Bi-Clustering of Bipartite Graphs - a New Objective Function and Complexity Analysis

Shai Ben-David and Sharon Wulff
School of Computer Science
University of Waterloo
shai@cs.uwaterloo.ca
www.cs.uwaterloo.cs/∼ shai

Common clustering tasks get as input a data set with some similarity (or distance) function over it, and aim to find a partitioning of the data into groups mutually similar elements. Bi-clustering is a variant of this general task, in which the input data comes from two domain sets, and instead of having a distance over its elements the input contains some relation over the cartesian product of these sets. For example, a set of documents and a set of words and the relationship indicating the membership of words in the documents. In this setting, the clustering task is to find partitions of each of these domain sets, so that the relation values in each of the resulting blocks (i.e., a product of two groups, one from each domain set)are as homogeneous as possible.

Bi-clustering is quite common in practice. Geneticists apply bi-clustering to detect groups of similar genes based on the gene expression matrix of their levels of interactions over a set of treatments. In recommender systems people apply bi-clustering to determine groups of similar customers based on the matrix of their preferences over some set of products. Bi-clustering is applied in text categorization and in many other diverse real life data mining settings.

Naturally, there exist a host of bi-clustering algorithms [], [], []. Yet, somewhat surprisingly, there is very little by way of a formal definition of bi-clustering tasks and, as far as we know, in terms analysis of their complexity, there are some NP-hardness results but no positive approximation guarantees for bi-clustering algorithms.

In this work we formalize a new natural objective (or cost) function for bi-clustering. Our objective function is suitable for detecting clusters based on binary valued input matrices (or, equivalently, for clustering bipartite (undirected) graphs based on the edge structure). We analyze the complexity of the resulting optimization problems, showing that on one hand finding optimal solutions is NP-hard. On the other hand, we introduce an approximation algorithm, prove that it is a polytime approximation scheme for this bi-clustering

task and demonstrate its effectiveness. We also show that bi-clustering with our objective function can be viewed as a generalization of correlation clustering.