Computational Statistics and Data Analysis 103 (2016) 350–366

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Comparing classical criteria for selecting intra-class correlated features in Multimix

Lynette A. Hunt [a,*], Kaye E. Basford [b]

[a] *University of Waikato, Hamilton, New Zealand*

[b] *University of Queensland, Brisbane, Australia*

ARTICLE INFO

Article history:
Received 1 May 2015
Received in revised form 26 May 2016
Accepted 27 May 2016
Available online 3 June 2016

Keywords:
Model selection criteria
Finite mixture models
Mixed data
Multimix

ABSTRACT

The mixture approach to clustering requires the user to specify both the number of components to be fitted to the model and the form of the component distributions. In the Multimix class of models, the user also has to decide on the correlation structure to be introduced into the model. The behaviour of some commonly used model selection criteria is investigated when using the finite mixture model to cluster data containing mixed categorical and continuous attributes. The performance of these criteria in selecting both the number of components in the model and the form of the correlation structure amongst the attributes when fitting the Multimix class of models is illustrated using simulated data and a real medical data set. It is found that criteria based on the integrated classification likelihood have the best performance in detecting the number of clusters to be fitted to the model and in selecting the form of the component distributions. The performance of the Bayesian information criterion in detecting the correct model depends on the partitioning structure among the attributes while the Akaike information criterion and classification likelihood criterion perform in a less satisfactory way.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Finite mixture models are widely used in a variety of applications to model the distributions of various events and to cluster data sets, see for example McLachlan and Peel (2000), McLachlan and Chang (2004), Everitt et al. (2011), Stahl and Sallis (2012) and Melnykov (2013). This paper focuses on the use of the mixture model approach to clustering which provides a formal statistical framework on which the clustering can be based. The procedure gives a probabilistic clustering that allows for overlapping clusters which correspond to the components in the model, and where each component in the finite mixture model corresponds to a cluster in the data. The probability that an observation belongs to each of the clusters can be obtained from the estimates of the posterior probabilities of cluster membership. A definitive partitioning of the observations into components (clusters) is obtained by assigning each observation to the component to which it has highest probability of belonging. The finite mixture model requires the specification of both the form of the density function of each of the underlying components and the number of components to be fitted in the model.

[*] Corresponding author.
E-mail addresses: lah@waikato.ac.nz (L.A. Hunt), k.e.basford@uq.edu.au (K.E. Basford).

http://dx.doi.org/10.1016/j.csda.2016.05.018
0167-9473/© 2016 Elsevier B.V. All rights reserved.

1.1. Number of components

Prior knowledge concerning the number of components, K, in the mixture reduces the complexity of the analysis. However there are many situations where there is no *a priori* knowledge of the number of components to be fitted, and thus finding the number of components present in the data becomes part of the clustering problem.

An obvious way of approaching this problem is to use the likelihood ratio statistic λ to test for the smallest value of K compatible with the data. However when testing for the number of components in a mixture, the usual regularity conditions do not hold for $-2\log\lambda$ to have its standard asymptotic null distribution of χ^2 with the degrees of freedom equal to the difference between the number of parameters under the full and reduced models. Accounts of the breakdown of the regularity conditions are given for example, by Hartigan (1977, 1985a,b), Titterington (1981), Titterington et al. (1985), Ghosh and Sen (1985), McLachlan and Basford (1988), and McLachlan and Peel (2000).

An alternative procedure is to use a bootstrap approach. McLachlan (1987) proposed a resampling procedure that involves a bootstrapped likelihood ratio test. Bootstrap samples are generated from the finite mixture model fitted under the null hypothesis of K components, where the parameters of the mixture are the likelihood estimates after fitting a K component model to the original sample. The value of the likelihood ratio statistic is computed for each of the bootstrap samples generated after fitting mixtures with K and K' components, where $K' > K$. The process is repeated independently B times. The replicated values of $-2\log\lambda$ formed from the successive bootstrap samples provide an assessment of the true null distribution of $-2\log\lambda$: see also Feng and McCulloch (1996), and McLachlan and Peel (1997). However, the problem with bootstrap methods is that they can be computationally intensive when the number of components is large, and little is known about the performance of the test when the distributional and model assumptions are violated (see Nyland et al., 2007). See also Lo et al. (2001) for details on another approach called the Lo–Mendell–Rubin likelihood ratio test which uses an approximation of the distribution of the difference of the two log likelihoods.

The use of information criteria to estimate the number of components of a finite mixture has become increasingly popular in model based cluster analysis. Information criteria allow the user to quantify the differences between a candidate set of models and help determine the number of components to be fitted to the mixture model. Many criteria have been proposed with some criteria derived within a Bayesian framework. The authors of this paper have used criteria that are Bayesian based, information criteria and classification criteria. See for example, McLachlan and Peel (2000, chapter 6), Fraley and Raftery (2002), Miloslavsky and van der Laan (2003) and McLachlan and Rathnayake (2014) plus the references therein for discussions on other approaches to the problem of determining the number of components.

The specification of the component distributions is also required in the fitting of a mixture model. There has been extensive use of mixtures where the component distributions are multivariate normal and there has been much interest in determining the number of components to be fitted to this model. McLachlan and Ng (2000) report Monte Carlo simulations to compare the performance of some criteria with that of classical criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), for determining the number of components in mixtures of multivariate normals.

1.2. Covariance structure for models

While it is common to take the component distributions to be multivariate normal, decisions still need to be made on the structure of the components' covariance matrices. There is the unrestricted case where the component covariances Σ_k are unequal, however this may be too general for many situations in practice. Often the component covariances are restricted to being the same ($\Sigma_k = \Sigma$ for $k = 1, \dots, K$), but this can have an adverse effect on the resulting clustering (Chapter 3, McLachlan and Peel, 2000).

Another way of proceeding is to adopt some model that is intermediate between homoscedasticity and the general unrestricted heteroscedastic case. Several authors (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Bensmail et al., 1997) have used the eigenvalue decomposition of the component covariance matrices in Gaussian mixtures to propose models for clustering. The covariance matrix Σ_k can be written in the form $\Sigma_k = \lambda_k D_k A_k D_k'$ where D_k is the matrix of eigenvectors of Σ_k, A_k is a diagonal matrix with the normalized eigenvectors of Σ_k on the diagonal in decreasing order with $|A_k| = 1$, and $\lambda_k = |\Sigma_k|^{\frac{1}{d}}$ where d denotes the number of variables. The volume, orientation and shape of the kth component are determined by λ_k, D_k and A_k respectively. Celeux and Govaert (1995) and Bensmail and Celeux (1996) consider 14 different models corresponding to different assumptions on the components' covariance matrices.

Biernacki and Govaert (1999) performed Monte Carlo simulations using two component bivariate Gaussian mixtures with different covariance matrices to compare the performance of several classical criteria in selecting a relevant and parsimonious model. The covariance matrices for the component distributions were determined using the 14 models relating to different assumptions on the component covariance matrix. They performed simulations using small ($n = 40$) and larger ($n = 200$) samples where the components were mixed in both equal and different proportions.

Hunt (1996) and Hunt and Jorgensen (1996, 1999) proposed a set of models that they termed the Multimix class of mixture models. The Multimix approach uses a form of conditional independence within the components, and can be used to cluster data containing both categorical and continuous attributes. When using the Multimix approach to clustering, Hunt (1996) suggested that a form of forward selection of covariates be used for selecting the correlation structure in the model.

Download English Version:

https://daneshyari.com/en/article/415282

Download Persian Version:

https://daneshyari.com/article/415282

Daneshyari.com