Applied Mathematics and Computation 290 (2016) 208–223

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Using extension sets to aggregate partial rankings in a flexible setting

Juan A. Aledo[a,*], Jose A. Gámez[b], David Molina[a]

[a] Departamento de Matemáticas, Universidad de Castilla-La Mancha, Albacete 02071, Spain
[b] Departamento de Sistemas Informáticos, Universidad de Castilla-La Mancha, Albacete 02071, Spain

ARTICLE INFO

Keywords:
Rank aggregation problem
Kemeny ranking problem
Borda method
Extension set
Partial ranking
Ranking with ties

ABSTRACT

This paper deals with the rank aggregation problem in a general setting; in particular, we approach the problem for any kind of ranking: complete or incomplete and with or without ties. The underlying idea behind our approach is to take into account the so-called extension set of a ranking, that is, the set of permutations that are *compatible* with the given ranking. In this way we aim to manage the uncertainty inherent to this problem, when not all the items are ranked and/or when some items are equally preferred. We develop two approaches: a general one based on this idea, and a constrained version in which the extension set is limited by not allowing non-ranked items to be placed in an existent bucket of tied items. We test our proposal by coupling it with two different algorithms. We formalize our approaches mathematically and also carry out an extensive experimental evaluation by using 22 datasets. The results show that the use of our extension sets-based approaches to compute the precedence matrix, clearly outperforms the standard way of computing the preference matrix by only using the information explicitly provided by the rankings in the sample.

1. Introduction

Dealing with rankings is currently a hot topic in statistics and machine learning research, due to the increasing availability of problems whose basic data are rankings (e.g. recommender systems, combinatorial optimization, preferences, etc.).

Rankings are a natural way to express preferences. Specifically, given a set of items $[[n]] = \{1, 2, \ldots, n\}$, a ranking π is an order of preference over (some of) these items. The case that has received more attention in the literature is the one in which all the items are ranked, that is, rankings are permutations of n elements [22,27]. However, real world problems usually deal with *incomplete* rankings, i.e. only p items are ranked, $2 \leq p < n$. This is the case of users expressing preferences about a set of movies, books, etc. when they have no opinion about some items. In these cases it is also usual to use a preference system, so obtaining (incomplete) rankings with ties (see Section 2 for details). It is important not to confuse our object of study (i.e. rankings expressing preferences about a set of elements) with *ratings*, when some (maybe all of them) elements of a set are graded according a given scoring system.

Our aim is not to use the preferences given by a set of users to make recommendations, but to obtain a ranking (permutation) which summarizes the information given by them. In other words, we want to get a *consensus* permutation which

* Corresponding author. Tel.: +34 967599200; fax: +34 967599224.
E-mail addresses: juanangel.aledo@uclm.es (J.A. Aledo), jose.gamez@uclm.es (J.A. Gámez), david.molina@uclm.es (D. Molina).

http://dx.doi.org/10.1016/j.amc.2016.06.005
0096-3003/

Table 1
Example of different types of rankings ($n = 4$).

Rankings	Complete	Incomplete
Without ties	(1|2|3|4)	(1|3|4)
With ties	(1|2, 3|4)	(1|2, 4)

represents a given set of rankings as well as possible. In the case of complete rankings, i.e. permutations, this problem is known as the *Kemeny ranking problem* [5,31]. In the more general case when dealing with incomplete rankings (with or without ties), this problem is known as the *rank aggregation problem* [16,17,39] and has applications to many real-world problems [10,13,16,19,28]. In all the cases the solution to the problem is a permutation (complete ranking without ties). Recall that the rank aggregation problem (and in particular the Kemeny ranking problem) is an NP-hard problem [7,8,25]. For this reason, it is usual to approach this problem by using greedy algorithms. Among them, the Borda algorithm is undoubtedly the preferred one, because of its good trade-off between efficiency and accuracy [5,9]. In particular, it is frequently used as a building-block to deal with problems that require to estimate the consensus permutation many times, e.g. optimization [10] and machine learning [13].

Given a dataset of rankings, the *Borda count* algorithm assigns points to the items according to their positions in the rankings of the dataset (the more preferred an item is, the more points it gets) and finally it computes the consensus ranking by ordering the items from the most valued to the less valued one. For permutations, the points are assigned easily by giving $n-1$ points to the first ranked item, $n-2$ to the second one, and so on. Alternatively the Borda algorithm can be implemented by using a precedence matrix (see Section 4). Some other algorithms (see for instance [2,11,16,35,41]) can also be adapted to use a matrix of preference.

When dealing with arbitrary rankings (i.e. complete or incomplete rankings with or without ties) two different approaches can be followed: (i) to ignore, when assigning points, those items not included in the incomplete ranking (as the Modified Borda algorithm does [18,39]), or (ii) to deal with the uncertainty associated to the items not appearing in a given ranking, that is, taking into account the positions of the ranking in which they could be placed. Our proposal belongs to the second approach and, in particular, it uses the concept of extension sets (see Section 2.1) to manage the unobserved information (see [13,33] for related research on this idea). Moreover, to evaluate how good the obtained consensus ranking is, we introduce a similarity measure based on an extension of the Kendall tau distance [32] that allows to compute the *distance* between any two arbitrary rankings. Different *distances* can be found in the literature for particular cases [1,15–17,20,33], although the Kendall distance and several extensions of it are the most used in this setting.

The contributions of this paper are: (1) to introduce two new preference matrices based on using the extension sets of a ranking, designed to cope with any type of rankings; (2) to provide an efficient method to compute these matrices; and (3) to show how these matrices can be coupled with different algorithms, in particular, we use the Borda and the GreedyOrder algorithms. To test the goodness of our proposal we conduct a broad experimental study, which shows the advantage of using the proposed matrices.

The paper is organized as follows. In Section 2 we provide the notation and necessary definitions over rankings and extension sets. Section 3 is devoted to revise the rank aggregation problem and the Borda algorithm. In Section 4 we detail our proposal and in Section 5 we develop the tools needed to efficiently carry out the computations. Section 6 describes the experimental evaluation and we finish with some conclusions in Section 7.

2. Rankings

Given a set $[[n]] = \{1, 2, \dots, n\}$ of items, a *ranking* π is an order of preference of these (or some of these) items.

Rankings can be classified as *complete* (the n items are ranked) or *incomplete* (only p items are ranked, $2 \le p < n$). On the other hand, rankings can also be classified as *with ties* or *without ties*. Conceptually, a *tie* consists in a lack of preference information among some ranked items. Tied items constitute a *bucket*. Thus, a ranking π can be understood as an order of preference among its (disjoint) buckets. The set of (complete or incomplete, with or without ties) rankings of the elements $1, 2, \dots, n$ will be denoted as $\widetilde{\mathbb{S}}_n$.

A ranking is denoted as a list of items, from most to least preferred, separated by vertical bars. Items between two consecutive vertical bars constitute a *bucket*, so all of them are *tied* regarding the preference criterion. Items in a bucket are separated by commas.

Thus, by

$$\pi = (x_{11}, x_{12} \dots, x_{1j_1} \,|\, x_{21}, x_{22}, \dots, x_{2j_2} \,|\, \dots \,|\, x_{k1}, x_{k2}, \dots, x_{kj_k})$$

with $1 \le j_i$, $1 \le k \le n$ and $2 \le \sum_{i=1}^{k} j_i \le n$, we denote a ranking where $x_{11}, x_{12}, \dots, x_{1j_1}$ are the items in the first bucket (i.e. they are preferred over the remaining items but tied among them), $x_{21}, x_{22, \dots}, x_{2j_2}$ are tied in the second bucket and so on. In particular, we will denote by $B_i(\pi) = \{x_{i1}, x_{i2}, \dots, x_{ij_i}\}$ the ith bucket of π. When no confusion is possible, we will just write B_i instead of $B_i(\pi)$ to simplify the notation.

Table 1 shows examples of the different types of rankings for $n = 4$.