

A Clustering Algorithm based on Graph Connectivity

Sukhendu Chakraborty

31 May, 2004

Abstract

Cluster analysis is a fundamental problem in experimental science, where one wishes to classify observations into groups or categories. It has applications in biology (e.g. gene expression), medicine, economics, psychology etc. Given M data points, $X = \{x_1, x_2, \dots, x_M\}$, the objective of clustering is to partition the data set into K non-empty subsets such that *alike* data are grouped together and data in different subsets or clusters are not alike.

Many graph theoretic techniques have been proposed for cluster analysis. Commonly known techniques include single-link and complete-link hierarchical algorithms formulated and implemented using a threshold graph forming clusters by breaking inconsistent arcs in the minimum spanning tree of the proximity graph or graphs constructed based on limited neighborhood sets. I would like to present a new clustering algorithm based on graph theoretic approach developed by Hartuv and Shamir[1]. The similarity data is used to form a *similarity graph* in which vertices correspond to elements and edges connect elements with values above some threshold. In that graph, clusters are highly connected subgraphs. Using minimum cut algorithms such subgraphs can be computed efficiently. It will also be shown that the solution produced by the algorithm possesses several properties that are desirable for clustering.

1 References

1. Erez Hartuv and Ron Shamir. *A Clustering algorithm based on Graph Connectivity*. Information Processing Letters, 76(4-6):175-181, 2000.