Computational Statistics and Data Analysis 71 (2014) 3–13

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Model based clustering of customer choice data

Donatella Vicari *, Marco Alfó

Dipartimento di Scienze Statistiche, "Sapienza" Universitá di Roma, P.le Aldo Moro, 5 - 00185 Rome, Italy

HIGHLIGHTS

- A two-level finite mixture model for clustering customers and products is proposed.
- Clusters of products nested in segments of customers are determined.
- Customer/product features influence the allocation to segments/clusters.
- An appropriate EM algorithm is presented for the special case of purchase counts.
- The application shows high purchase rate segments linked with clusters of products.

ARTICLE INFO

Article history:
Received 18 September 2012
Received in revised form 16 September 2013
Accepted 16 September 2013
Available online 23 September 2013

Keywords:
Model-based clustering
Conditional logit
Multinomial logit
Co-clustering
Bi-clustering

ABSTRACT

In several empirical applications analyzing customer-by-product choice data, it may be relevant to partition individuals having similar purchase behavior in homogeneous segments. Moreover, should individual- and/or product-specific covariates be available, their potential effects on the probability to choose certain products may be also investigated. A model for joint clustering of statistical units (customers) and variables (products) is proposed in a mixture modeling framework, and an appropriate EM-type algorithm for ML parameter estimation is presented. The model can be easily linked with similar proposals appeared in various contexts, such as co-clustering of gene expression data, clustering of words and documents in web-mining data analysis.

1. Introduction

We propose a model-based approach to cluster individuals and products in disjoint individual- and product-specific groups, where the corresponding partitions are dependent. We will refer to individual-specific groups as *segments*, while the product-specific groups will be referred to as *clusters*. The motivation arises from empirical situations where customer data are analyzed to investigate on factors affecting the purchase behavior towards several products. The idea is to define individual-specific segments which are homogeneous in terms of customer product choices; the prior (conditional) probability for an individual to belong to a given segment is assumed to be a function of individual-specific covariates, and we are interested in investigating how such characteristics affect the segment memberships. We can also imagine that, within an individual-specific segment, a partition of the products may be identified depending on their characteristics. For example, customers with a given purchase profile may prefer a particular subset of products because of their features and such preferences may vary within segments of customers. In this view, we may be interested in studying whether individuals in a specific segment (representing a prototypical purchase behavior) choose specific subsets of products for their features. In this perspective, we aim at jointly partitioning customers and products to investigate about the determinants of the customer

* Corresponding author. Tel.: +39 0649690210.
E-mail address: donatella.vicari@uniroma1.it (D. Vicari).

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.09.014

choices. This purpose might be linked with methods for joint partitioning of genes and tissues (or experimental conditions) in microarray data analysis (see e.g. Martella et al. (2008)), of words and documents in web data analysis (see e.g. Li and Zha (2006)), or, in general, when latent block-based clustering is pursued (see e.g. Govaert and Nadif (2003)). Further interesting links can be established with multi-layer mixtures, see e.g. Li (2005), and with hierarchical mixture of experts models, see e.g. Titsias and Likas (2002). Such methodological connections will be further discussed to better focus and motivate our proposal.

The plan of the paper is as follows. In Section 2, the model is introduced in a general framework, and, in Section 3, a ML approach to parameter estimation is described. An EM-type algorithm is detailed in Section 4 in the context of observed count data. In Section 5, the analysis of a benchmark data set is proposed. In the last section, concluding remarks and the future research agenda are discussed.

2. The model

Let $\mathbf{Y}_i$, $i = 1, \dots, n$, be a p-dimensional random vector and let $\mathbf{y}_i$, $i = 1, \dots, n$ represent the corresponding realization in a sample of size n; let $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)^\top$ denote the (n, p) matrix of the observed values y_{ij}, for individual $i = 1, \dots, n$ and variable $j = 1, \dots, p$. Just to give an example, and without loss of generality, we may suppose to consider n customers and p products, where y_{ij} represents the number of items of the j-th product the i-th customer has bought in a given time interval.

In addition, we assume that a set of outcome-specific (price, weight, type of package, etc.) and of individual-specific (age, gender, educational level, income, etc.) covariates have been also recorded. Let $\mathbf{x}_i$ and $\mathbf{z}_j$ denote the vectors containing the characteristics of the i-th individual, and of the j-th product, $j = 1, \dots, p$, respectively. In the following, for the sake of clarity, groups of individuals and products will be termed *segments* and *clusters*, respectively.

We adopt a mixture model framework and start by assuming that the population consists of G segments in proportions $\pi_1, \dots, \pi_G$, $\sum_{g=1}^{G} \pi_g = 1$, $\pi_g \geq 0$, $\forall g = 1, \dots, G$. An unobservable G-dimensional binary indicator vector $\mathbf{u}_i = (u_{i1}, \dots, u_{iG})$ is associated with each individual and has a unique non null element, indicating whether the i-th individual belongs to the g-th segment or not, $i = 1, \dots, n$, $g = 1, \dots, G$, see e.g. Titterington et al. (1985). In such a mixture sampling scheme, the sample is obtained by first drawing, independently for each unit, the corresponding segment label, u_{ig}, $g = 1, \dots, G$ from the population with probability density function (pdf) $h(\mathbf{u}_i \mid \boldsymbol{\pi})$; then, values of the outcome variables are drawn from the population with pdf given by

$$f_g\left(\mathbf{y}_i \mid \boldsymbol{\theta}_g\right) = f\left(\mathbf{y}_i \mid u_{ig} = 1\right), \tag{1}$$

where $f_g = f\left(\mathbf{y}_i \mid \boldsymbol{\theta}_g\right)$ is the g-th segment-specific density with indexing parameter vector given by $\boldsymbol{\theta}_g$. As usual, the individual segment indicators $\mathbf{u}_i = (u_{i1}, \dots, u_{iG})$ are assumed to be independent multinomial random variables with probabilities given by $\boldsymbol{\pi} = (\pi_1, \dots, \pi_G)$. Thus, each observation $\mathbf{y}_i$, $i = 1, \dots, n$, is sampled from the finite mixture density

$$f(\mathbf{y}_i \mid \boldsymbol{\pi}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G) = \sum_{g=1}^{G} \pi_g f_g\left(\mathbf{y}_i \mid \boldsymbol{\theta}_g\right). \tag{2}$$

Let $(\mathbf{y}_i, \mathbf{u}_i)$, $i = 1, \dots, n$ be an observed sample drawn under such a sampling scheme; the so-called *complete data* density function is given by

$$f(\mathbf{y}_i, \mathbf{u}_i \mid \boldsymbol{\theta}, \boldsymbol{\pi}) = f(\mathbf{y}_i \mid \boldsymbol{\theta}, \mathbf{u}_i)\, h(\mathbf{u}_i \mid \boldsymbol{\pi}) = \prod_{g=1}^{G} \left[\pi_g f_g\left(\mathbf{y}_i \mid \boldsymbol{\theta}_g\right)\right]^{u_{ig}}, \tag{3}$$

and the resulting *complete-data* log-likelihood may be expressed as follows

$$\begin{aligned} \ell_c(\boldsymbol{\theta}, \boldsymbol{\pi}) &= \sum_{i=1}^{n} \sum_{g=1}^{G} u_{ig} \log\left[\pi_g f_g\left(\mathbf{y}_i \mid \boldsymbol{\theta}_g\right)\right] \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} u_{ig} \left\{\log\left(\pi_g\right) + \log\left[f_g\left(\mathbf{y}_i \mid \boldsymbol{\theta}_g\right)\right]\right\}. \end{aligned} \tag{4}$$

The estimation of the segment-specific parameters $\boldsymbol{\theta}_g$, $g = 1, \dots, G$ and the priors, π_g, $g = 1, \dots, G$, is usually based on an EM-type algorithm. Such estimates help us to identify the segment-specific densities and, as a byproduct, to assign each individual to a segment, through a maximum a posteriori (MAP) rule. That is, the i-th individual is assigned to the g-th segment if the following condition on the posterior probabilities holds:

$$\Pr\left(\boldsymbol{\theta} = \boldsymbol{\theta}_g \mid \mathbf{y}_i\right) = \max_{l} \Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_l \mid \mathbf{y}_i) \quad l = 1, \dots, G.$$

Let us assume that, within the g-th individual-specific segment ($g = 1, \dots, G$), we may identify a partition of the products in K_g clusters. Fig. 1 may clarify what kind of partition of the observed data we are discussing about.