Weighted and unweighted distances based decision tree for ranking data.

Alberi decisionali per ranking data basati su distanze pesate e non pesate

Antonella Plaia, Simona Buscemi, Mariangela Sciandra

Abstract Preference data represent a particular type of ranking data (widely used in sports, web search, social sciences), where a group of people gives their preferences over a set of alternatives. Within this framework, distance-based decision trees represent a non-parametric tool for identifying the profiles of subjects giving a similar ranking. This paper aims at detecting, in the framework of (complete and incomplete) ranking data, the impact of the differently structured weighted distances for building decision trees. The traditional metrics between rankings don't take into account the importance of swapping elements similar among them (element weights) or elements belonging to the top (or to the bottom) of an ordering (position weights). By means of simulations, using weighted distances to build decision trees, we will compute the impact of different weighting structures both on splitting and on consensus ranking. The distances that will be used satisfy Kemenys axioms and, accordingly, a modified version of the rank correlation coefficient τ_x, proposed by Edmond and Mason, will be proposed and used for assessing the trees' goodness.

Abstract *I dati di preferenza rappresentano un particolare tipo di ranking data (ampiamente usati nello sport, nella ricerca sul web, nelle scienze sociali), dove un gruppo di persone dá le sue preferenze su un set di alternative. In questo contesto, gli alberi decisionali basati sulle distanze rappresentano uno strumento non parametrico per identificare i profili di soggetti che forniscono un ranking simile. Questo "articolo" mira ad indagare, nel contesto di ordinamenti completi e incompleti, quale sia l'impatto delle differenti distanze pesate sulla costruzione di alberi decisionali. Le tradizionali metriche tra ordinamenti non prendono in considerazione l'importanza di scambiare elementi simili tra di loro (pesi di item) o elementi che stanno in cima o in coda a una classifica (pesi di posizione). Usando le distanze pesate per la costruzione degli alberi, condurremo uno studio di sim-*

Antonella Plaia, Simona Buscemi, Mariangela Sciandra
Dipartimento di Scienze Economiche, Aziendali e Statistiche, University of Palermo, Viale delle Scienze, Edificio 19, 90128 Palermo, Italy, e-mail: `antonella.plaia(simona.buscemi, mariangelasciandra)@unipa.it`

ulazione per misurare l'impatto di differenti sistemi di peso sia sugli splitting sia sull'individuazione del consensus ranking. Le distanze che saranno usate rispettano gli assiomi di Kemeny.

Key words: weighted distances, ranking, Kemeny, consensus, trees

1 Introduction

Distances between rankings and the rank aggregation problem have received a growing consideration in the past few years. Ranking and classifying are two simplified cognitive processes usefull for people to handle many aspects in their life. When some subjects are asked to indicate their preferences over a set of alternatives (items), ranking data are called preference data. One great issue of interest in literature is: what can be done to identify, through subject-specific characteristics, the profiles of subjects having similar preferences? In order to answer to this question, different solutions have been proposed: distance-based tree models [15], distance-based multivariate trees for ranking [4], log-linearized Bradley-Terry models [9] and a semi-parametric approach for recursive partitioning of Bradley-Terry models for incorporating subject covariates [18]. Lee and Yu (2010) [15] investigated the development of distance-based models using decision tree with weighted distances, where weights are related to items. The traditional metrics between rankings don't take into account the importance of swapping elements similar among them (element weights) or elements belonging to the top (or to the bottom) of an ordering (position weights). Kumar and Vassilvitskii (2010) [14] provided an extended measure for Spearman's Footrule and Kendall's τ, embedding weights relevant to the elements or to their position in the ordering. The purpose of this paper is to investigate the effect of different weighting vectors on the tree. A particular attention is given to the weighted Kemeny distance and to the consensus ranking process for assigning a suitable label to the leaves of the tree. The stopping criterion for detecting the optimum tree is a properly modified Tau_x [10].

The rest of the paper is organized as follows: Section 2 introduces different metrics between rankings, their properties and their weighted extension; Section 3 introduces the weighted correlation coefficient; after a brief view on decision trees, in Section 4 we perform our analysis through a simulation study and, in the end, a short conclusion is presented (Section 5).

2 Distances between rankings

Ranking data arise when a group of n individuals (experts, voters, raters etc) shows their preferences on a finite set of items (k different alternatives of objects, like movies, activities and so on). If the k items are ranked in k distinguishable ranks, a complete ranking or linear ordering is achieved [8]. A ranking π is, in this case, one of the $k!$ possible permutations of k elements, containing the preferences given by the judge to the k items. When some items receive the same preference, then a tied ranking or a weak ordering is obtained. In real situations, many times it happens that not all items are ranked: partial rankings, when judges are asked to rank only a subset of the whole set of items, and incomplete rankings, when judges can freely choose to rank only some items. In order to get homogeneous groups of subjects having similar preferences, it's natural to measure the spread between rankings through dissimilarity or distance measures among them. Within the metrics proposed in literature to compute distances between rankings, the Kemeny distance will be here considered [13]. The Kemeny distance (K) between two rankings π and σ is a city-block distance defined as:

$$K(\pi,\sigma) = \frac{1}{2}\sum_{r=1}^{k}\sum_{s=1}^{k}|a_{rs} - b_{rs}| \tag{1}$$

where a_{rs} and b_{rs} are the generic elements of the $k \times k$ score matrices associated to π and σ respectively, assuming value equal to 1 if the item r is preferred to or tied with the item s, -1 if the item s is preferred to the item r and 0 if $r = s$.
K is in a one-to-one correspondence, $\tau = 1 - 2d/D_{max}$, to the rank correlation coefficient τ_x proposed by [10] defined as:

$$\tau_x(\pi,\sigma) = \frac{\sum_{r=1}^{k}\sum_{s=1}^{k} a_{rs}b_{rs}}{k(k-1)}. \tag{2}$$

2.1 Weighted distances

Kumar and Vassilvitskii (2010) [14] introduced two aspects essential for many applications involving distances between rankings: positional weights and element weights. In short, i) the importance given to swapping elements near the head of a ranking could be higher than the same attributed to elements belonging to the tail of the list or ii) swapping elements similar between themselves should be less penalized than swapping elements which aren't similar. In this paper, we deal with case i) and consider the weighted version of the Kemeny metric. For measuring the weighted distances, the non-increasing weights vector $w = (w_1, w_2, ..., w_{k-1})$ constrained to $\sum_{p=1}^{k-1} w_p = 1$ is used, where w_p is the weight given to position p in he ranking.

Given two generic rankings of k elements, π and σ, the Weighted Kemeny distance was provided by [11] as follows:

$$K^{w}(\sigma,\pi) = \frac{1}{2}\left[\sum_{\substack{r,s=1\\ r<s}}^{k} w_r|a_{rs}^{(\sigma)} - b_{rs}^{(\sigma)}| + \sum_{\substack{r,s=1\\ r<s}}^{k} w_r|b_{rs}^{(\pi)} - a_{rs}^{(\pi)}|\right], \tag{3}$$

where (σ) states to follow the σ ranking and (π), similarly, orders according to π (see [16] for more details).

3 A new suitable rank correlation coefficient

In this paper we propose a new consensus measure, suitable for position weighted rankings. It represents an extension of τ_x proposed by [10] that handles linear and weak rankings when the position occupied by the items is relevant. It is defined as:

$$\tau_x^w(\pi,\sigma) = \frac{\sum_{r<s}^{k} a_{rs}^{\sigma} b_{rs}^{\pi} w_r + \sum_{r<s}^{k} a_{rs}^{\pi} b_{rs}^{\sigma} w_r}{Max[K^w(\sigma,\pi)]}, \tag{4}$$

where the denominator represents the maximum value for the Kemeny weighted distances, equal to:

$$Max[K^w(\pi,\sigma)] = \sum_{r=1}^{m-1} (m-1) w_r \cdot n. \tag{5}$$

A consensus measure has to satisfy conditions like unanimity, anonymity and neutrality, i.e. the consensus in every subset of individuals is maximum if and only if all opinions are the same and the degree of consensus is not affected by permutations of the voters or permutations of the alternatives, respectively. Furthermore, it could fulfill some other properties such as maximum dissension, reciprocity and homogeneity, i.e.: in each subset of two subjects, the minimum consensus is achieved if the preferences are linear orderings and each one is the reverse of the other one; if all individual orderings are reversed then level of consensus doesn't change and, in the end, if a subset of agents is replicated, then the consensus in that group doesn't change [11]. By simulations, we verified the fulfillment of these properties.

4 Decision Trees and Simulation Study

Decision trees are non parametric recursive statistical tools used for classification and prediction issues. The most known decision tree methodology is applied when the response variable is categorical or quantitative. Recently the procedure has been extended to rankings as response variable. For more details see [16]. In this paper we will use the weighted Kemeny distance (3) as impurity function and τ_x^w (4) as a measure of goodness of the tree. In particular, we are interested in evaluating the effect both on the splits and on the leaf labels of different weighting vectors w. For this reason, following [7], we consider a theoretical population partition of the predictor space (X_1 and X_2): Fig. 1 shows one of the nine datasets considered in the simulation plan, with $X_1 \sim U(0,10)$ and $X_2 \sim U(0,6)$. The number of rankings falling in each group was defined by a random number drawn from a normal distribution $N(10,2)$ and each number was divided by the summation of all of them, obtaining a relative frequency distribution for each sub-partition.

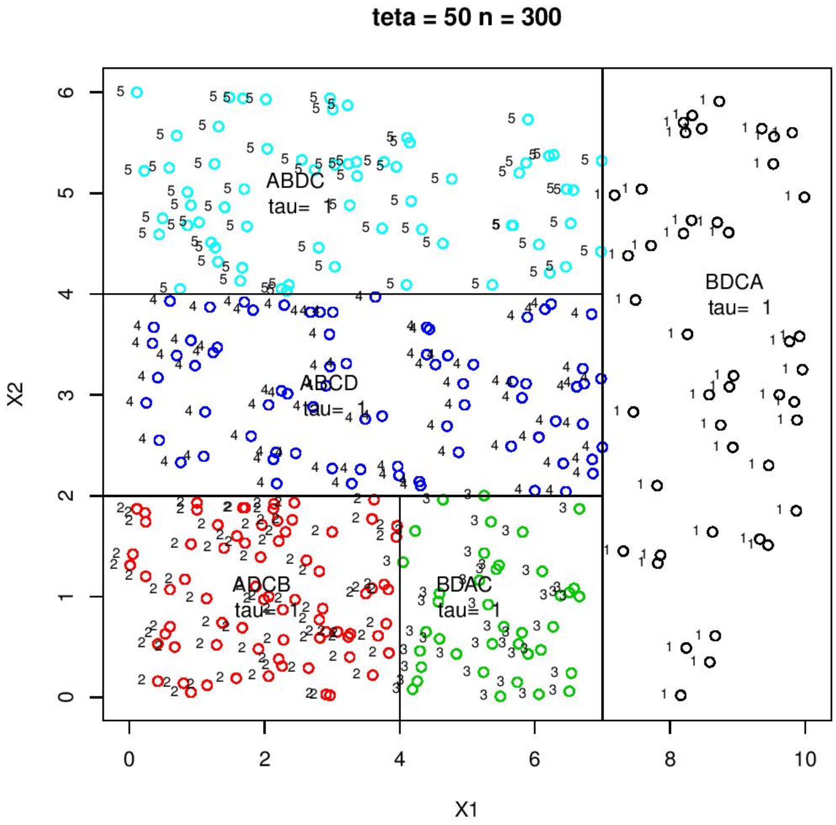
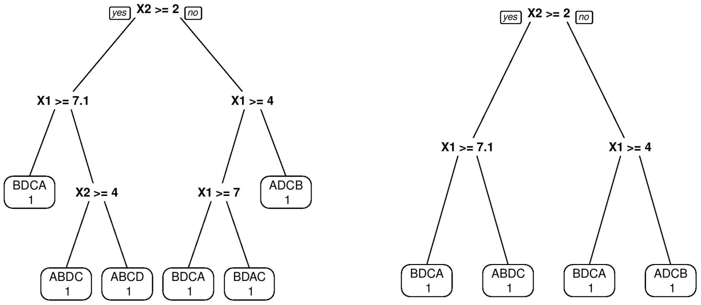


Fig. 1: Generation of homogeneous groups of ranking

The rankings of $k = 4$ items of each sub-partition were generated from a Mallows Model [12], varying the dispersion parameter θ, according to three different level of noise (low with $\theta = 50$, medium with $\theta = 2$ and high with $\theta = 1$). Considering three levels for the sample size (50, 100 and 300), the experimental design counts 3x3=9 different experiments. For each dataset, five different weighting vectors are considered : $w_1 = (1/3,1/3,1/3)$, $w_2 = (3/6,2/6,1/6)$, $w_3 = (1/2,1/2,0)$, $w_4 = (2/3,1/3,0)$ and $w_5 = (1,0,0)$.

With reference to the data in Fig. 1 (corresponding to $\theta = 50$ and $n = 300$), Fig. 2 reports two of the five trees obtained: in particular, Fig. 2a shows the tree corresponding to w_1, which perfectly recreates the original partition of the predictor space; Fig. 2b corresponds to w_3 and, as expected, does not perform the two splits $X \gtreqless 4$ and $X \gtreqless 7$ (the couples of rankings below each of the split in fig. 2a do not differ for the first two positions).



(a) Decision tree with w_1 weights (b) Decision tree with w_3 weights

Fig. 2: Decision tree models for weighted rankings

5 Conclusion

In this paper, we have focused on distance-based decision trees for ranking data, when the position occupied by items is relevant. We have proposed the weighted Kemeny distance as impurity function and a relative proper weighted consensus measure to be computed in each leaf and for the whole tree. Our methodology found to be capable of identifying correctly homogeneous groups of rankings for the relevant positions (according to the weighting structure). Future developments could be an analytical study of the properties of the new consensus measure and a replication of the same analyses both with an increasing number of items and in the case of weak orderings.

References

1. Amodio, S. and D'Ambrosio, A. and Siciliano, R.: Accurate algorithms for identifying the median ranking when dealing with weak and partial rankings under the Kemeny axiomatic approach. European Journal of Operational Research, 249(2), 667-676 (2016)

2. Breiman, L. and Friedman, J. and Olshen, R. and Stone, C.: Classification and Regression Trees. Wadsworth and Brooks, 1984.
3. Cheng, W and Hühn, J. and Hüllermeier, E.: Decision Tree and Instance-Based Learning for Label Ranking. In:Léon Bottou and Michael Littman, Proceedings of the 26th International Conference on Machine Learning, pages 161-168, Montreal. Omnipress. (2009)
4. D'Ambrosio, A.: Tree based methods for data editing and preference rankings. Ph.D. thesis, Universitá degli Studi di Napoli "Federico II" (2007)
5. D'Ambrosio, A. and Amodio, S.: ConsRank: Compute the Median Ranking(s) According to the Kemeny's Axiomatic Approach. R package version 1.0.2. (2015)
6. D'Ambrosio, A. and Amodio, S. and Iorio, C.: Two algorithms for finding optimal solutions of the Kemeny rank aggregation problem for full rankings. Electronic Journal of Applied Statistical Analysis, 8(2). (2015)
7. DAmbrosio, A., and Heiser, W. J.: A recursive partitioning method for the prediction of preference rankings based upon kemeny distances. Psychometrika 81.3 774-794. (2016)
8. Cook, W.D.: Distance based and ad hoc consensus models in ordinal preference ranking. European Journal Operation Research 172:369385. (2006)
9. Dittrich, R., Hatzinger, R., Katzenbeisser, W.: Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. Journal of the Royal Statistical Society C (Appl Stat) 47(4):511525 (1998)
10. Edmond, E. J. and Mason, D. W.: A new rank correlation coefficient with application to the concensus ranking problem. Journal of Multi-criteria decision analysis, 11, 17-28. (2002)
11. García-Lapresta, J. L. and Pérez-Román, D.: Consensus measures generated by weighted Kemeny distances on weak orders. In: Procceedings of the 10th International Conference on Intelligent Systems Design and Applications, Cairo. (2010)
12. Irurozki, E., Calvo, B., Lozano, J. A.: PerMallows: An R Package for Mallows and Generalized Mallows Models. Journal of Statistical Software, 71(12), 1-30. doi:10.18637/jss.v071.i12 (2016)
13. Kemeny, J. G. and Snell, J. L.: Preference rankings an axiomatic approach. MIT Press. (1962)
14. Kumar, R. and Vassilvitskii, S.:Generalized Distances Between Rankings. In Proceedings of the 19th International Conference on World Wide Web, WWW '10, pages 571-580, New York, NY, USA. ACM (2010)
15. Lee, P. H. and Yu, P. LH.: Distance-based tree models for ranking data. Computational Statistics & Data Analysis, 54(6), 1672-1682. (2010)
16. Plaia, A. and Sciandra, M.: Weighted distance-based trees for ranking data. Advances in Data Analysis and Classification, pages 1-18. Springer, https://doi.org/10.1007/s11634-017-0306-x (2017)
17. Sciandra, M. Plaia, A. and Capursi, V.: Classification trees for multivariate ordinal response: an application to Student Evaluation Teaching. Quality & Quantity, pages 1-15. (2016)
18. Strobl, C., Wickelmaier, F., Zeileis, A.: Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. Journal of Educational and Behavioral Statistics 36(2):135153. (2011)
19. Therneau, T. and Clinic, M.: User written splitting functions for RPART (2015)
20. Therneau, T. and Atkinson, B. and Ripley, B: rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10 (2015)
21. Yu, P. LH. and Wan, W. M. and Lee, P. H.: Decision tree modeling for ranking data. In Preference Learning, pages 83-106. Springer. (2010)