Multiple Heterogeneous Data Sources Integration for Discovery of Multiple Clustering Solutions

Yale Chang, Jennifer Dy

February 13, 2014

1 Abstract

Dataset collected from multiple heterogeneous sources usually contains different data types and a complex dataset could be multi-faced by nature. Most existing clustering algorithms, however, can only find a single solution either by maximizing the clustering quality given one source or minimizing the disagreement between different solutions given multiple sources. Moreover, existing work on exploring multiple clustering solutions can only use one single source as input. We introduce a novel approach that provides multiple clustering solutions by integrating information from multiple heterogeneous data sources through multiple kernel learning. The objective is to preserve the global structure of data similarity matrix in dimensionality reduction as well as finding good and novel clustering solutions. The data similarity matrix is constructed from convex linear combination of the similarity matrix of each source. The novelty of a new clustering solution is measured by Hilbert-Schmidt independence criterion. The algorithm can automatically find the weight of each source for every novel clustering solution. Experiment results show the effectiveness of our algorithm.