Journal of Statistical Planning and Inference 142 (2012) 2976–2984

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

Model-based clustering, classification, and discriminant analysis of data with mixed type

Ryan P. Browne, Paul D. McNicholas*

Department of Mathematics and Statistics, University of Guelph, ON, Canada N1G 2W1

ARTICLE INFO

Article history:
Received 20 January 2011
Received in revised form
28 March 2012
Accepted 1 May 2012
Available online 9 May 2012

Keywords:
Classification
Clustering
Discriminant analysis
Latent variables
Mixed type
Mixture models

ABSTRACT

We propose a mixture of latent variables model for the model-based clustering, classification, and discriminant analysis of data comprising variables with mixed type. This approach is a generalization of latent variable analysis, and model fitting is carried out within the expectation-maximization framework. Our approach is outlined and a simulation study conducted to illustrate the effect of sample size and noise on the standard errors and the recovery probabilities for the number of groups. Our modelling methodology is then applied to two real data sets and their clustering and classification performance is discussed. We conclude with discussion and suggestions for future work.

1. Introduction

The purpose of this paper is to introduce methodology to perform model-based clustering, classification, and discriminant analysis on data sets comprising variables of mixed type. We say that data comprise variables of mixed type when more than one type of variable is present; e.g., data may contain both categorical and interval variables (cf. Section 4.1). Within the literature, little work has been done on the clustering or classification of such data. Our approach is based on extending the latent variable analysis approach to handle variables of mixed type (cf. Bartholomew and Knott, 1999). In this paper, we propose a way to incorporate methodologies from latent variable mixture models.

The idiom 'latent variable model' is a blanket term for a class of models that includes latent class analysis, latent trait analysis, factor analysis, and latent factor models. Most latent variable models (cf. Eq. (1)) assume that the observed (or manifest) variables within an observation $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ are independent given an unobservable, or latent, variable $\mathbf{Y}_i = \mathbf{y}_i$ with dimension $q < p$. However, some latent variable models allow dependence between variables conditional upon the latent variable. Herein, we will assume conditional independence. In both latent trait and factor analyses, the latent variable is assumed to have a standard Gaussian distribution; the observed data are either categorical (latent trait analysis) or Gaussian (factor analysis).

Within this framework, the density function of p-dimensional random vector $\mathbf{X}_i$ has the form:

$$f(\mathbf{x}_i) = \int f(\mathbf{x}_i|\mathbf{y},\boldsymbol{\theta})h(\mathbf{y})\,d\mathbf{y} = \int \prod_{j=1}^{p} g_i(x_{ij}|\mathbf{y},\boldsymbol{\theta}_j)h(\mathbf{y})\,d\mathbf{y}, \tag{1}$$

* Corresponding author. Tel.: +1 519 824 4120x53136; fax: +1 519 837 0221.
E-mail addresses: rbrowne@uoguelph.ca (R.P. Browne), paul.mcnicholas@uoguelph.ca (P.D. McNicholas).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.05.001

where $g_i(x_{ij}|\mathbf{y},\boldsymbol{\theta}_j)$ is the conditional distribution of X_{ij} given $\mathbf{Y}=\mathbf{y}$ and parameters $\boldsymbol{\theta}_j$, and $h(\mathbf{y})$ is the marginal distribution of $\mathbf{Y}$. If the variable X_{ij} is categorical with levels $0,1,\ldots,c_j-1$, then the conditional distribution $X_{ij}|\mathbf{y},\boldsymbol{\theta}_j$ is Bernoulli with success probability:

$$g(x_{ij}=s|\mathbf{y},\boldsymbol{\theta}_j)=\frac{\prod_{s=0}^{c_i-1}[\exp\{\beta_{js0}+\boldsymbol{\beta}_{js}'\mathbf{y}\}]^{x_{ij}(s)}}{1+\sum_{k=1}^{c_j-1}\exp\{\beta_{jk0}+\boldsymbol{\beta}_{jk}'\mathbf{y}\}}, \tag{2}$$

where $x_{ij}(s)=1$ if the response falls into category s and $x_{ij}(s)=0$ otherwise, $\boldsymbol{\beta}_{js}=(\beta_{js1},\ldots,\beta_{jsq})$ is a vector of coefficients, $\beta_{js0}\in\mathbb{R}$, and $\boldsymbol{\theta}_j$ again denotes the parameters for variable j. The category associated with $s=0$ is called the reference category. On the other hand, if X_{ij} is an interval variable then it is assumed to have a Gaussian conditional distribution given by

$$g(x_{ij}|\mathbf{y},\boldsymbol{\theta}_j)=\frac{1}{\sqrt{2\pi\sigma_j}}\exp\left\{-\frac{1}{2}\left(\frac{x_{ij}-\beta_{j0}-\boldsymbol{\beta}_j'\mathbf{y}}{\sigma_j}\right)^2\right\}, \tag{3}$$

where $\boldsymbol{\beta}_j=(\beta_{j1},\ldots,\beta_{jq})$. This methodology can be seen as combining latent trait and latent factor analyses into a single model. Note that we are using a linear latent variable model (cf. Eqs. (2) and (3)). Although not considered herein, one could extend our approach to accommodate non-linear models (e.g., Lawrence, 2005).

In clustering and classification applications, the goal is to find sub-populations within a given data set. Finite mixture models assume that a population is a convex combination of a finite number of densities and so they are naturally suited to clustering and classification problems. Finite mixture models have been used for clustering for almost 50 years (cf. Wolfe, 1963; Fraley and Raftery, 2002) and have received renewed attention over the past decade or so. A p-dimensional random vector $\mathbf{X}$ is said to arise from a parametric finite mixture distribution if, for all $\mathbf{x}\subset\mathbf{X}$, we can write $\xi(\mathbf{x}|\boldsymbol{\vartheta})=\sum_{g=1}^{G}\eta_g\psi_g(\mathbf{x}|\boldsymbol{\theta}_g)$, where $\eta_g>0$ such that $\sum_{g=1}^{G}\eta_g=1$ are the mixing proportions, $\boldsymbol{\vartheta}=(\eta_1,\ldots,\eta_G,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_G)$ is the vector of parameters, and $\psi_1(\mathbf{x}|\boldsymbol{\theta}_1),\ldots,\psi_G(\mathbf{x}|\boldsymbol{\theta}_G)$ are the component densities. Gaussian mixture models have garnered most of the attention within the literature due to their mathematical tractability (recent examples include Dean et al., 2006; McNicholas, 2010). However, the efficacy of the Gaussian approach is confined to continuous variables; i.e., to interval data.

The remainder of this paper is laid out as follows. In Section 2, we introduce our modelling framework before discussing parameter estimation and model selection. We also explain why mixtures of latent class analyzers are not considered. In Section 3, we present a simulation study to illustrate our mixture modelling approach to clustering and another to illustrate classification and discriminant analysis. Our approach is then applied to real data (Section 4), where clustering, classification, and discriminant analysis are again considered. The paper concludes in Section 5 with a summary and suggestions for future work.

2. Methodology

2.1. Model-based clustering framework

Model-based clustering, model-based classification, and model-based discriminant analysis are similar frameworks, with model fitting in one case analogous to the other. First, we shall illustrate our proposed method within the model-based clustering paradigm; we extend the latent variable model depicted in Eq. (1) by applying mixture model methodology. We assume that the independent observations $\mathbf{x}_1,\ldots,\mathbf{x}_n$ arise from a finite mixture model and that the effect of the latent variable $\mathbf{y}$ is different for each component. If the observed data have mixed type, then the mixture of latent variables model has the form

$$f(\mathbf{x}_i)=\int\left[\sum_{g=1}^{G}f(\mathbf{x}_i|\mathbf{y},\boldsymbol{\theta}_g)\eta_g\right]h(\mathbf{y})\,d\mathbf{y}=\sum_{g=1}^{G}\eta_g\left[\int\prod_{j=1}^{p}f(x_{ij}|\mathbf{y},\boldsymbol{\theta}_{gj})h(\mathbf{y})\,d\mathbf{y}\right] \tag{4}$$

and the log-likelihood is

$$l(\boldsymbol{\vartheta}|\mathbf{x}_1,\ldots,\mathbf{x}_n)=\sum_{i=1}^{n}\log\left\{\sum_{g=1}^{G}\eta_g\left[\int\prod_{j=1}^{p}f(x_{ij}|\mathbf{y},\boldsymbol{\theta}_{gj})h(\mathbf{y})\,d\mathbf{y}\right]\right\}. \tag{5}$$

When the manifest variables in Eq. (4) have a categorical conditional distribution, the integral cannot be solved analytically; however, when they have a Gaussian conditional distribution, this integral can be solved analytically.

2.2. Model fitting

Parameter estimation is carried out using an expectation-maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm is an iterative technique that facilitates maximum likelihood estimation when data are incomplete or treated as being incomplete. In our case, the missing data comprise the group memberships and the latent variables. We denote the