Recent Advances in Model-Based Clustering: Variable Selection and Social Networks

Adrian E. Raftery[1,2]

[1] Center for Statistics and the Social Sciences, University of Washington, Box 354320, Seattle, WA 98195-4320
[2] UTIA, Prague

Abstract. Cluster analysis is the automated search for groups of related observations in a dataset. Model-based clustering bases cluster analysis on a finite mixture probabilistic model, allowing inference to be put on a formal statistical basis. This leads to maximum likelihood and Bayesian estimation of the model parameters, assessment of uncertainty about group classifications, formal inference about the number of groups present and the best clustering models, as well as robust methods for dealing with outliers.

I will describe a method for deciding which variables should be used for clustering. This recasts the variable selection problem as a model selection one, leading to a solution based on approximate Bayes factors. In experiments, we found that removing irrelevant variables often improved performance, and led to more parsimonious clustering models and easier visualization of results.

I will also describe the application of model-based clustering to social network data. Network models describe ties among interacting units or actors. Network data often exhibit transitivity, meaning that two actors that have ties to a third actor are more likely to be tied than actors that do not, homophily by attributes, meaning that actors with similar values of variables are more likely to be tied, and clustering. Interest often focuses on finding clusters of actors or ties, and the number of groups in the data is typically unknown. We propose a new model, the Latent Position Cluster Model, under which the probability of a tie between two actors depends on the distance between them in an unobserved Euclidean "social space," and the actors' locations in the latent social space arise from a mixture of distributions, each one corresponding to a cluster. It models transitivity, homophily by attributes and clustering simultaneously, and does not require the number of clusters to be known. The model makes it easy to simulate realistic networks with clustering, potentially useful as inputs to models of more complex systems of which the network is part, such as epidemic models of infectious disease.

This is joint work with Nema Dean, Mark Handcock, Jeremy Tantrum and Chris Fraley.

Keywords