A Framework for Simultaneous Co-clustering and Modeling of Dyadic Data

Meghana Deodhar and Joydeep Ghosh

Department of Electrical and Computer Engg.
University of Texas at Austin, Austin, TX, USA.
deodhar,ghosh@ece.utexas.edu

Abstract. For difficult classification or regression problems, practitioners often segment the data into relatively homogenous groups and then build a model for each group. This two-step procedure usually results in simpler, more interpretable and actionable models without any loss in accuracy. We consider problems such as predicting customer behavior across products, where the independent variables can be naturally partitioned into two groups. A pivoting operation can now result in the dependent variable showing up as entries in a "customer by product" data matrix. We present a model-based co-clustering (meta)-algorithm that interleaves clustering and construction of prediction models to iteratively improve both cluster assignment and fit of the models. This algorithm provably converges to a local minimum of a suitable cost function. The framework not only generalizes co-clustering and collaborative filtering to model-based co-clustering, but can also be viewed as simultaneous co-segmentation and classification or regression, which is better than independently clustering the data first and then building models. Moreover, it applies to a wide range of bi-modal or multimodal data, and can be easily specialized to address classification and regression problems. We demonstrate the effectiveness of our approach on both these problems through experimentation on real and synthetic data. Further, we realize that in several datasets not all the data is useful for the learning problem and ignoring outliers and non-informative values while learning models may lead to better models. We explore extensions of the simultaneous co-clustering and learning framework to automatically identify and discard irrelevant data points and features while modeling in order to improve prediction accuracy.

Topic: data mining
Preference: oral/poster