F_{jL}) \, . \tag{8}$$

Piccarreta [2008] shows that $-1 \leq A \leq +1$ with $|A| = 1$ iff the split s originates two non-overlapping subnodes i.e. s induces a so-called *ordinal exclusive split*. Thus the use of the following splitting criterion is proposed:

$$C_A(s) = \pi_L \pi_R \; |A|. \tag{9}$$

This latter proposal appears to be very attractive when dealing with ordinal variable because, conditionally on $\pi_L \pi_R$, (9) prefers splits that place individuals with a level on the response up to a certain class in one subnode, say L, and the remaining individuals in the other subnode R. This means that, considering again the credit ranking example, loan applicants with a risk level up to (for example) "medium" are mainly placed in the left subnode while applicants with risk level higher than "medium"

Table 1: Agresti index and splits evaluation given π_L and π_R

Split	$\mathbf{p}_L$	$\mathbf{p}_R$
s_1	(0.60, 0.40, 0, 0, 0)	(0, 0, 0.50, 0.25, 0.25)
s_2	(0.90, 0.10, 0, 0, 0)	(0, 0, 0.50, 0.25, 0.25)
s_3	(0.90, 0.08, 0.02, 0, 0)	(0, 0, 0.50, 0.25, 0.25)
s_4	(0.90, 0, 0.10, 0, 0)	(0, 0.30, 0, 0, 0.70)
s_5	(0.90, 0, 0, 0.10, 0)	(0, 0.30, 0, 0, 0.70)

stay in the right subnode. Moreover Piccarreta [2008] showed, via simulations, that this splitting criterion has a good performance with respect to the other ordinal criteria, when the comparison is based on the miscalssification cost.

As a final remark, recall that Cerioli [1990] (see also Cerioli [1988, 1990]) proved an interesting relation between the Agresti index (8) and the transvariation probability measure proposed by Gini [1916]. Without entering into details, recall that if $Y_t,\ t \in \{L, R\}$ are the categories of two randomly selected units from subnodes L and R and $Me_t,\ t \in \{L, R\}$ denote the medians of the corresponding distribution, the two units are said to be *transvariant* (with respect to the median) if either of the following relations holds:

$$\{Y_L \succ Y_R | Me_L \prec Me_R\} \qquad \text{or} \qquad \{Y_L \prec Y_R | Me_L \succ Me_R\}\,.$$

4 A new splitting criterion based on Gini's mean difference decomposition

Piccarreta [2008] showed that, given π_L and π_R, the splitting rule (9) evaluates all the exclusive splits as equivalent. Consider for example splits s_1 and s_2 reported in Table 1: the Agresti rule does not allow us to distinguish between the two splits even if split s_2 seems to be preferable to s_1 because of the lower dispersion of individuals among the categories in the left subnode. Compare now splits s_1 and s_3: the Agresti rule favors again the exclusive split s_1, but actually s_3, in spite of a slight overlapping between the subnodes, appears to be better. These examples underline a weakness of the Agresti splitting criterion: given π_L and π_R, (9) does not give the right weight to the dispersion within subnodes, a relevant dimension when classifying individuals among the categories of an ordinal response. Another flaw regards the evaluation of the degree of overlapping (transvariation) between the subnodes obtained by the split. Compare for instance splits s_4 and s_5 reported in Table 1: the Agresti rule evaluates these splits as equivalent and hence it is not able to capture two main differences. First, as above noticed, it does not account for the different dispersion in the left subnode, that is lower for s_4 than for s_5. Secondly, the degree of overlapping appears to be higher for s_5 than for s_4 because the 10% left-subnode individuals "entering" in the right subnode distribution lie in a higher level for s_5 than for s_4.

The main idea of this paper is to overcome the drawbacks of the Agresti splitting criterion by introducing a new rule based on the same logic used in the decomposition by subgroups of one the most used and widespread inequality measures, the *Gini's concentration ratio*. Details of such a decomposition can be found in Dagum [1997] and in Costa [2008] for the case of two subgroups.

As shown in section 2, the dispersion measure (2) is equivalent to the gmd (5) computed when the ordered categories $y_1 \prec \ldots \prec y_k$ are replaced by the set of integers $1, \ldots, k$.

The gmd of the parent node can be rewritten as follows:

$$\begin{aligned}
\Delta(\mathbf{p}) &= \frac{1}{N^2}\sum_j\sum_i |j-i|\ n_j\ n_i \\
&= \frac{1}{N^2}\sum_j\sum_i |j-i|\ (n_{jL}+n_{jR})\ (n_{iL}+n_{iR}) \\
&= \frac{1}{N^2}\left[\sum_j\sum_i |j-i|\ n_{jL}n_{iL} + \sum_j\sum_i |j-i|\ n_{jR}n_{iR} + \sum_j\sum_i |j-i|\ n_{jL}n_{iR}\right. \\
&\qquad \left. + \sum_j\sum_i |j-i|\ n_{jR}n_{iL}\right] \\
&= \Delta_{LL}\pi_L^2 + \Delta_{RR}\pi_R^2 + \Delta_{LR}\pi_L\pi_R + \Delta_{RL}\pi_R\pi_L \\
&= \Delta_{LL}\pi_L^2 + \Delta_{RR}\pi_R^2 + 2\Delta_{LR}\pi_L\pi_R \qquad (10)
\end{aligned}$$

where$\Delta_{LL} = \Delta(\mathbf{p}_L)$ and$\Delta_{RR} = \Delta(\mathbf{p}_R)$ are the gmd evaluated *within* the two subnodes distributions, respectively, and

$$\Delta_{LR} = \frac{1}{n_L\ n_R}\sum_j\sum_i |j-i|\ n_{jL}n_{iR} = \sum_j\sum_i |j-i|\ p_{jL}p_{iR} = \Delta_{RL} \qquad (11)$$

is the mean difference *between* the left and the right subnodes as introduced by Dagum [1980].

The first two terms in (10) are the weighted sum of the offspring dispersion measure with weights given by the corresponding squared proportion of cases while the third term is the mean of the $n_L n_R$ differences between the observation of the two groups. This latter term can be further decomposed as in the following (see Gini [1916]). Denote with

$$\mu_t = \frac{1}{n_t}\sum_j j\ n_{jt}, \qquad t \in \{L, R\} \qquad (12)$$

the arithmetic means of the two offspring distributions and suppose, without loss of generality, that $\mu_L > \mu_R$. It can be easily shown that:

$$\begin{aligned}
n_L n_R \Delta_{LR} &= \sum_j\sum_i |j-i|\ n_{jL}n_{iR} \\
&= (\mu_L - \mu_R)\, n_L n_R + 2\sum_j\sum_{i>j}(i-j)n_{jL}n_{iR} \\
&= (\mu_L - \mu_R)\, n_L n_R + 2T_{LR} \qquad (13)
\end{aligned}$$

where T_{LR} denotes the sum of the transvariations (with respect to the arithmetic mean) between the two offsprings. In other words there is transvariation between two individuals, one from each distribution L and R, if (given $\mu_L > \mu_R$) the ordering between the corresponding categories on the response, has opposite sign of the one between the arithmetic means. The amount of this transvariation is measured by the difference $|i-j|$, i.e. the number of categories existing between the two individuals. According to (10) and (13) we thus obtain a three terms additive decomposition of the parent's node gmd:

$$\Delta(\mathbf{p}) = \Delta_{LL}\pi_L^2 + \Delta_{RR}\pi_R^2 + 2\,(\mu_L - \mu_R)\,\pi_L\pi_R + \frac{4T_{LR}}{n_L n_R}\pi_L\pi_R. \qquad (14)$$

In the case $\mu_L < \mu_R$ the term T_{LR} is evaluated as

$$\sum_j\sum_{i>j}(i-j)n_{iL}n_{jR}$$

and the gmd decomposition is given by

$$\Delta(\mathbf{p}) = \Delta_{LL}\pi_L^2 + \Delta_{RR}\pi_R^2 + 2\left(\mu_R - \mu_L\right)\pi_L\pi_R + \frac{4T_{LR}}{n_L n_R}\pi_L\pi_R. \tag{15}$$

By defining

$$T_{LR} = \begin{cases} \sum_j \sum_{i>j}(i-j)n_{jL}n_{iR}, & \text{if } \mu_L > \mu_R \\ \sum_j \sum_{i>j}(i-j)n_{iL}n_{jR}, & \text{if } \mu_L < \mu_R \end{cases}$$

the following unique expression for the gmd decomposition is thus obtained:

$$\Delta(\mathbf{p}) = \Delta_{LL}\pi_L^2 + \Delta_{RR}\pi_R^2 + 2\left|\mu_R - \mu_L\right|\pi_L\pi_R + \frac{4T_{LR}}{n_L n_R}\pi_L\pi_R. \tag{16}$$

Our proposal is then to select the split that minimizes both$\Delta_{LL}\pi_L^2 + \Delta_{RR}\pi_R^2$, accounting for the dispersion within the subnodes, and $\frac{4T_{LR}}{n_L n_R}\pi_L\pi_R$, penalizing the split for the amount of transvariation between the subnodes. This goal is achieved by adopting the following splitting criterion:

$$C_{\Delta,T}(s) = \pi_L\pi_R\left|\mu_R - \mu_L\right| \tag{17}$$

i.e. by selecting the split that maximizes the weighted absolute difference between the subnodes arithmetic means, with weight given by the so called *anti-end-cut factor* $\pi_L\pi_R$, that forces the criterion to prefer splits resulting in subnodes of similar size. Contrary to the Agresti splitting rule, criterion (17) enables to distinguish between ordinal exclusive splits because of the presence of the dispersion measures of the subnodes (even if the transvariation term vanishes). Moreover, the use of T_{LR}, measuring the amount rather than the number of transvariations, can overcomes the above drawback of giving equivalent evaluation to all transvarying pairs, regardless of the number of categories they transvary for.

5 Application: the Eurobarometer survey.

In order to prove the capability of the proposed method (17), we apply it to the Eurobarometer survey and we compare the results with those of the Gini ordinal method (2) and Agresti method (9). The data regards the Standard Eurobarometer 71.2 [2]. This survey was conducted between May and June 2009 in 31 states of the European community for a total of 29,768 respondents living in European countries and older than 15 years. As part of the analysis we consider as explanatory variables the answers relating to economic and social situation and some variables about socio-demographic respondents informations (that is sex, age,living area, country etc). The dependent variable is referred to the overall satisfaction of the current life of the respondent. This variable has 4 ordered categories of levels: "Very satisfied", "Fairly satisfied", "Not very satisfied" and "Not at all satisfied". The analysis shows very encouraging results as reported in Table 2.

Table 2: Comparation among the different method

Method	n. splits	n. variables	Misclassification rate
New Method	12	7	0.38435
Agresti	11	6	0.38531
Nominal Gini	16	10	0.38441

Based on the obtained results, the best method in terms of simplicity and forecasting ability is the tree based on the minimization of (17).

[2]The data can be downloaded on the website http://zacat.gesis.org/webview/index.jsp

References

Agresti, A. (1981). Measures of Nominal-ordinal Association. *Journal of the American Statistical Association*, 76(375):524–529.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Cerioli, A. (1988). Indici di Associazione di Agresti e Transvariazione. Quaderni dell'Istituto di Statistica 6, Istituto di Statistica - Universitá degli Studi di Parma.

Cerioli, A. (1990). Transvariation Measures in Nominal-Ordinal Association. In Momirovic, K., editor, *Compstat 1990: Proceedings in Computational Statistics 9th Symposium Held at Dubrovnik Yugoslavia, 1990*, pages 73–74. Springer Verlag.

Costa, M. (2008). Gini Index Decomposition for the Case of Two Subgroups. *Communications in Statistics - Simulation and Computation*, 37(4):631–644.

Dagum, C. (1980). Inequality Measures Between Income Distributions with Applications. *Econometrica*, 48(7):1791–1803.

Dagum, C. (1997). A New Approach to the Decomposition of the Gini Income Inequality Ratio. *Empirical Economics*, 22(4):515–531.

Gini, C. (1912). Variabilitá e Mutabilitá. *Studi Economici e Giuridici della Facoltá di Giurisprudenza. Cagliari*, 2(3). Reprinted in Gini (1955).

Gini, C. (1916). Il Concetto di Transvariazione e le Sue Prime Applicazioni. *Giornale degli Economisti e Rivista di Statistica*, (52):13–43. Reprinted in Gini (1959).

Gini, C. (1959). Il Concetto di Transvariazione e le Sue Prime Applicazioni. In Ottaviani, G., editor, *Memorie di Metodologia Statistica. Transvariazione*, volume II. Libreria Goliardica. Roma.

Leti, G. (1983). *Statistica Descrittiva*. Il Mulino.

Piccarreta, R. (2008). Classification Trees for Ordinal Variables. *Computational Statistics*, 23(3):407–427.

Xia, F., Zhang, W., and Wang, J. (2006). An Effective Tree-Based Algorithm for Ordinal Regression. *IEEE Intelligent Informatics Bulletin*, 7(1):22–26.