Clustering of preference rankings: a non-parametric soft-clustering approach

Clustering di dati di preferenza: un approccio di soft clustering non parametrico

Antonio D'Ambrosio and Willem J. Heiser

Abstract Typically, rank data consist of a set of individuals, or judges, who have ordered a set of items or objects according to their overall preference or some prespecified criterion. When each judge has expressed his or her preferences according to his own best judgment, such data are characterized by systematic individual differences. In the literature several approaches have been proposed in order to decompose heterogeneous populations into a defined number of homogeneous groups. Often, these approaches work by assuming that the ranking process is governed by some distance-based models.
We use the flexible class of methods proposed by Ben-Israel and Iyigun, which consists in a probabilistic-distance clustering approach, and define the disparity between a ranking and the center of a cluster as the Kemeny distance. This class of methods allows for probabilistic allocation of cases to classes, being a form of fuzzy clustering, rather than hard clustering, where the probability is unequivocally related to the chosen distance measure.
Abstract *In genere, i 'rank data' consistono in una serie di individui, o giudici, che hanno espresso le loro preferenze su un set di oggetti, o item, ordinando questi ultimi sulla base delle loro preferenze dal piú preferito al meno preferito. In letteratura sono stati proposti diversi approcci volti a decomporre popolazioni di giudici nel complesso eterogenee in termini di preferenze espresse in un numero ristretto di sotto-popolazioni internamente omogenee. Molto spesso tali approcci seguono approcci basati su misture di modelli basati su distanze, che prevedono sia, in taluni casi, la scelta della distanza piú adeguata, sia la stima di massima verosimglianza di una serie di parametri. In questo lavoro si propone un approccio di 'soft clustering' che si basa sul concetto di probabilistic distance clustering, utilizzando la distanza di Kemeny come metrica di riferimento.*

Antonio D'Ambrosio
University of Naples Federico II, Italy, e-mail: antdambr@unina.it

Willem J. Heiser
Leiden University, The Netherlands e-mail: heiser@fsw.leidenuniv.nl

Key words: Preference rankings, Soft clustering, Kemeny distance

1 Introduction

Preference rankings are data expressing individual's preferences over a set of available alternatives. Statistical methods and models for the analysis of preference rankings can be distinguished in methods based on badness-of-fit functions, which aim to describe the structure of rank data, and methods based on probabilistic models, which aim to model either the ranking process or the population of judges (Marden, 1995). Within this latter category heterogeneity among the judges is assumed, and the goal is generally the identification of homogeneous sub-populations.
When, in addition to rank data, also some information about the judges are known, a variety of models and methods have been introduced, such as generalized linear-like models for rank data (Ditrich, Hatzinger and Katzenbeisser, 1998; Ditrich, Katzenbeisser and Hatzinger, 2000; Böckenholt, 2001; Gormley and Murphy 2008a) and recursive partitioning methods (D'Ambrosio, 2008; Lee and Yu, 2010; Strobl, Wickelmaier and Zeileis, 2011; D'Ambrosio and Heiser, 2016; Plaia and Sciandra, 2017).
Among clustering methods for rank data, mixtures of Bradley-Terry-Luce models and mixtures of distance-based models were proposed (Croon, 1989; Murphy and Martin, 2003; Gormley and Murphy, 2008b, Jacques and Biernacki, 2014).
We focus our attention to clustering of preference rankings. Most of the existing clustering methods for rank data are based on (mixtures of) distance-based models. Other approaches work under (Bayesian) Placket-Luce models (Mollica and Tardella, 2017) or under some recently introduced models for rank data, such as the Insertion Sorting Rank model (Jacques and Biernacki, 2014). In all these cases, either the choice of the distance measure or the assumption of the right model play a key role in obtaining the solution.
We propose a clustering approach for rank data that belong to the probabilistic distance clustering class of methods defined by Ben-Israel and Iyigun (2008), that allows a flexible way to find homogeneous sub-groups without any assumption on the model that in the population generates the preferences. In following this approach, according to which the probability of each judge to belong to a given cluster is unequivocally related to the distance between himself and the cluster center, we use the Kemeny distance (Kemeny and Snell, 1962). This choice is due to the consideration that the Kemeny distance is defined in the space of both full and tied rankings, and that it is the unique distance measure defined on the extended permutation polytope, which is generally accepted as the geometrical space of preference rankings (Heiser and D'Ambrosio, 2013; D'Ambrosio et al., 2017).

References

1. Ben-Israel, A., and Iyigun, C. Probabilistic distance clustering. Journal of Classification, vol. 25(1), pp. 5–26. (2008)
2. Böckenholt, U. Mixed-effects analyses of rank-ordered data. Psychometrika, 66(1), 45–62. (2001)
3. D'Ambrosio, A. Tree-based methods for data editing and preference rankings. Doctoral dissertation. Naples, Italy: Department of Mathematics and Statistics. http://www.fedoa.unina.it/2746/
4. D'Ambrosio, 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)
5. D'Ambrosio, A., Mazzeo, G., Iorio, C., and Siciliano, R. A differential evolution algorithm for finding the median ranking under the Kemeny axiomatic approach. Computers & Operations Research, 82, 126–138. (2017)
6. Dittrich, R., Hatzinger, R., and Katzenbeisser, W. Modelling the effect of subjectspecific covariates in paired comparison studies with an application to university rankings. Journal of the Royal Statistical Society: Series C (Applied Statistics), 47(4), 511–525. (1998)
7. Dittrich, R., Katzenbeisser, W., and Reisinger, H. The analysis of rank ordered preference data based on Bradley-Terry Type Models Die Analyse von Prferenzdaten mit Hilfe von loglinearen Bradley-Terry Modellen. OR-Spektrum, 22(1), 117–134. (2000)
8. Gormley, I. C., and Murphy, T. B. A mixture of experts model for rank data with applications in election studies. The Annals of Applied Statistics, 4(2), 1452–1477. (2008a)
9. Gormley, I. C., and Murphy, T. B. Exploring voting blocs within the Irish electorate: A mixture modeling approach. Journal of the American Statistical Association, 103(483), 1014–1027. (2008b)
10. Heiser, W.J., and D'Ambrosio, A. Clustering and prediction of rankings within a Kemeny distance framework, in Lausen, B, Van den Poel, D. and Ultsch, A. (EDS.), Algorithms from and for Nature and Life, Springer series in Studies in Classification, Data Analysis, and Knowledge Organization, 19–31, Springer International Publishing Switzerland. (2013)
11. Kemeny, J. G. and Snell, L. Mathematical Models in the Social Sciences. Ginn and Company. (1962)
12. Lee, P. H., and Philip, L. H. Distance-based tree models for ranking data. Computational Statistics & Data Analysis, 54(6), 1672–1682. (2010)
13. Lee, P.H., and Yu, P.L.H. Mixtures of weighted distance-based models for ranking data with applications in political studies. *Computational statistics & data analysis*, vol. 56(8), pp. 2486–2500. (2012)
14. Marden, J.I. Analyzing and modelling rank data. Chapman & Hall, London. (1995)
15. Mollica, C., and Tardella, L. Bayesian PlackettLuce mixture models for partially ranked data. Psychometrika, 82(2), 442–458. (2017)
16. Murphy, T.B., and Martin,D. Mixtures of distance-based models for ranking data. *Computational statistics & data analysis*, vol. 41(3-4), pp. 645–655. (2003)
17. Jacques, J., and Biernacki, C. Model-based clustering for multivariate partial ranking data. *Journal of Statistical Planning and Inference*, vol. 149, pp. 201–217. (2014)
18. Plaia, A., and Sciandra, M. (2017). Weighted distance-based trees for ranking data. Advances in Data Analysis and Classification, https://doi.org/10.1007/s11634-017-0306-x
19. Strobl, C., Wickelmaier, F., and Zeileis, A. Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. Journal of Educational and Behavioral Statistics, 36(2), 135–153. (2011)