Image segmentation methods and the use of pairwise similarities in data mining and pattern recognition

Dorit Hochbaum, UC Berkeley

Image segmentation methods often rely on the use of pairwise similarities, typically computed for adjacent pixels. We study here an algorithm that is a polynomial time variant of normalized cut, HNC (Hochbaum's Normalized Cut) and demonstrate its effectiveness for image segmentation tasks as well as for other classification tasks, including as a general machine learning technique. We show that the use of priors (or given label with various degrees of certainty) along with the use of pairwise similarities enhances performance in classification tasks for general data mining. It is demonstrated, via an extensive empirical study, that incorporating the use of pairwise similarities improves accuracy of classification and clustering. However, a drawback of the use of similarities is the quadratic rate of growth in the size of the data. A methodology we devised, called ""Sparse Computation"", is to address and eliminate this quadratic growth by mimicking the neighborhoods of a p-dimensional image. The technique of Sparse Computation is shown to enable the scalability of similarity-based algorithms to very large-scale data sets while maintaining high levels of accuracy. Several applications of variants of HNC for data mining, medical imaging, and image segmentation tasks, including a recent one in which HNC is among the top performing methods in a benchmark for cell identification in calcium imaging movies for neuroscience brain research.