Quaderni di Statistica
Vol. 8, 2006

A tree-based method for selection of variables in models for ordinal data

Carmela Cappelli, Angela D'Elia
Dipartimento di Scienze Statistiche
Università di Napoli Federico II
E-mail: carmela.cappelli@unina.it; angela.delia@unina.it

Summary: In this paper we propose a strategy aimed at variable selection in models for ordinal data. The procedure core exploits the primary splitters, as they arise from a tree-based procedure. Indeed, the latter is based on a recursive partitioning approach that increasingly reduces the so called *impurity* of the response variable based on the covariate values. Since the reduction in impurity provides a natural importance ranking of the covariates, this allows to identify those most explanatory among the available ones. The proposed strategy is, then, applied to a case study concerning the consumers' preferences towards different types of smoked salmons.

Keywords: Tree-based methods, MUB model, Variable Selection, EuroSalmon survey.

1. Introduction

In studies on customers/users' satisfaction towards a given item, it is common to gather an ordered response variable, that expresses the consumer's liking on a hedonic scale, and a set of covariates (both categorical and quantitative) that characterize the raters and/or the item. In such situation, a relevant issue concerns the selection of the covariates that mostly explain different degrees of satisfaction among the raters.

The problem of searching for a parsimonious model involving a subset of the covariates has no default solution and since the searching through all possible sub-models is usually time consuming and, at an extreme, it

might be unfeasible, various selective strategies have been proposed in the literature, mainly in the context of the linear model (for discussion and comparison, see Friedman *et al.*, 2001, and for a review see Guyon and Elisseeff, 2003).

Here, we propose an empirical strategy (for a preliminary work, see Cappelli and D'Elia, 2006) that exploits tree-based methods, a non parametric tool that allows to model the relation between a response variable and a set of predictors by means of a recursive binary partitioning approach that increasingly reduces the so called *impurity* of the response variable based on the covariate values. In particular, we propose to consider the covariates that generate the so called *primary splitters*, i.e. the splitters that mostly reduce impurity, picking those at the root node that correspond to the whole data set.

Although the proposed strategy can be applied for any model, we focus on a mixture model for ordinal data recently proposed by D'Elia and Piccolo (2005). Indeed, in this framework the inclusion of covariate has been, so far, faced on a tentative basis, being the variable selection methods mainly developed in the context of the linear model. For an illustrative purpose, we have applied this strategy to analyze a data set concerning the preferences expressed by a sample of citizens (from some European Union countries) towards different smoked salmon: this choice is due to the great number of available covariates in the data set.

This paper is organized as follows: in section 2 the main features of the model are introduced; in section 3 the covariate selection strategy is illustrated. Evidence from the above mentioned case study is presented and discussed in section 4; brief final remarks conclude the paper.

2. Rationale and main features of the MUB model

Assume a sample of raters is asked to evaluate an item (product, service, etc.) by means of an ordinal score y ranging in $[1, m]$, for a given m. We also assume that the observed ordinal score (rating) y is the realization of a Mixture of a Uniform and a shifted Binomial distribution (D'Elia and Piccolo, 2005); the rationale for this class of models (MUB)

arises from the following considerations.

Firstly, the degree of liking for a given product, depending on several causes, may be thought to follow a Gaussian distribution, but the rating $y\,(=1,2,...,m)$ assigned to the product can be analyzed as a realization of a discrete random variable Y: a suitable probabilistic model for achieving the mapping of a continuous variable into a discrete set of values $y=1,2,...,m$, is the shifted Binomial distribution.

Secondly, it seems sensible to assume that the preference expression is also characterized by an uncertainty component, that adds up to the basic liking: this uncertainty/fuzziness component in the preferences expression can be adequately described by means of the discrete Uniform distribution.

Thus, for a given m we get $Y \sim MUB(\pi,\xi)$ with probability mass function:

$$Pr(Y=y)=\pi\left(\begin{array}{c} m-1 \\ y-1 \end{array}\right)(1-\xi)^{y-1}\xi^{m-y}+(1-\pi)\frac{1}{m},\quad y=1,2,\ldots,m;$$

where $\pi \in [0,1]$ and $\xi \in [0,1]$.

The π parameter is inversely related to the Uniform component of the MUB model; thus, the estimate of $(1-\pi)$ is a measure of the uncertainty in the expression of the ratings. On the other hand, the meaning of ξ depends on how the ratings have been codified: in general, being greater the score greater the liking, the estimate of $(1-\xi)$ is a preference measure.

By letting:

$$\pi=\frac{1}{1+exp(-\boldsymbol{z}_i\boldsymbol{\beta})};\quad \xi=\frac{1}{1+exp(-\boldsymbol{w}_i\boldsymbol{\gamma})},$$

where $\boldsymbol{z}_i=(1,z_{i1},...,z_{ip_\pi})'$ is the i-th row of a design matrix **Z** of $(p_\pi+1)$ explanatory variables related to the parameter π and $\boldsymbol{w}_i=(1,w_{i1},...,w_{ip_\xi})'$ is the i-th row of a design matrix **W** of $(p_\xi+1)$ explanatory variables related to ξ, we allow the inclusion of the i-th rater ($i=1,2,...,n$) covariates in the MUB model, preserving at the same time -by means of the logistic mapping- the admissible parameter values: $\pi \in [0,1]$ and $\xi \in [0,1]$.

The inclusion of covariates greatly improves the usefulness of the MUB model, since it makes it possible to study the impact of raters/item features on different liking patterns.

In particular, we can study the effect of the covariates $\boldsymbol{x}_i = \boldsymbol{z}_i;\ \boldsymbol{w}_i$ on the expected rating (that is, on the basic feeling):

$$E(Y \mid \boldsymbol{x}_i) = \frac{m-1}{2}\left[\frac{tgh(-\boldsymbol{w}_i\boldsymbol{\gamma}/2)}{(1+exp(-\boldsymbol{z}_i\boldsymbol{\beta}))}\right] + \frac{m+1}{2}\ ,$$

as well as on the heterogeneity of the ratings themselves (as in D'Elia and Piccolo, 2006):

$$\mathcal{A}(Y \mid \boldsymbol{x}_i) = \frac{1}{m-1}\left\{\left(\sum_{y=1}^{m}\left[Pr(Y=y \mid \boldsymbol{z}_i\ ,\ \boldsymbol{w}_i)\right]^2\right)^{-1} - 1\right\},$$

by means of the normalized heterogeneity index $\mathcal{A}$, originally proposed by Laakso and Taagepera (1979).

For the class of MUB models, as for any statistical model involving covariate effect, the selection of those relevant is an issue. At this aim, we propose a strategy that exploits the selection criterion in tree methods, as it will be explained in the next section.

3. Variable importance ranking via tree-based methods

Tree-based methods are a nonparametric tool for modelling the relationship between a response variable and a set of predictors. In the last decades they have shown to be a useful approach for high dimensional data analysis able to capture nonlinear structures and interactions effects leading to a blossoming of methodological proposals and practical applications (Friedman *et al*, 2001).

In the basic method (Morgan and Sonquist, 1963; Breiman *et al*, 1984), the covariate space is recursively partitioned into two subregions containing units that are as homogeneous as possible with respect to the response variable. The data partitioning is based on a splitting criterion which allows to select at each tree node the best "splitter", i.e. the best

covariate and cut point along it. Indeed, since the aim is making the distribution of the response variable as "pure" as possible within the nodes, splitters are evaluated by means of an *impurity measure* $i(t)$ that expresses how homogeneous a node t is with respect to the response variable.
Let $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ be a training set described in terms of a response variable and a set of p covariates and denote by s_{x_j} a splitter based on covariate x_j. It is possible to evaluate the decrease in impurity generated by splitting a node t into its left and right descendants (t_l and t_r, respectively) using splitter s_{x_j} as follows:

$$\Delta I(s_{x_j}, t) = i(t) - [i(t_l)p_l + i(t_r)p_r],$$

where p_l and p_r are the proportions of node t units falling into each descendant.

In the present case, as we are dealing with an ordinal response variable, impurity need to be evaluated by means of heterogeneity indexes. The most common measure employed in tree construction is the Gini index of heterogeneity but, in this context we suggest the Shannon's entropy index because it is more sensitive to changes in the distributions and more appropriate for factors with several levels (see Zhang and Singer, 1998).

Let $f_y(t),\ y = 1, 2, ..., m$ be the observed frequencies of the ordinal scores (ratings) at node t, the Shannon index is defined as

$$H(t) = -\sum_{k=1}^{m} f_k(t) \log(f_k(t)).$$

Due to their flexibility, tree-based methods have been employed for different purposes: in the present case we suggest to employ trees for variable selection. Indeed, the values ΔI, computed by using $H(t)$ as impurity measure, express how effective a splitter is in reducing the heterogeneity of the response variable, i.e., in explaining the response variable and thus they can be employed for variable importance ranking. In particular, a normalized measure of the covariate importance can be defined as:

$$\frac{\Delta I(s_{x_j}, t)}{\max_{x_j} \Delta I(s_{x_j}, t)}.$$

Then, our idea is to estimate nested MUB models in which the covariates are included according to the ranking provided by the above criterion at the root node that contains all the observations. This choice arises because we are interested in ranking the covariates with respect to the whole data set; besides, it is well known that, in practical applications, competing covariates at the root node explains most of the heterogeneity of the response variable (see Holte, 1993).

In various applications (see for example Cappelli and D'Elia, 2006) we found this strategy effective in the detection of relevant covariates, especially when their amount is not small; here, for illustrative purposes, we presents and discuss the results of a real data set analysis.

4. An application

In 2002 a survey was conducted in some European countries to assess consumers' preferences towards smoked salmons (EuroSalmon project): indeed, in the seafood market, smoked salmon has become a product of a wide range consumption in Europe during the last years and, for this reason, the interest in its quality and in how this is related to consumers' preferences is greatly increasing (Cardinal *et al*., 2004).

In the survey, 1063 consumers had to express their appreciation towards a given salmon product by a score on a 9 point scale (ranging from dislike extremely to like extremely); additional information on consumers features (gender, age, country of origin, occupation, marital status, etc.) and on the frequency of consumption was also collected (EuroSalmon Final Report, 2004).

For an illustrative purpose, here we focussed on product 27 (labelled PROD27) that shows the highest heterogeneity of scores, with a bimodal pattern as shown in Figure 1. In such a case it is interesting to investigate whether this variability in the consumers' preferences can be explained by means of the available covariates: in particular the detection of significant features of the raters may explain different degrees of preferences and, then, the observed bimodal distribution of the ratings.

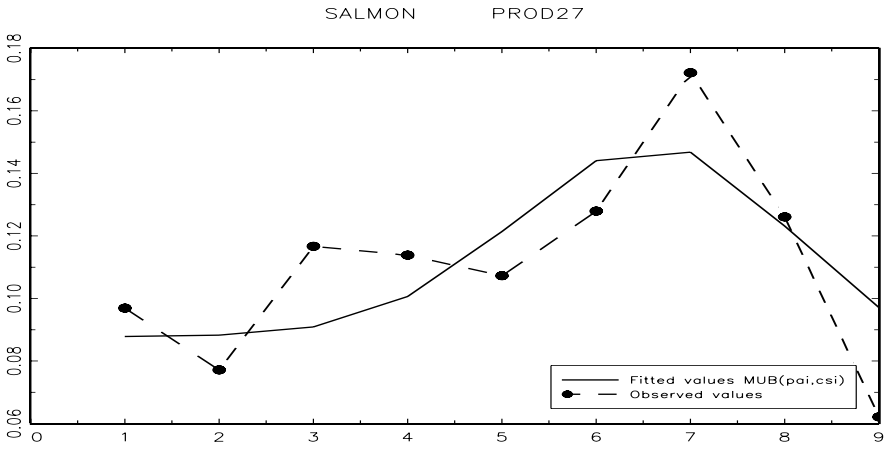

Figure 1. Estimated and observed ratings distribution for salmon product 27.

Following the procedure described in Section 3, we found that the variables that mostly explain the heterogeneity of the response variable are: *age*, *country of origin* and *marital status*.

Then, four MUB models were estimated by adding one covariate at a time according to the above ordering. As expected the sequential inclusion of covariates progressively increases the log-likelihood, also, the incresae has been found significant. tests

Table 1. Inclusion of covariates.

MUB models	covariates	log-likelihood
model 1	no covariates	-2314.49
model 2	age	-2310.84
model 3	age + country	-2290.04
model 4	age + country + marital status	-2289.18

For the purpose of model comparison we have applied the likelihood ratio test; in particular, we have found the difference in loglikelihood among model 2 and model 3 (with the inclusion of the 3-levels covariate *country*) is the most significant. Thus, we have investigated model 3 more deeply, getting the findings reported in Table 2.

As far as it concerns the uncertainty ($1-\hat{\pi}$), for each country of origin, older consumers exhibit lower heterogeneity in their judgements with respect to younger ones: it appears that the fuzziness in assigning the scores diminishes with the age.

On the other hand, French consumers have the expected lowest preference ($1-\hat{\xi}$) towards salmon product 27, whatever their age; on the contrary, Belgian and German consumers have high and very high estimated scores, respectively.

Table 2. Profiles of consumers.

Age	Country	$1-\hat{\pi}$ (uncertainty)	$1-\hat{\xi}$ (liking)	$E(Y)$	$\mathcal{A}(Y)$
18-34	Germany	0.87791	0.77643	5.26998	0.96096
18-34	Belgium	0.93955	0.71756	5.10522	0.99599
18-34	France	0.77738	0.41230	4.84381	0.95724
35-49	Germany	0.81162	0.73253	5.35044	0.96085
35-49	Belgium	0.90302	0.66706	5.12961	0.99073
35-49	France	0.67659	0.35619	4.62793	0.90805
50-65	Germany	0.76643	0.82019	5.59829	0.92210
50-65	Belgium	0.87643	0.76943	5.26636	0.98093
50-65	France	0.61441	0.47957	4.93699	0.88457

However, it is to be stressed that focussing only on the expectations $E(Y)$ may be misleading for such kind of data, as it is confirmed in Figure 2 where the estimated probability mass functions of the ratings are displayed for 6 chosen profiles of consumers: indeed, the distributions of German and Belgian consumers are much more shifted on higher ratings (that is higher preferences), whereas the $E(Y)$'s are not so high because they reflect also the weight of the uncertainty (e.g. heavier tails). This evidence supports the need for analyzing together the effect of the covariates on both location and heterogeneity of the ratings.

On the other hand, it should be noticed that the detection and the inclusion of relevant covariates, by means of the tree-based strategy, greatly improves the fitting of the MUB model to the observed ratings distribution

with respect to the model without covariates (as it can be noticed comparing Figure 2 versus Figure 1). In particular, the observed bimodal pattern is now explained by means of the covariate *country of origin*, opposing French consumers to Belgian and German ones.

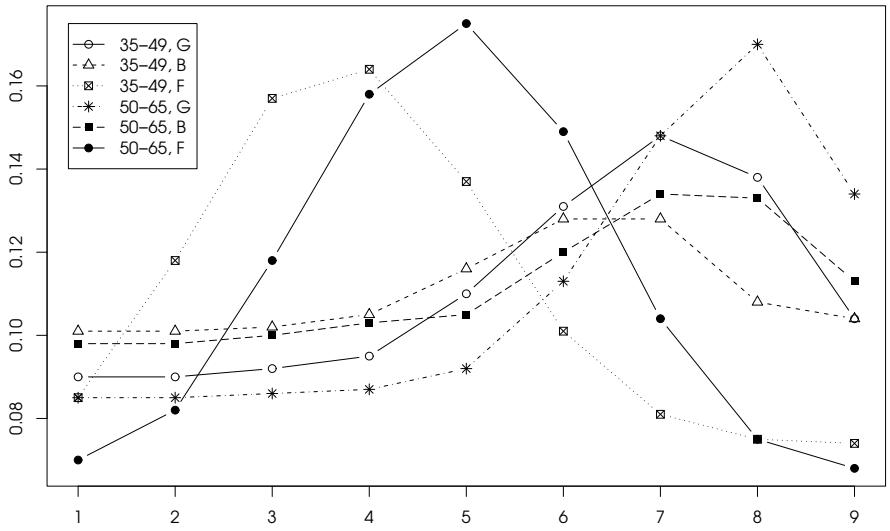

Figure 2. Estimated ratings distributions for different consumers' profiles.

5. Concluding remarks

In the paper we have presented an operative strategy for dealing with the issue of the covariate selection: a tree based method based on the Shannon' entropy index (as impurity measure) has been then exploited for the detection of significant covariates in a mixture model for ordinal response variables.

We have tested the proposed strategy on various data sets on customers/users' satisfaction towards a given item; here we have presented the results consumers' preferences towards smoked salmon: in this case, the available data set consist of a large amount of variables (the liking score, gender, age, country of origin, marital status, occupation, fre-

quency of consumption, number of family's member, etc.) and, by means of the tree-based method, we were able to quickly select those highly significant in explaining different degrees of preferences among the raters; in particular, the *country of origin* and the *age* were found relevant.

In other words, the application has shown that the proposed selection strategy improves the performance of the MUB model by identifying a candidate subset of covariates that may explain the different degree of satisfaction providing a better understanding of the underlying data-generating mechanism.

At this stage of the research, the proposed procedure has been used for simultaneously detect significant covariates with respect to both the components of the MUB model (uncertainty and basic liking, as expressed by $1-\pi$ and $1-\xi$, respectively). A possible further development would be to develop a tree-based method which uses different impurity measures for the two components, in order to adequately consider their different role and to detect separate (disjoint) sets of significant covariates for them. Future research will also address the comparison of the proposed strategy with classical selection methods such as forward stepwise selection and backward elimination.

Acknowledgements. The authors acknowledge financial support from the Dipartimento di Scienze Statistiche, Università di Napoli Federico II. Although this is a joint research, C. Cappelli wrote sections 1 and 3 and A. D'Elia wrote sections 2 and 5.

References

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

Cappelli C., D'Elia A. (2006), Preliminary identification of relevant covariates in models for ordinal data: a strategy based on tree methods, *ROBCLA-06 Book of Abstracts*, (Croux C., Riani M., Chiandotto B., Bini M. and Bertaccini B. Eds), Università degli Studi di Firenze, 23-24.

Cardinal M., Gunnlaugsdottir H., Bjoernevik M., Ouisse A., Vallet J.L. and Leroi F. (2004), Sensory characteristics of cold-smoked Atlantic salmon (*Salmo salar*) from European market and relationships with chemical, physical and microbiological measurements, *Food Research International*, 37, 181-193.

D'Elia A., Piccolo D. (2005), A mixture model for preferences data analysis, *Computational Statistics & Data Analysis*, 49, 917-934.

D'Elia A., Piccolo D. (2006), Uno studio sulla percezione delle emergenze metropolitane: un approccio modellistico, *Quaderni di Statistica*, 7, 121-161.

EuroSalmon (2004), Improved quality of smoked salmon for the European consumer, *Final Report for the EC "Quality of life and management of living resources" programme*, www.mmedia.is/matra/eurosalmon.

Friedman J., Tibshirani R., Hastie T. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York.

Guyon, I., Elisseeff, A. (2003), An introduction to variable and feature selection, *Journal of Machine Learning Research*, 3, 1157-1182.

Holte R.C. (1993), Very simple classification rules perform well on mostly commonly used datasets, *Machine Learning*, 11, 63-90.

Laakso M., Taagepera R. (1979), Effective number of Parties: a measure with application to West Europe, *Comparative Political Studies*, 12, 3-27.

Morgan J.N., Sonquist J.A. (1963), Problems in the analysis of survey data and a proposal, *Journal of American Statistical Association* 58, 415-434.

Zhang H., Singer B. (1998), *Recursive Partitioning in the Health Science*, Springer, New York.