Measurement of quality in cluster analysis

Christian Hennig
Department of Statistical Science, University College London, UK
c.hennig@ucl.ac.uk

There is much work on benchmarking is supervised classification, where "quality" can generally be measured as a function of misclassification probabilities. In unsupervised classification (cluster analysis), the measurement of quality is much more problematic, because in reality there is no "true" class label which can be used for cross-validation and the like. Furthermore, there is no guarantee that in situations where there is a true classification (for example, where benchmark data sets from supervised classification are used to assess clustering methods, or where data is simulated from a mixture distribution), this classification is unique. There can be a number of different reasonable clusterings of the same data, depending on the research aim.

I will discuss the use of statistics for the assessment of clustering quality that can be computed from classified data without making reference to "the true clusters". Such statistics have traditionally been called "cluster validation indexes" (such as the average silhouette width), and sometimes been used for estimating the number of clusters. Most of the traditional statistics try to balance various aspects of a clustering against each other (such as within-cluster homogeneity and between-cluster separation), but in order to characterize what advantages and disadvantages a clustering has, it is useful to formalize different aspects of cluster quality separately. This can also be used to explain misclassification rates in cases where "true" clusterings exist as function of the features of these clusterings.

Key Words: benchmarking, cluster validity, misclassification rate, homogeneity, separation, stability