Bayesian hierarchical clustering: agglomerative clustering meets divisive clustering

Kevin J. Dawson[1] and Khalid Belkhir[2]

[1] Centre for Mathematical and Computational Biology, Rothamsted Research, UK
[2] Institut des Sciences de l'Evolution, CNRS, Montpellier, France

In a *clustering* problem we have a collection of individuals (the sample) which must be classified into mutually exclusive categories. In general we are uncertain about the number of categories, and their precise nature. The parameter of interest is the *sample partition*, - the partition of the label set of the sample induced by classifying the individuals correctly into mutually exclusive categories. From a Bayesian perspective we are interested in the posterior distribution of the sample partition, given all the available data. Various Monte Carlo methods are available for sampling from the posterior distribution of the sampling partition.
Since the size of the (discrete) parameter space $\mathcal{P}(S)$ (the set of all partitions of the label set S) grows very rapidly with the size $n = |S|$ of the sample, it is only in cases where n is small, that a probability distribution over the space $\mathcal{P}(S)$ can be visualised in its entirety. As a consequence, if many of the individuals in the sample are difficult to classify, then there will be many plausible partitions, each one having an individually low posterior probability. So, regions where the probability is elevated will be difficult to identify. This visualisation problem has received surprisingly little attention.
Here we approach the problem of visualisation of a (posterior) probability distribution on $\mathcal{P}(S)$, by constructing a sequence of nested credible sets on $\mathcal{P}(S)$, - always of some special type which can be readily comprehended and visualised. We show how simple hierarchical clustering algorithms arise naturally as procedures for constructing these sequences of nested credible sets. In this way we establish a straightforward Bayesian interpretation for the hierarchical clustering algorithms.
These methods will be illustrated using clustering problems in population genetics, where the data consists of the genotypes of the individuals in the sample, at multiple genetic markers. Markov chain samplers are available for the Bayesian discovery of populations of origin, for individuals sampled from an outcrossing species [1, 2]. A Markov chain sampler is also available for the Bayesian discovery of selfing lines, among individuals sampled from a partially selfing species [3]. We will use our clustering algorithms to process the output from these Markov chain samplers.

References

[1] Pritchard JK, Stephens M, Donnelly PJ (2001) Inference of population structure using multilocus genotype data. Genetics 155:945-959

[2] Dawson KJ, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genetical Research 78:59-77

[3] Wilson IJ, Dawson KJ (2007) A Markov chain Monte Carlo strategy for sampling from the joint posterior dsitribution of pedigrees and popualtion parameters under a Fisher-Wright model with partial selfing. Theoretical Population Biology 72:436-458