CLUSTERING MULTINOMIAL OBSERVATIONS VIA FINITE AND INFINITE MIXTURE MODELS AND MCMC ALGORITHMS

Mario Medvedovic, University of Cincinnati Medical Center
Mario Medvedovic, Department of Environmental Health, University of Cincinnati Medical Center, 3223 Eden Av., Cincinnati, OH, 45267-0056 (Email: medvedm@email.uc.edu).

Key Words: Cluster Analysis; Multinomial Distribution; Gibbs Sampling; Finite Mixture Model; Infinite Mixture Model; Dirichlet Process Mixtures

ABSTRACT

Traditional statistical clustering procedures based on finite mixtures model require the number of mixture components to be known prior to the analysis. Establishing the number of mixture components from the data is generally a difficult problem involving the comparison of models with different number of parameters. We used the Bayesian infinite mixture approach to devise a clustering procedure that does not require the number of mixture components to be specified in advance. The performance of this model is compared to the performance of the finite mixture approach when the number of components is known as well as when the number of components is estimated using AIC criterion. We showed that the infinite mixture procedure offers comparable results to the finite mixtures approach.

Introduction

Suppose that T multinomial observations $\mathbf{X}=(\mathbf{x}_1`,\ldots,\ \mathbf{x}_T`)$ were generated as independent observations from Q (Q<T) different multinomial random variables defined with parameter vectors $\mathbf{p}_1,\ldots,\mathbf{p}_Q$. That is, for each $i\in\{1,\ldots,T\}$ there exists a unique $j\in\{1,\ldots,Q\}$ such that $\mathbf{x}_i$ is a realization of the multinomial random variable with the probability density function

$$f_{mult}(\mathbf{y}\ \ |\mathbf{p}_j) = \frac{N!}{y_1!\ldots y_M!} p_{j1}^{y_1}\ldots p_{jM}^{y_M} \qquad (1)$$

where $N = y_1+\ldots+y_M = x_{i1}+\ldots+x_{iM}$ for all i=1,..,T. Each of the parameter vectors $\mathbf{p}_1,\ldots,\mathbf{p}_Q$ of probability density functions in (1) defines a cluster of mutually similar multinomial observations. The data can be regarded to be generated by the random process in which the parameter vector $\mathbf{p}_j$ is selected with the probability $\text{Prob}(\mathbf{p}_j)=\pi_j$; then the observation $\mathbf{x}$ is generated with the probability $f_{mult}(\mathbf{x}|\mathbf{p}_j)$. That is, data is considered to be a random sample from the Q-component mixture distribution

$$f_{mixture}(\mathbf{x}|\mathbf{p}_j) = \overset{Q}{\underset{j=1}{}}\ \pi_j\ f_{mult}(\mathbf{x}|\mathbf{p}_j) \qquad (2)$$

In a clustering problem, the goal is to identify groups of observation that are generated by the same mixture component. Let c_i be the classification variable indicating the cluster to which the i^th^ observation belongs ($c_i=k$ means that the i^th^ observation belongs to the k^th^ cluster). The assignment vector $\mathbf{c}=(c_1,\ldots,c_Q)`$ defines completely the distribution of all T spectra among Q clusters. Let π_j denote the proportion of data coming from the j^th^ cluster, $\pi_j =(\overset{T}{\underset{i=1}{}} I(c_j = j)\ )/T$. Prior to taking into account $\mathbf{x}_i$, the probability of observing the classification variable c_i is $p(c_i)=\prod_{j=1}^{Q}\pi_j^{I(c_i=j)}$. It is further assumed that $\mathbf{x}_1,\ldots,\mathbf{x}_T$ given $c_1,\ldots,z_T$, respectively, are conditionally independent, and

$$p(\mathbf{x}_i|c_i) = \mathop{}_{j=1}^{Q} I(c_i = j)\ f_{mult}(\mathbf{x}_i \mid \mathbf{p}_j)\text{, for any } i = 1,\ldots,T.$$

Hence, the Classification Likelihood for the data is given by

$$L_C(\mathbf{c}, \mathbf{P}) = \prod_{i=1}^{T} \mathop{}_{j=1}^{Q} I(c_i = j)\, \pi_j\, f_{mult}(\mathbf{x}_i \mid \mathbf{p}_j)\ ,$$

where $\mathbf{P}$=($\mathbf{p}_1$\`,…,$\mathbf{p}_Q$\`). The Classification Likelihood based clustering procedure consists of finding ($\mathbf{c}$, $\mathbf{P}$) that maximize L_C(.)

In terms of the EM algorithm for finite mixtures (Dempster, Laird and Rubin, 1977), ($\mathbf{c}$, $\mathbf{X}$) represents the complete data and L_C(.) is the probability distribution of the complete data. It can also be shown that clustering obtained by maximizing L_C is identical to the clustering generated by the classical finite mixture approach (Celeux and Govaert, 1992). Therefore we refer to this approach as the finite mixture clustering. We have previously shown (Medvedovic, et al. 2000) that this approach can recreate original structure of data given that the number of clusters is known. However, the identification of the classification vector $\mathbf{c}^*$ maximizing L_C turned out to be a difficult problem due to inability of available maximization algorithms to avoid sub-optimal local maxima (Medvedovic, et al. 2000). Furthermore, assessing the true number of clusters from the data is a difficult problem on its own adding an additional level of uncertainty to the clustering result. In this article, we applied the Infinite Mixtures model as described by Neal, 2000 to clustering multinomial observations and compared its performance to the finite mixture approach. The major advantage of this approach is that the number of clusters needs not to be known. The clustering procedure consists of sampling from the posterior distribution of classification vectors using a Gibbs sampler. We used the sequence of clustering assignments generated by the Gibbs sampler to identify the optimal assignment vector from the distribution of all pair-wise assignments of individual observations.

Infinite Mixtures Model

Consider the following formulation of the problem in terms of a hierarchical model (Neal, 1998)

$$f(\mathbf{x}_i \mid c, \mathbf{P}) = f_{mult}(\mathbf{x}_i \mid \mathbf{p}_{c_i})$$

$$f(c_i) = \prod_{j=1}^{Q} \pi_j^{I(c_i=j)} \qquad (3)$$

$$\mathbf{p}_j \sim \text{Dirichlet}(\beta,\ldots,\beta)$$

$$\pi_j \sim \text{Dirichlet}(\alpha/Q,\ldots,\alpha/Q)$$

Except for the prior distributions for the parameters of individual mixture components and the prior distribution for the mixing proportions, this model is equivalent to the mixture model (2). When Q approaches infinity, the conditional probability distributions for individual classification variables are (Neal, 2000)

$$\Pr ob(c_i = c \mid \mathbf{c}_{-i}, \mathbf{x}_i) =$$

$$b\frac{n_{-i,c}}{T-1+\alpha}\ f_{mult}(\mathbf{x}_i \mid \mathbf{p})\, dH_{-i,c}(\mathbf{p})$$

for c=c_j for some j≠i, and

$$\Pr ob(c_i \neq c_j \text{ for all } j \neq i \mid \mathbf{c}_{-i}, \mathbf{x}_i) =$$

$$b\frac{n_{-i,c}}{T-1+\alpha}\ f_{mult}(\mathbf{x}_i \mid \mathbf{p})\, dD(\mathbf{p} \mid \beta,\ldots,\beta)$$

where, $n_{-i,c}$ is the number of observations classified in c, not counting the ith observation, $\mathbf{c}_{-i}$ is the classification vector for all observation except ith, $H_{-i,c}$ is the posterior distribution of $\mathbf{p}$, based on the prior Dirichlet distribution given in (3) and all observations $\mathbf{x}_j$ for which j≠i and c_j=c and b is the normalizing constant. The Gibbs sampler for sampling from the posterior distribution of the classification vector consists of drawing individual classification variables according to these conditional probabilities. Although the number of components is theoretically infinite, the maximum number of nonempty components is T. After each step, components with zero associated observations are removed from the list of current components. New components are added to the list whenever a $c_i \neq c_j$ for all i≠j is drawn. If the total number of simulated assignment vectors, after "burn-in", is G, the T by T assignment matrix $\mathbf{Z}$ is created by setting $\mathbf{Z}$[i,j] to the number of simulated assignment vectors for which $c_i=c_j$. Clusters are then created by putting together observations that had equal assignments in more than 50% of generated assignment vectors.

Simulation Study

Previously we showed (Medvedovic, et al. 2000) that the finite mixture approach performed reasonably well when the components of the mixture generating data are well separated and the number of events in multinomial observations is relatively high. Now we applied the infinite mixture approach to some of the same simulated data sets and compared its performance with respect to the number of misclassified observations to the results we obtained by the finite mixture approach. The simulated data consisted of 10 data sets simulated from the model with three clusters under three different scenario. Each cluster had equal number of observations. Observations in a cluster were generated by a multinomial distribution with 10 possible outcomes.

Scenario 1.

Multinomial parameters for the three clusters represented well separated multinomial distributions were:

$\mathbf{p}_1$=(0.37, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07)

$\mathbf{p}_2$=(0.07, 0.37, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07)

$\mathbf{p}_3$=(0.07, 0.07, 0.37, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07)

Number of events in each multinomial observation was 20.

Scenario 2.

Multinomial parameters for the three clusters represented poorly separated multinomial distributions were:

$\mathbf{p}_1$=(0.28, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08)

$\mathbf{p}_2$=(0.08, 0.28, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08)

$\mathbf{p}_3$=(0.08, 0.08, 0.28, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08)

Number of events in each multinomial observation was 20.

Scenario 3.

Multinomial parameters for the three clusters represented poorly separated multinomial distributions were:

$\mathbf{p}_1$=(0.28, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08)

$\mathbf{p}_2$=(0.08, 0.28, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08)

$\mathbf{p}_3$=(0.08, 0.08, 0.28, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08)

Number of events in each multinomial observation was 10.

Results obtained by applying the infinite mixture approach are contrasted to the results obtained by the finite mixture approach assuming that the number of components is known (Table 1) in terms of the percentage of misclassified observations. For each simulated data set the Gibbs sampler described in the previous section was started from the classification vector corresponding that lumps all observations in a single cluster. 10,000 full iterations of the sampler were run with only last 9,000 (after 1,000 "burn-in" iterations) being used in the analysis. α and β parameters in both prior distributions were set to 1.

As it can be seen, the finite mixture model under the correct number of components performed better than the infinite model. This could hardly by a surprise considering the fact that the information about the number of components was not used in the infinite mixture approach. When the number of mixture components was estimated from the data, finite mixture component performed better only in Scenario 1. By manipulating the cut-off point for putting observations in the same cluster we were able to somewhat trim misclassification percentages (results not shown).

Generally, it is possible to hypothesize the number of components from the distribution of the number of different components in simulated classification vectors. These distributions for first data sets under each scenario are given in Figures 1,2 and 3.

Table 1

Percent of Misclassified Observations			
Scenario	Finite Mixture		Infinite Mixture
	Known Q	AIC	
1	4.7	4.7	10.7
2	22	30	28
3	40	53	56

Figure 1 (Scenario 1)

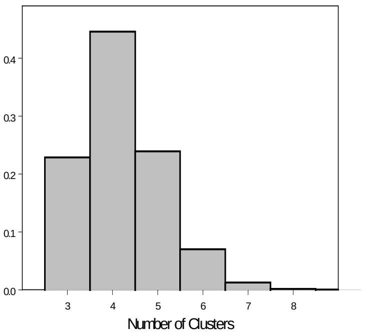


Figure 2 (Scenario 2)

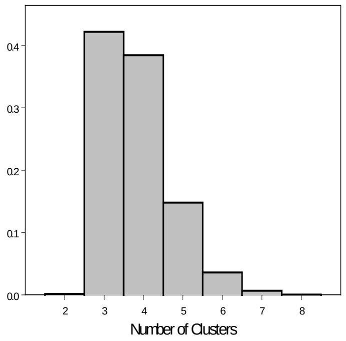
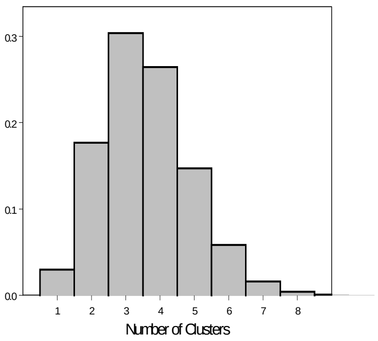


Figure 3 (Scenario 3)



Since the data in all three situations was simulated under a model with three clusters, simulations do not seem to be very useful in assessing the number of mixture components present in the data.

Conclusions

We showed that the clustering method based on the infinite mixture approach was able to moderately well recreate underlying structure of the data. This method performed slightly worse than the clustering procedure based on the finite mixture model that requires the number of clusters to be specified. It also did slightly worse than the procedure in which the number of mixture components was estimated using AIC criterion in the situation when the mixture components are well separated and a high number of events is observed in each multinomial observation. That is, in two out of three situations, uncertainties in the process of determining the number of components have undermined the slight advantage that the finite mixture model had over the infinite model approach. Furthermore, the process of maximizing likelihood in the traditional finite mixture approach is generally a difficult problem involving uncertainties about whether the global maximum has been identified. All these issues make the infinite mixture approach a valid alternative.

Further examinations of the sensitivity of the procedure with respect to parameters specifying prior distributions in the model is needed. A natural approach to this problem would be to consider parameters α and β to be random and adding another hierarchy to the model. Such an expended model has bee described in the context of Gaussian mixtures by Rasmussen, 2000. Also, finding better ways of transforming the list of clusterings generated by the Gibbs sampler into a single optimal clustering is likely to significantly improve performance of the infinite mixture approach to clustering.

Acknowledgement

This work has been supported by the NIEHS grant ES06096-09.

References

(1) Celleux, G., Govaert, G. (1992), “A Classification EM algorithm for clustering and two stochastic versions,” *Computational Statistics & Data Analysis*,14, 315-332.

(2) Dempster. A. P., Laird, N. M., Rubin, D. B. (1977), “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society, B*, 39, 1-38.

(3) Medvedovic, M., Succop, P., Shukla, R., Dixon, K. (2000), “Clustering Mutational Spectra via Classification Likelihood and Markov Chain Monte Carlo Algorithms,” *Journal of Agricultural, Biological, and Environmental Statistics*, To appear in the December Issue.

(4) Neal, R.M. (2000), “Markov Chain Sampling Methods for Dirichlet Process Mixture Models”, *Journal of Computational and Graphical Statistics*, 9, 249-265.

(5) Rasmussen, C.E. (2000), “The Infinite Gaussian Mixture Model”, *Advances in Neural Information Processing Systems 12,MIT Press,* 554-560.