Summary 7: Bayesian Hierarchical Clustering

Susan Cherry

This paper presents a new algorithm for agglomerative hierarchical clustering with the following advantages over traditional distance-based agglomerative clustering algorithms. 1) It presents a probabilistic model of data that can be used to compute the predictive distribution of a test point and the probability that a test point belongs to any existing clusters in the tree, 2) it uses model-based criterion to decide whether to merge existing clusters, 3) it uses bayesian hypothesis testing to decide whether merges are advantageous, 4) it can be interpreted as a fast bottom-up approximate inference method for a Dirichlet process.

First, the paper presents the Bayesian hierarchical clustering algorithm, which is similar to the traditional agglomerative clustering model. The main difference is that the algorithm presented in the paper uses statistical hypothesis test to choose which clusters to merge. When considering a merge, the following hypotheses are compared: The first hypothesis is that all the data in the union of two clusters were generated iid from the same probabilistic model. The alternative hypothesis is that the union of two clusters has or more clusters. The algorithm is quite simple and is outlined in Section 2 of the paper. The algorithm has several desirable qualities that the traditional algorithm does not have. It allows us to define predictive distributions for new data points, decides which merges are good, suggests natural places to cut the tree using statistical model comparisons, and can be customized to different sets by choosing appropriate models for the mixture components.

Next the paper discusses approximate inference for a Dirichlet Process Mixture model. The algorithm described is an approximate inference method for DPMs. The algorithm is a fast and deterministic alternative to MCMC approximation, as it computes the sum over exponentially many tree-consistent partitions for a tree that is built greedily bottom-up. The paper presents the marginal likelihood of a DPM and presents the proposition that the number of tree-consistent partitions is exponential in the number of data points for balanced binary trees.

Then the paper discusses learning and prediction. For a given setting of hyperparameters, the root node of the tree approximates the probability of the data given hyperparameters. We can use the root node marginal likelihood to compare different settings of hyperparameters. The paper also briefly talks about the predictive distribution, saying that the probability of a new test point given the data can be computed by recursing through the tree starting at the root node. Finally the paper talks about results comparing Bayesian hierarchical clustering to traditional clustering on several data sets. BHC nearly always finds highest purity trees. BHC also tends to create hierarchies with good structure.

This paper's algorithm is different from the related work because it is not a hierarchical generative model of data but is a hierarchical way of organizing nested clusters. It is also derived from Dirichlet Process mixtures and uses the hypothesis test to decide about merges. Finally, it doesn't uses iterative methods so is faster than most of the previous algorithms. However there are some limitations, including its greediness, computational complexity, and lack of incorporation of tree uncertainty.