58d Hierarchical K-Means Clustering Using Principal Components to Solve the Unsupervised Multi-Class Classification Problem

Syed B. Mohiddin, James Rathman, and Chihae Yang

Current clustering techniques can be grouped as either supervised or unsupervised. In a supervised method, each observation in the training dataset is pre-assigned to a class based on prior knowledge, while an unsupervised method uses no prior knowledge of class distinction. Numerous supervised techniques have been demonstrated to work well for binary classification and a few of these are reasonably good at making supervised multi-class predictions. However, techniques for unsupervised binary and multi-class predictions have not been fully developed. In this work, we present an analysis technique based on hierarchical K-means using differentially weighted principal component analysis to address unsupervised classification for both binary and multi-class problems. Application of this methodology to biological datasets (e.g., microarray gene expression data) has already been demonstrated and is extended to chemical datasets in this work with the objectives of predicting class membership and identifying non-redundant features most responsible for differentiating the observed classes.