

Scaling Soft Clustering to Very Large Data Sets

Lawrence O. Hall
Department of Computer Science and Engineering, ENB118
University of South Florida
4202 E. Fowler Ave.
Tampa, FL 33620-9951
hall@cse.usf.edu

There are an increasing number of large unlabeled data sets available. Some of these may have billions of examples or feature vectors. Partitioning such data sets into groups can be done by clustering algorithms. However, classical clustering algorithms do not scale well to tens of thousands of examples much less millions or billions. This talk introduces algorithms that can scale to arbitrarily large data sets. They can be used for data that flows as a stream or for online clustering. We show adaptations that are based on fuzzy c-means, possibilistic c-means and the Gustafson-Kessel clustering algorithms. Approaches applied to the EM algorithm and k-means are also covered. Results on real data sets showing that it is possible to obtain a final data partition which is almost identical to that obtained with the original algorithms on data that will fit in the memory of one computer are given.