A Bayesian Model-based Clustering with Dissimilarities

Man-Suk Oh
Ewha Womens University

Adrian Raftery
University of Washington

October 17, 2003

Abstract

A Bayesian model-based clustering method is proposed for clustering objects on the basis of dissimilarites. This combines two basic ideas. The first is that the objects have latent positions in a Euclidean space, and that the observed dissimilarities are measurements of the Euclidean distances with error. The second idea is that the latent positions are generated from a mixture of multivariate normal distributions, each one corresponding to a cluster. We estimate the resulting model in a Bayesian way using Markov chain Monte Carlo. The method carries out multidimensional scaling and model-based clustering simultaneously, and yields good object configurations and good clustering results with reasonable measures of clustering uncertainties. In the examples we studied, the clustering results based on low-dimensional configurations were almost as good as those based on high-dimensional ones. Thus the method can be used as a tool for dimension reduction when clustering high-dimensional objects, which may be useful especially for visual inspection of clusters.

We also propose a Bayesian criterion for choosing the dimension of the object configuration and the number of clusters simultaneously. This is easy to compute and works reasonably well in simulations and real examples.

1 Introduction

Cluster analysis is the automatic grouping of objects into groups on the basis of numerical data consisting of measures either of properties of the objects, or of the dissimilarities between them. It was developed initially in the 1950s (e.g. Sneath 1957; Sokal and Michener 1958), and the early development was driven by problems of biological taxonomy and market segmentation. More recently, clustering has attracted a great deal of attention as a useful

tool for grouping genes and samples in DNA microarray experiments, clustering documents on the World Wide Web and in other text databases, and grouping pixels in medical images so as to identify features of clinical interest.

Model-based clustering is a framework for putting cluster analysis on a principled statistical footing; for reviews see McLachlan and Peel (2000) and Fraley and Raftery (2002). It is based on probability models in which objects are assumed to follow a finite mixture of probability distributions such that each component distribution represents a cluster. The model-based approach has several advantages over heuristic clustering methods. First, it clusters objects and estimates component parameters simultaneously, avoiding well-known biases that exist when they are done separately. Second, it provides clustering uncertainties which is important especially for objects close to cluster boundaries. Third, the problems of determining the number of components and the component probability distributions can be recast as statistical model selection problems, for which principled solutions exist. Unlike the previously mentioned heuristic clustering algorithms, however, model-based clustering requires object coordinates rather than dissimilarities between objects as an input. Thus, despite the important advantages of model-based clustering, it can be used only when object coordinates are given, and not when dissimilarities are provided.

Even when object coordinates are given, visual display of clusters in low dimensional space is often desired since it may provide useful information about the relationships between the clusters and the underlying data generation process (Hedenfalk et al, 1999; Yin, 2002; Nikkila, 2002). One way to reduce the dimensionality of objects for visual display in lower dimensional space is multidimensional scaling (MDS). In MDS, objects are placed in a Euclidean space while preserving the distance between objects in the space as well as possible.

In this paper, we develop a model-based clustering method for dissimilarity data. We assume that an observed dissimilarity measure is equal to the Euclidean distance between the objects plus a normal measurement error. We model the unobserved object configuration as a realization of a mixture of multivariate normal distributions, each one of which corresponds to a different cluster. We carry out Bayesian inference for the resulting hierarchical model using Markov chain Monte Carlo (MCMC). The resulting method combines MDS and model-based clustering in a coherent framework.

Other important issues are the choice of the number of clusters and of the dimension of the objects. Oh and Raftery (2001) proposed an easily computed Bayesian criterion called MDSIC for choosing object dimension. We extend this to determine the number of clusters

as well. The resulting criterion can be computed easily from MCMC output.

2 Model for Clustering with Dissimilarities

Let δ_{ij} denote the dissimilarity measure between objects i and j, which is assumed to be functionally related to p unobserved attributes of the objects. Let $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ denote an unobserved vector representing the values of the attributes possessed by object i.

As in Oh and Raftery (2001), we model the true dissimilarity measure δ_{ij} as the distance between objects i and j in a Euclidean space, i.e., $\delta_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$. In practical situations, the true dissimilarity measure can be different from Euclidean distance and there can be measurement error in observations. We therefore assume that the observed dissimilarity measure, d_{ij}, is equal to the true measure, δ_{ij}, plus a Gaussian error. In addition, since dissimilarity measures are typically given as positive values we restrict the observed dissimilarity to be positive. Thus, given the Euclidean distance δ_{ij}, the observed dissimilarity measure d_{ij} is assumed to follow the truncated normal distribution $d_{ij} \sim N(\delta_{ij}, \sigma^2)\ I(d_{ij} > 0), \quad i \neq j, i, j = 1, ..., n$. Note that d_{ij} is related to $\mathbf{X} = \{\mathbf{x}_i\}$, called the object configuration, only through δ_{ij}. To represent clustering, we assume that the object configuration is a sample from a mixture of multivariate normal distributions,

$$\mathbf{x}_i \sim \sum_{j=1}^{G} \varepsilon_j\ N(\mu_j, T_j), \tag{1}$$

where each component normal distribution represents a cluster.

We use the following priors for the model parameters: $\sigma^2 \sim IG(a, b), \quad (\varepsilon_1, \ldots \varepsilon_g) \sim$ Dirichlet$(1, ..., 1)$,$\mu_j \sim N(\mu_{j0}, T_j)$,$T_j \sim IW(\alpha, B_j)$, where $IG(a, b)$ is the inverse Gamma distribution with mode $b/(a+1)$ and IW is the inverse Wishart distribution.

3 Posterior Inference

It is well known that inference for mixture models can be simplified with latent variables which indicate the group memberships of objects. We define latent variables K_i such that $P(K_i = j) = \varepsilon_j$ and $\mathbf{x}_i$ belongs to class j if $K_i = j$, so that $\mathbf{x}_i | K_i = j \quad \sim \quad N(\mu_j, T_j)$. From the prior and the model, the full conditional posterior distributions (densities)of each parameter given all the other unknowns are given in simple forms. Iterative generation of the unknown parameters from their full conditional distributions for a sufficiently long

time yields samples of the parameters from the joint posterior distribution, and posterior inference can be done by using the samples.

4 A Bayesian Selection Criterion for Configuration Dimension and the Number of Clusters

Posterior inference as described in the previous section presumed that the dimension, p, of the object configuration, and the number of clusters, G, are given. These are typically unknown, however, and we now propose a statistical method for choosing p and G. Oh and Raftery (2001) suggested a dimension selection criterion for MDS, called MDSIC, which works well with Euclidean distance measures with small or moderate error size. In this section, we extend MDSIC and propose a new Bayesian selection criterion, named MIC, for choosing both p and G simultaneously.

We view the overall goal of our analysis as being to choose the best object configuration across the dimension p and the number of clusters G. We therefore base our model selection criteria on $\pi(\mathbf{X}_{pG}, p, G|D)$, the posterior density function of $\mathbf{X}, p, G$, given data D at $\mathbf{X} = \mathbf{X}_{pG}$, where $\mathbf{X}_{pG}$ is the best object configuration given p and G.

We propose a selection criterion, which we call MIC, as follows. Let

$$
\begin{aligned}
MIC_{1G} &= (m-2)\log SSR_{1G} - 2\log \pi(\mathbf{X}_{1G}) \\
MIC_{pG} &= \sum_{q=1}^{p} -2\log \frac{\pi(\mathbf{X}_{qG}|D)}{\pi(\mathbf{X}^*_{qG}|D)} \qquad (2) \\
&= \sum_{q=1}^{p} -2\log \frac{l(\mathbf{X}_{qG}|D)}{l(\mathbf{X}_{q-1,G}|D)} \frac{\pi(\mathbf{X}_{qG})}{\pi(\mathbf{X}_{q-1,G})} - 2\log A_q \qquad (3) \\
&= (m-2)\log(SSR_{pG}) - 2\log \pi(\mathbf{X}_{pG}) - 2\sum_{q=1}^{p} \log A_q. \qquad (4)
\end{aligned}
$$

Note that $(m-2)\log(SSR_{pG})$ can be considered as a measure of fit, $-2\log\pi(\mathbf{X}_{pG})$ plays the role of a penalty for complexity, and $-2\sum_{q=1}^{p}\log A_q$ is a cumulative correction factor for the shrinking effect. The values of p and G that yield the minimum of MIC_{pG} are viewed as best.

5 Discussion

We have proposed a model-based clustering method for the situation where the data consist of dissimilarity measures between pairs of objects. It is also useful for clustering objects in low dimensional space for visual display and parsimony even when the object coordinates are given, but are high-dimensional.

When the estimated dimension is high, we have compared BMDS with the selected dimension with BMDS with low dimension (2 or 3). We found that the clustering results were very similar, and that those misclassified in the low-dimensional analysis had high clustering uncertainties, which is good. Thus, in practice BMDS low dimensional configurations may be good enough for many purposes, especially if it is followed up with more intensive investigation of objects with high clustering uncertainty.

We have proposed a Bayesian criterion, MIC, for simultaneously selecting the object dimension and the number of clusters, which is easy to compute from MCMC output. In our simulations and in real examples, it worked reasonably well in all cases. MIC varied more between dimensions than between numbers of clusters, and the choice of dimension was not affected by the choice of the number of clusters. Thus, as an approximation we suggest selecting the dimension assuming one cluster (i.e. using BMDS), and then choosing the number of clusters given the selected dimension. This greatly reduces computation time.

References

[1] Fraley, C. and Raftery, A.E. (2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

[2] Hedenfalk, I.A., Ringer, M., Trent, J., Borg, A. (1998). Gene Expression in Inherited Breast Cancer,

[3] Hoff, P., Raftery, A.E. and Handcock, M. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

[4] Nikkila, J., Toronen, P., Kaski, S., Venna, J., Castren, E., and Wong, G. (2002), Analysis and visualization of gene expression data using Self-Organizing Maps, *Neural Networks*, **15**, 953-966.

[5] Oh, M-S. and Raftery, A. (2001), Bayesian Multidimensional Scaling and Choice of Dimension, *Journal of the American Statistical Association*, **28**, 259-271.

[6] Sneath, P.H.A. (1957), "The Application of Computers to Taxonomy," *Journal of General Microbiology*, 17, 201–206.

[7] Sokal, R.R. and Michener, C.D. (1958), "A Statistical Method for Evaluating Systematic Relationships," *University of Kansas Scientific Bulletin*, 38, 1409–1438.

[8] Yin, H. (2002), Data visualization and manifold mapping using the ViSOM, *Neural Networks*, **15**, 1005-1016.