A two level co-clustering algorithm for very large data sets

Bartcus Marius, Boullé Marc, Clérot Fabrice

Orange Labs
prenom.nom@orange.com

Abstract. Co-clustering is a data mining technique that aims at identifying the underlying structure between the rows and the columns of a data matrix in the form of homogeneous blocks. It has many real world applications, however many current co-clustering algorithms are not suited on large data sets. One of the successfully used approach to co-cluster large data sets is the MODL co-clustering method that optimizes a criterion based on a regularized likelihood. However, difficulties are encountered with huge data sets. In this paper, we present a new two-level co-clustering algorithm, given the MODL criterion allowing to efficiently deal with very large data sets that does not fit in memory. Our experiments, on both simulated and real world data, show that the proposed approach dramatically reduces the computation time without significantly decreasing the quality of the co-clustering solution.

1 Introduction

Co-clustering (Hartigan, 1972), also named block clustering (Govaert and Nadif, 2008) or two-mode clustering (Mechelen et al., 2004) is a data mining technique. It aims at identifying the underlying structure between the rows and the columns of a data matrix in the form of homogeneous blocks. Whereas, the principle of standard clustering is to group similar individuals (observations) with respect to a set of features, the task of co-clustering is to simultaneously group similar individuals with respect to variables and similar variables with respect to observations, thus extracting the correspondence structure between the objects and features. Another advantage of co-clustering over standard clustering techniques is its matrix reduction capacity, where a large data table can be reduced into a significantly smaller one yet having the same structure as the original matrix. Indeed, this technique finds its use in many applications like in telecommunications (Guigourès et al., 2015), text mining (Dhillon et al., 2003; Li and Abe, 1998), graph mining (Guigourès et al., 2015), etc.

Several co-clustering approaches have been proposed in the literature (Bock, 1979; Dhillon et al., 2003; Govaert and Nadif, 2008). These methods differ mainly according to the type of analyzed data (categorical or numerical), the underlying hypothesis, the extraction method and the expected results. Several families of approaches have then been proposed to perform co-clustering. Govaert and Nadif (2013, 2008) investigated probabilistic models with use of latent variables in mixture models. Difficulties arise on initialization, large number of parameters to estimate and computational efficiency, therefore large data are hard to manage. Indeed,