Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5368–5376
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A hierarchical mixture model for clustering three-way data sets

Jeroen K. Vermunt*

Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Available online 24 August 2006

Abstract

Three-way data sets occur when various attributes are measured for a set of observational units in different situations. Examples are genotype by environment by attribute data obtained in a plant experiment, individual by time point by response data in a longitudinal study, and individual by brand by attribute data in a market research survey. Clustering observational units (genotypes/individuals) by means of a special type of the normal mixture model has been proposed. An implicit assumption of this approach is, however, that observational units are in the same cluster in all situations. An extension is presented that makes it possible to relax this assumption and that because of this may yield much simpler clustering solutions. The proposed extension—which includes the earlier model as a special case—is obtained by adapting the multilevel latent class model for categorical responses to the three-way situation, as well as to the situation in which responses include continuous variables. An efficient EM algorithm for parameter estimation by maximum likelihood is described and two empirical examples are provided.

Keywords: Clustering; Three-way data; Finite mixture model; Longitudinal data; EM algorithm; Multilevel latent class model

1. Introduction

An example of a three-way data set is data collected in plant experiments where various attributes are measured on genotypes grown in several environments. This would be a genotype by environment by attribute data set. Basford and McLachlan (1985) proposed a variant of the normal mixture model for the analysis of such three-way data, where the aim is to cluster genotypes by explicitly taking into account the information on attributes and environments simultaneously. This is achieved by a multivariate normal mixture model with cluster-, environment-, and attribute-specific means, and with non-zero cluster-specific covariances between attributes within environments. More recently, Hunt and Basford (1999, 2001) extended the approach to cases with categorical attributes and with not all attributes observed on all genotypes. Meulders et al. (2002) proposed a restricted latent class model for the analysis of three-way dichotomous attribute data.

Other examples of three-way data include longitudinal data on multiple response variables—person by time point by response data—or data from experiments in which individuals provide multiple ratings for multiple objects (products, brands) or report on possible behaviors shown in multiple situations, yielding person by object by attribute and person by situation by behavior data, respectively. Other examples consist of data sets in which objects are rated on multiple attributes by multiple experts, such as exams with multiple questions corrected by multiple raters or products evaluated on multiple attributes by multiple raters. In the remaining, I will refer to the three ways of the data sets as cases,

* Tel.: +31 13 4662748; fax: +31 13 4663002.
E-mail address: j.k.vermunt@uvt.nl.

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.08.005

situations, and attributes, respectively. The aim of the application of a mixture model is to cluster cases based on measured attributes in various situations. Clusters will also be referred to as (latent) classes and groups.

An important characteristic of the Basford and McLachlan (B&M) mixture model for three-way data, as well as of the other variants mentioned above, is that cases are assumed to belong to the same cluster in all investigated situations. I propose an alternative mixture model for three-way data that relaxes this assumption: cases may be in a different latent class depending on the situation or, more specifically, cases are clustered with respect to the probability of being in a particular latent class at a certain situation. The basic idea is to treat the three ways as hierarchically nested levels and assume that there is a mixture distribution at each of the two higher levels; i.e., one at the case and one at the case-in-situation level. The proposed model is an adaptation of the multilevel latent class model by Vermunt (2003) to continuous responses, as well as to the specific model structures needed for dealing with three-way data. A nice feature is that it has the B&M three-way mixture model as a special case.

An important advantage of the proposed modelling approach is that it may yield more parsimonious solutions—solutions with less clusters—with an even better description of the data than the B&M model. Moreover, interpretation of results may be easier and the model may be more in agreement with reality and thus more meaningful. For example, in a longitudinal data application it is unrealistic to assume that individuals are in the same latent class at each time point or in a multiple experts study it is unrealistic to assume that each expert classifies an object in the same latent class.

Böhning et al. (2000) proposed a state–space mixture model in which, in fact, two ways (case by time point) are collapsed into one way. A standard mixture model is subsequently adopted, which implies that observations of the same case at different time points are assumed to be independent of one another. An advantage of the hierarchical mixture model described below is that it can take into account dependencies between repeated observations within cases. It should be noted that the hierarchical mixture model has the state–space mixture model of Böhning et al. (2000) as a special case; that is, as the limiting case in which there is only one higher-level mixture component.

The remaining of this article is organized as follows. Using B&M's model as the starting point, I first describe the simplest form of the new model, and subsequently introduce variants such as restricted multivariate normals, models for categorical and mixed responses, and models with covariates and regression type constraints. Subsequently, I show how parameter estimation can be performed using a special variant of the EM algorithm which is implemented in the Latent GOLD mixture modelling software (Vermunt and Magidson, 2005). The new approach is illustrated with two empirical examples.

2. Mixture models for three-way data

2.1. Basford and McLachlan's mixture model

Following a similar notation as in McLachlan and Peel (2000, p. 114) and Hunt and Basford (2001), suppose that the responses on P attributes were recorded in N cases, each of which was observed in R situations. Let $\mathbf{y}_{ir}$ be a $P \times 1$ vector containing the values of the P attributes of case i in situation r, for $i = 1, \ldots, N$; $r = 1, \ldots, R$. The $RP \times 1$ observation vector $\mathbf{y}_i$ is given by

$$\mathbf{y}_i = \left(\mathbf{y}'_{i1}, \ldots, \mathbf{y}'_{iR}\right)',$$

where $\mathbf{y}_i$ contains the multi-attribute responses of the ith case in all R situations. Under the mixture model proposed by Basford and McLachlan (1985), it is assumed that cases belong to one of K possible groups or latent classes $G_1, \ldots, G_K$ in proportions $\pi_1, \ldots, \pi_K$, respectively, where $\sum \pi_k = 1$ and $\pi_k \geqslant 0$ for $k = 1, \ldots, K$. The responses of case i in situation r have a multivariate normal distribution conditional on group G_k; i.e., $\mathbf{y}_{ir} \sim N(\boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}_k)$. The mixture model for three-way data proposed by Basford and McLachlan (1985) has the following form:

$$f\left(\mathbf{y}_i\right) = \sum_{k=1}^{K} \pi_k \prod_{r=1}^{R} f_k\left(\mathbf{y}_{ir}; \boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}_k\right). \tag{1}$$

Note that the values of the within-class covariance matrices are constant across situations, whereas the class-specific attribute means differ across situations. An important assumption is that conditional on the class membership of case i the responses in the different situations are independent of one another. It is, however, impossible to relax that