ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Model-based co-clustering for ordinal data

Julien Jacques [a,c,*], Christophe Biernacki [b,c]

[a] Université de Lyon, Lyon 2, ERIC EA 3083, Lyon, France
[b] Laboratoire Paul Painlevé, UMR CNRS 8524, Université de Lille, Lille, France
[c] MODAL team, Inria Lille-Nord Europe, France

ARTICLE INFO

Article history:
Received 27 January 2017
Received in revised form 24 January 2018
Accepted 25 January 2018
Available online xxxx

Keywords:
Latent block model
EM algorithm
Gibbs sampler

ABSTRACT

A model-based co-clustering algorithm for ordinal data is presented. This algorithm relies on the latent block model embedding a probability distribution specific to ordinal data (the so-called BOS or Binary Ordinal Search distribution). Model inference relies on a Stochastic EM algorithm coupled with a Gibbs sampler, and the ICL-BIC criterion is used for selecting the number of co-clusters (or blocks). The main advantage of this ordinal dedicated co-clustering model is its parsimony, the interpretability of the co-cluster parameters (mode, precision) and the possibility to take into account missing data. Numerical experiments on simulated data show the efficiency of the inference strategy, and real data analyses illustrate the interest of the proposed procedure.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Historically, clustering algorithms are used to explore data and to provide a simplified representation of data with a small number of homogeneous groups of individuals (i.e. clusters). With the big data phenomenon, the number of features becomes itself larger and larger, and traditional clustering methods are no more sufficient to explore such datasets. Indeed, the interpretation of a cluster of individuals using for instance a representative of this cluster (mean, mode, ...) is unfeasible since this representative is itself described by a very large number of features. Consequently, there is also a need to summarize the features by grouping them together into clusters.

Two approaches exist: bi-clustering and co-clustering. On the one hand, bi-clustering aims to identify blocks (or bi-clusters) defined as a subset of observations described by a subset of variables. These subsets can overlap. On the other hand, co-clustering aims to define both a partition of the observations and of the variables, and the blocks (or co-clusters) are obtained by crossing both partitions. The main differences are that blocks can overlap in bi-clustering and not in co-clustering, and moreover all features and observations have to belong to a block in co-clustering whereas not necessarily in bi-clustering. Fig. 1 illustrates the differences between both approaches. This work focuses on the co-clustering problem as a natural extension of traditional partition clustering.

Co-clustering algorithms have been introduced to provide a solution by gathering into homogeneous groups both the observations and the features. Thus, the large data matrix can be summarized by a reduced number of blocks of data (or co-clusters). If the earliest (and most cited) methods are probably due to Hartigan (1972, 1975), the model-based approaches have recently proven their efficiency either for continuous, binary, count or contingency data (Govaert and Nadif, 2013; Pledger and Arnold, 2014).

This work focuses on particular type of categorical data, ordinal data, occurring when the categories are ordered (Agresti, 2010). Ordinality is a characteristic of the meaning of measurements (Stevens, 1946), and distinct levels of an ordinal variable

* Corresponding author at: Université de Lyon, Lyon 2, ERIC EA 3083, Lyon, France.
E-mail addresses: julien.jacques@univ-lyon2.fr (J. Jacques), christophe.biernacki@univ-lille1.fr (C. Biernacki).

https://doi.org/10.1016/j.csda.2018.01.014
0167-9473/© 2018 Elsevier B.V. All rights reserved.

UNCORRECTED PROOF

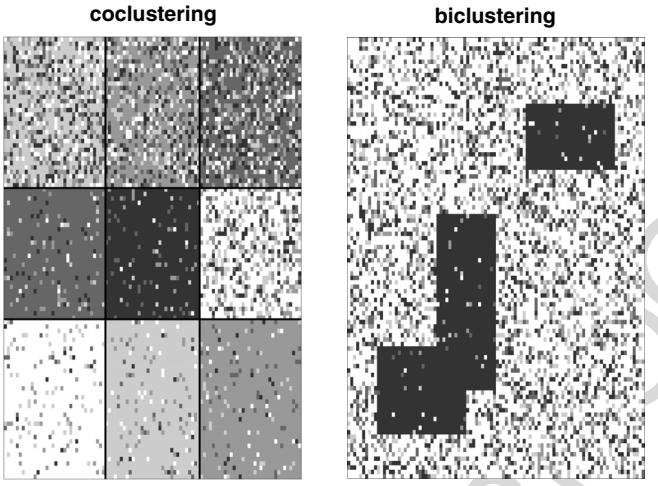

Fig. 1. Co-clustering versus bi-clustering.

differ in degree of dissimilarity more than in quality (Agresti, 2010). Such data are very frequent in practice, as for instance in marketing studies where people are asked through questionnaires to evaluate some products or services on an ordinal scale (Dillon et al., 1994). Another example can be found in medicine, when patients are asked to evaluate their quality of life on a Likert scale (see for instance Cousson-Gélie(2000)), or in vegetation sciences with the Braun-Blanquet scale (Podani, 2006).

However, contrary to nominal categorical data, studied for instance in Celeux and Govaert (2015), ordinal data have received less attention from a clustering point of view, and then, in face of such data, the practitioners often transform them into either quantitative data (associating an arbitrary number to each category, see Kaufman and Rousseeuw (1990) or Lewis et al. (2003) for instance) or into nominal data (ignoring the order information, see the Latent GOLD software (Vermunt and Magidson, 2005)) in order to "recycle" easily related distributions. In order to avoid such extreme choices, some recent works have contributed to define clustering algorithms specific for ordinal data (Gouget, 2006; D'Elia and Piccolo, 2005; Podani, 2006; Giordan and Diana, 2011; Jollois and Nadif, 2011; Biernacki and Jacques, 2016; Ranalli and Rocci, 2016; Fernández et al., 2016). Nevertheless, when the number of features is large, the clustering of observations can be insufficient to summarize the data and a simultaneous clustering of the features could be meaningful.

In a co-clustering context Matechou et al. (2016) recently proposed an approach relying on the proportional odds model, itself assuming that the ordinal response has an underlying continuous latent variable. Unfortunately, the authors did not provide any code or package for their method and thus numerical comparisons are not possible. Let notice that the R package *biclust* (Kaiser et al., 2015) proposes several bi-clustering algorithms, whose bi-clustering goal is not the same than co-clustering (cf. Fig. 1).

In this work, we propose a model-based co-clustering algorithm relying on a recent distribution for ordinal data (BOS for *Binary Ordinal Search* model, Biernacki and Jacques (2016)), which has proven its efficiency for modeling and clustering ordinal data. One of the main advantage of the BOS model is its parsimony and the significance of its parameters. Indeed, in the present work, each co-cluster of data is summarized with only two parameters, one position parameter and one precision parameter. Another advantage of the proposed co-clustering model is its ability to take into account missing data by estimating them during the inference algorithm. Thus, the proposed co-clustering algorithm can be also used in a matrix completion task (see Candès and Recht (2009) for instance).

The paper is organized as follows. Section 2 proposes the co-clustering model whereas its inference and tools for selecting the number of co-clusters are presented in Section 3. Numerical studies (Section 4) show the efficiency of the proposed approach, and two real data applications are presented in Section 5. A discussion concludes the paper in Section 6.

2. Latent block model for ordinal data

The dataset is composed of a matrix of n observations (rows or individuals) of d ordinal variables (columns or features): $\mathbf{x} = (x_{ih})_{1\leq i\leq n, 1\leq h\leq d}$. For simplicity, the ordered levels of x_{ih} will be numbered $\{1, \ldots, m_h\}$, and all m_h's are assumed to be equal: $m_h = m\,(1 \leq h \leq d)$. A natural approach for model-based co-clustering is to consider the latent block model (Govaert and Nadif, 2013), which itself relies on a probability distribution for the data. In the following, the BOS model for ordinal data is presented, then the latent block model and finally their combination for providing the proposed model.

Please cite this article in press as: Jacques J., Biernacki C., Model-based co-clustering for ordinal data. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.01.014.

UNCORRECTED PROOF

Download English Version:

https://daneshyari.com/en/article/6868735

Download Persian Version:

https://daneshyari.com/article/6868735

Daneshyari.com