Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Distance-based tree models for ranking data

Paul H. Lee [*], Philip L.H. Yu

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 6 November 2008
Received in revised form 1 January 2010
Accepted 21 January 2010
Available online 1 February 2010

Keywords:
Decision tree
Ranking data
Distance-based model

ABSTRACT

Ranking data has applications in different fields of studies, like marketing, psychology and politics. Over the years, many models for ranking data have been developed. Among them, distance-based ranking models, which originate from the classical rank correlations, postulate that the probability of observing a ranking of items depends on the distance between the observed ranking and a modal ranking. The closer to the modal ranking, the higher the ranking probability is. However, such a model basically assumes a homogeneous population and does not incorporate the presence of covariates.

To overcome these limitations, we combine the strength of a tree model and the existing distance-based models to build a model that can handle more complexity and improve prediction accuracy. We will introduce a recursive partitioning algorithm for building a tree model with a distance-based ranking model fitted at each leaf. We will also consider new weighted distance measures which allow different weights for different ranks in formulating more flexible distance-based tree models. Finally, we will apply the proposed methodology to analyze a ranking dataset of Inglehart's items collected in the 1999 European Values Studies.

1. Introduction

Ranking data frequently occurs where judges (or individuals) are asked to rank a set of items, which may be types of soft drinks, political goals, candidates in an election, etc. By studying ranking data, we can understand judges' perception and preferences on the ranked alternatives.

Over the years, various statistical models for ranking data have been developed such as order statistics models, rankings induced by paired comparisons (for instance, Bradley–Terry Model), distance-based models and multistage models; see Critchlow et al. (1991) and Marden (1995) for more details of these models. Among these models, distance-based models have the advantages of being simple and elegant. However, they have received much less attention than what they should deserve, probably because they have two major weaknesses: the assumption of homogeneous population and difficulties in incorporating covariates. These have greatly limited the usefulness of the model.

Distance-based models (Fligner and Verducci, 1986) assume a modal ranking π_0 and the probability of observing a ranking π is inversely proportional to its distance from the modal ranking. The closer to the modal ranking π_0, the more frequent the ranking π is observed. Many distance measures have been proposed in the literature. Typical examples of distances are Kendall, Spearman and Cayley's distances; see Mallows (1957), Critchlow (1985) and Diaconis (1988). The models with Kendall's distance is sometimes referred as Mallows' ϕ-model (Mallows, 1957). The models consist of only two parameters, modal ranking π_0 and dispersion λ, and yet can provide a useful descriptive summary to a set of ranking data. Simplicity is obvious.

[*] Tel.: +852 96846294.
E-mail addresses: honglee@hku.hk (P.H. Lee), plhyu@hku.hk (P.L.H. Yu).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.01.027

The distance-based models assume a homogeneous population, all individuals will have a consensus view on the ranking of the items which is summarized by the modal ranking π_0. However, this may not be always true. Recently, Murphy and Martin (2003) extended the use of mixtures to distance-based models to describe the heterogeneity among the judges. By relaxing the homogeneous assumption, this leads to a significant improvement in the model, but the model still does not incorporate the presence of covariates.

There are quite a number of developments for including covariates. For example Beggs et al. (1981), Chapman and Staelin (1982), Hausman and Ruud (1987) and Train (2003) discussed the rank-ordered logit model (originated from order statistic model) which can incorporate covariates. Yu (2000) developed the multivariate order statistic models for ranking data, which can incorporate covariates as well. Gormley and Murphy (2008) adopted the mixture of experts model introduced by Jacobs et al. (1991) which allows covariates for mixture model, but again with an order statistic model. For distance-based models, the inability to incorporate covariates still remains a major inadequacy. This paper aim at developing a distance-based model by incorporating covariates, and hence addressing the heterogeneity population, by making use of a decision tree approach.

Decision trees are statistical models designed for classification and prediction problems. A decision tree is so called because the prediction rules generated from a set of the covariates can be displayed in a tree-like structure. Because of their ease of model interpretation, and the automatic detection of important covariates and interaction effects, tree-based models have been developed successfully in extending classical statistical models including logistic regression tree (Chan and Loh, 2004), Poisson regression tree (Chaudhuri et al., 1995), log-normal regression tree (Ahn, 1996) and generalized autoregressive conditional heteroscedastic (GARCH) tree (Audrino and Bühlmann, 2001). In this paper, we will combine a decision tree and a distance-based model so as to develop a more flexible distance-based model which can allow the presence of covariates.

The remainder of this paper is organized as follows. Section 2 reviews the distance-based models for ranking data and proposes the new weighted distance-based models. Section 3 proposes an algorithm of building distance-based tree models for ranking data. To illustrate the feasibility of the proposed algorithm, a simulation study and a case study of real data are presented in Sections 4 and 5 respectively. Finally, some concluding remarks are given in Section 6.

2. Distance-based models for ranking data

2.1. Distance-based models

Some notations are defined here for better description of ranking data. When ranking k items, labeled $1, \ldots, k$, a ranking $\boldsymbol{\pi}$ is a mapping function from $1, \ldots, k$ to $1, \ldots, k$, where $\pi(i)$ is the rank given to item i. For example, $\pi(2) = 3$ means that item 2 is ranked third.

Distance function is useful in measuring the discrepancy in two rankings. The usual properties of a distance function are:

- $d(\boldsymbol{\pi}, \boldsymbol{\pi}) = 0$,
- $d(\boldsymbol{\pi}, \boldsymbol{\sigma}) > 0$ if $\boldsymbol{\pi} \neq \boldsymbol{\sigma}$,
- $d(\boldsymbol{\pi}, \boldsymbol{\sigma}) = d(\boldsymbol{\sigma}, \boldsymbol{\pi})$.

For ranking data, we require the distance, apart from the usual properties, to be right invariant, i.e. $d(\boldsymbol{\pi}, \boldsymbol{\sigma}) = d(\boldsymbol{\pi} \circ \boldsymbol{\tau}, \boldsymbol{\sigma} \circ \boldsymbol{\tau})$, where $\boldsymbol{\pi} \circ \boldsymbol{\tau}(i) = \boldsymbol{\pi}(\boldsymbol{\tau}(i))$. This requirement makes sure relabeling of items has no effect on the distance.

Some popular distances are Spearman's rho, given by

$$R(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \left(\sum_{i=1}^{k} [\pi(i) - \sigma(i)]^2 \right)^{0.5}. \tag{1}$$

Spearman's rho square, given by

$$R^2(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \sum_{i=1}^{k} [\pi(i) - \sigma(i)]^2. \tag{2}$$

Spearman's footrule, given by

$$F(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \sum_{i=1}^{k} |\pi(i) - \sigma(i)|, \tag{3}$$

and Kendall's tau, given by

$$T(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \sum_{i<j} I\{[\pi(i) - \pi(j)][\sigma(i) - \sigma(j)] < 0\}, \tag{4}$$

where $I\{\}$ is the indicator function. Apart from these distances, there are other distances for ranking data, and readers can refer to Critchlow et al. (1991) for details.