Subset Clustering of Binary Sequences with an Application to Genomic Abnormality Data

P.D. Hoff[✧]

University of Washington, Seattle WA USA

[✧]Email: *hoff@stat.washington.edu*

We discuss a model-based approach to identifying clusters of objects based on subsets of attributes, so that the attributes that distinguish a cluster from the rest of the population may depend on the cluster being considered. The method is based on a Polya urn cluster model for multivariate means and variances, resulting in a multivariate Dirichlet process mixture model. This particular model-based approach accommodates outliers, a variety of data types (continuous, count and binary) and allows for the incorporation of application-specific data features into the clustering scheme. For example, in an analysis of genetic CGH array data we are able to design a clustering method that accounts for spatial dependence of chromosomal abnormalities.