Classification trees for preference data: a distance-based approach

Mariangela Sciandra[1], Antonella Plaia[1]

[1] Univerisitá degli Studi di Palermo, Italy

E-mail for correspondence: mariangela.sciandra@unipa.it

Abstract: In the framework of preference rankings, when the interest lies in explaining which predictors and which interactions among predictors are able to explain the observed preference structures, the possibility to derive consensus measures using a classification tree represents a novelty and an important tool given its easy interpretability. In this work we propose the use of a multivariate decision tree where a weighted Kemeny distance is used both to evaluate the distances between rankings and to define an impurity measure to be used in the recursive partitioning. The proposed approach allows also to weight differently high distances in rankings in the top and in the bottom alternatives.

Keywords: MRT; distance-based methods; preference data, Kemeny distance.

1 Introduction

In every day life ranking and classification are basic cognitive skills that people use in order to graduate everything that they experience. Moreover, many data collection methods in the social sciences often rely on ranking and classification. Grouping and ordering a set of elements is considered easy and communicative, so often it happens to observe rankings of sport-teams, universities, countries and so on. A particular case of ranking data is represented by preference data, in which individuals show their preferences over a set of alternatives, *items* from now on.

Preference rankings, for example consumers' preferences, are indicator of individual behaviours and, if subject-specific characteristics are available, besides a principal preference structure, interactions among predictors (of preference rankings) can be discerned. This allows to identify profiles of respondents giving same/similar rankings.

From a methodological point of view, preference analyses often model the probability for certain preference structures, finally providing the probabilities for choosing one single object. Such a problem has been widely explored in literature and many models have been proposed over the years, such as order statistics models (Dwass, 1957), distance-based models (Lee

and Yu, 2010) and some log-linear version of standard Bradley-Terry models (Dittrich, R. *et al.*,2002). Compared to parametric ranking models, the approach by Lee and Yu (2010), being based on the definition of a decision tree, is characterized by a simpler interpretability. Aim of this work is to extend the idea of Lee and Yu and to build a decision tree model where the response variable is represented by the subject specific preference rankings. The resulting decision tree will not be a classical Multivariate Regression Tree (MRT), because now each ranking vector should be considered as a unique multidimensional entity. Therefore, in this context, techniques known in literature to define splits for multivariate response variables are not feasible, because they should not take into account the ordinal structure of a ranking. Building a tree-based structure with rankings as response variable requires the definition of an impurity measure, for example a suitable distance which is sufficiently discriminatory. Starting from the well-known Kemeny distance, in this work we try to derive a decision tree through a recursive partitioning that uses, as impurity measure, the sum of these distances within node. In particular, we propose the use of a weighted version of the Kemeny distance (García-Lapresta, J.L., and Pérez-Román, D., 2010) to deal also with decision problems where it is important to emphasize differences between rankings that occur in particular items.
The paper is organized as follows. In the next section a brief introduction to MRT is presented together with the proposed approach based on the use of the weighted Kemeny distance for building a classification tree for preference data. Finally, an example on a real data set is presented.

2 Multivariate Regression Trees

Most of the literature on classification and regression tree deals with univariate response variables. Some recent attempts to develop regression tree methods handling multivariate responses are due to De'ath (2002), Larsen and Speckman (2004) and Lee and Lee (2005). In this section we assume the ordinary regression tree methodology as known, and give a brief overview of the class of the multivariate regression trees (MRT). A MRT can be seen as the natural extension of univariate classification and regression tree, where the multivariate response is predicted by some explanatory variables, both numeric and/or categorical. The main difference between the proposed extensions lies in the way in which they extend the definition of the partitioning metric . During the definition of a measure of distance between rankings many problems can arise. First of all, the measure should at least ensure that equal preference structures have zero distance; moreover, the distance shoud increase as the difference in these structures increases. Several measures have been proposed in literature in order to assess consensus: Bosch (2005) introduced the notion of consensus measures in the context of linear orders, while García-Lapresta, and Pérez-Román, (2008) extended Bosch's concept to the context of weak orders.

2.1 The proposed extension: a distance-based impurity measure

Weighted Kemeny distance has been proposed in order to face decision problem where it is not the same to have differences in the top alternatives than in the bottom ones. The use of the weighted Kemeny distance gives the possibility of introducing weights in order to distinguish where these differences occur. So, let V $= \{v_1, \ldots, v_m\}$ be a set of rankers with $m \geq 3$ and X $= \{x_1, \ldots, x_n\}$ a set of alternatives with $n \geq 3$. Let L(X) be the set of linear orders on X and R $\in L($X$)$. A profile vector is a vector of linear order such as

$$\mathbf{R} = (R_1, \ldots, R_m).$$

Given $R \in L(X)$ it is possible to define o_R as the position of each alternative in R, $o_R = (o_R(x_1), o_R(x_2), \ldots, o_R(x_n))$. In this way we can identify $L(X)$ with S_n (the set of permutations on the first n integers). Given $A \subseteq \mathbb{R}^n$ such that $S_n \subseteq A$ and a distance (metric) $d : A \times A \longrightarrow \mathbb{R}$, the distance of linear orders is is the mapping $\overline{d} : L(X) \times L(X) \longrightarrow \mathbb{R}$ defined by

$$\overline{d}(R_1, R_2) = d((o_{R_1}(x_1), \ldots, o_{R_1}(x_n)), (o_{R_2}(x_1), \ldots, o_{R_2}(x_n))) \quad \forall R_1, R_2 \in L(X)$$

The Kemeny metric on $L(X)$ is the mapping $d^K : L(X) \times L(X) \longrightarrow \mathbb{R}$ defined as the cardinality of the symmetric difference between the linear orders. So let $R_1 \equiv (a_1, \ldots, a_n) \in L(X)$ and $R_2 \equiv (b_1, \ldots, b_n) \in L(X)$

$$d^K(R_1, R_2) = \overline{d}_K(R_1, R_2) = d_K((a_1, \ldots, a_n), (b_1, \ldots, b_n)) =$$

$$\sum_{i,j=1, i<j}^{n} |sgn(a_i - a_j) - sgn(b_i - b_j)|$$

Let $\mathbf{w} = (w_1, \ldots, w_{n-1}) \in [0, 1]^{n-1}$ be a weighting vector such that $w_1 \geq \cdots \geq w_{n-1}$ and $\sum_{i=1}^{n-1} w_i = 1$. The weighted Kemeny distance on $L(X)$ associated with $\mathbf{w}$ is the mapping

$$\overline{d}_{K,w}(R_1, R_2) = \frac{1}{2} \left[\sum_{i,j=1, i<j}^{n} w_i |sgn(a_i^{\sigma_1} - a_j^{\sigma_1}) - sgn(b_i^{\sigma_1} - b_j^{\sigma_1})| \right.$$
$$\left. + \sum_{i,j=1, i<j}^{n} w_i |sgn(b_i^{\sigma_2} - b_j^{\sigma_2}) - sgn(a_i^{\sigma_2} - a_j^{\sigma_2})| \right]$$

where $(a_1, \ldots, a_n) \equiv R_1 \in L(X), (b_1, \ldots, b_n) \equiv R_2 \in L(X)$ and $\sigma_1, \sigma_2 \in S_n$ are such that $R_1^{\sigma_1} = R_2^{\sigma_2} \equiv (1, 2, \ldots, n)$.
The recursive partitioning process creates a nested sequence of subtrees $T_m = \{root - node\} \subset \cdots \subset T_0 = \{full - tree\}$ maximazing, at each step, the decrease in the node impurity

$$i(t) = \sum_{\mathtt{p > q}} \overline{d}_{K,w}(R_p, R_q),$$

according to all covariates and respective split points. Accordingly, the impurity measure for a generic subtree T having $\{N_t\}$ terminal nodes is defined as

$$I(T) = \sum_{\texttt{all terminal nodes} N_t} \overline{d}_{K,w}(R_p, R_q), \quad with \quad p > q.$$

As example, a dataset concerning the ranks assigned by $n = 91$ students to six different platforms for computer games has been considered. For each respondent the *age*, the number of *hours* spent on gaming per week and a dummy *own* indicating if the platform is currently owned have been used as explanatory variables. Using the proposed MRT the profiles of respondents giving the same rankings have been identified, even if results cannot be reported due to lack of space.

Acknowledgments: Special thanks to Giovanni Boscaino and Enza Capursi for the intersting discussions on the topic.

References

De'ath, G. (2002). Multivariate regression trees: A new technique for modeling species-environmental relationships. *Ecology*, **83**, 1105 – 1117.

Dittrich, R., Hatzinger, R. and Katzenbeisser, W. (2002). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society, Series C*, **47 (4)**, 511 – 525.

Dwass, M. (1957). On the distribution of ranks and of certain rank order statistics. *The Annals of Mathematical Statistics*, **28**, 424 – 431.

García-Lapresta, J.L., and Pérez-Román, D. (2008) Some Measures of Consensus Generated by Distances on Weak Orders. In: *Proceedings of the XIV Congreso espaol sobre tecnologías y lógica fuzzy*, 477 – 483

García-Lapresta, J.L., and Pérez-Román, D. (2010) Consensus Measures Generated by Weighted Kemeny Distances on Linear Orders. In: *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications (ISDA)*, 463 – 468

Larsen, D. R., Speckman, P. L. (2004). Multivariate Regression Trees for Analysis of Abundance Data. *Biometrics*, **60**, 543 – 549.

Lee, S.K. Lee, J.C. (2005). On generalized multivariate decision tree by using GEE. *Computational Statistics and Data Analysis*, **49**, 1105 – 1119.

Lee, P.H. Yu, P.L.H. (2010). Distance-based tree models for ranking data. *Computational Statistics and Data Analysis*, **54**, 1672 – 1682.