Nonparametric Bayesian Methods for Relational Clustering

Pu Wang, Kathryn B. Laskey and Carlotta Domeniconi
George Mason University
pwang7@gmu.edu, klaskey@gmu.edu, carlotta@cs.gmu.edu

An important task in data mining is to identify natural clusters in data. Relational clustering [1], also known as co-clustering for dyadic data, uses information about related objects to help identify the cluster to which an object belongs. For example, words can be used to help cluster documents in which the words occur; conversely, documents can be used to help cluster the words occurring in them. Algorithms for co-clustering have become popular in data mining (cf., [2], [3]). Empirical evaluations have shown improvement in both model fit and predictive performance from leveraging relational information in clustering.

Bayesian methods for co-clustering (*e.g.*, [3], [4]) have advantages such as the ability to represent uncertainty about cluster assignments, theoretical soundness, and natural protection against overfitting [5]. Several studies demonstrate that these advantages can translate into performance improvements ([3], [4]).

Figure 1, taken from [6], shows a latent Dirichlet Bayesian co-clustering model for dyadic data. In this model, each entry x_{rc} of the matrix is sampled from a mixture of multinomial distributions, with mixture components indexed by latent row cluster and column cluster indicator variables z_{1_r} and z_{2_c}. The Dirichlet parameters π_{1_r} and π_{2_c} represent membership probabilities for row and column clusters. The common distribution for cluster indicators of objects in the same row [column] introduces dependence between entries in the same row [column]. Given observations x_{rc}, the row and column clusters z_{1_r} and z_{2_c} are correlated: thus, information about row objects influences the cluster assignments for column objects, and vice versa.

Exact inference in Bayesian models of any complexity is typically intractable. Inference commonly requires a tradeoff between accuracy and computational complexity. Shan and Banerjee [3] applied a variational approach to approximate posterior distributions for a model similar to Figure 1. Wang *et al.* [6] noted that the parameters $\vec{\pi}_{1_r}$, $\vec{\pi}_{2_c}$, and $\vec{\theta}$ can be marginalized out of the conditional distribution for the latent variables given their Markov blankets, giving rise to a collapsed variational Bayesian method that proved more accurate in empirical evaluation than standard variational inference. Wang *et al.* also considered a collapsed Gibbs sampling inference method, which was more accurate but slower than collapsed variational Bayes.

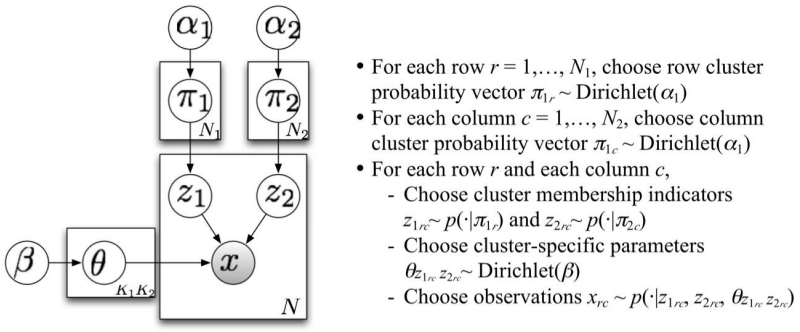

Figure 1: Generative Model for Latent Dirichlet Bayesian Co-Clustering

The model of Figure 1 is naturally extended in several directions. Extension to n-ary data is straightforward. Missing data can be treated via the EM algorithm or Gibbs sampling. The collapsed variational

Figure 2: Axis-Aligned Partitioning (left); Regular Grid Partitioning (right)

Bayes and collapsed Gibbs sampling algorithms can be applied to any conjugate pair of observation likelihood and parameter prior distributions. Information about attributes of the objects can easily be incorporated by adding additional variables into the model [1]. For example, consider a problem in which rows index customers, columns index products, and entries indicate customer satisfaction ratings. We might introduce a variable for a customer attribute such as gender; an arc from gender to z_{1_r} indicates that gender influences the cluster to which a customer belongs. Similarly, we can model multi-relational data, as well as co-clustering ensembles, for which a consensus co-clustering is formed from a set of base co-clusterings [7].

A nonparametric extension to the model of Figure 1 places no *a priori* bound on the number of latent row and column clusters. Replacing the Dirichlet prior distribution on the latent cluster indicators with a Dirichlet process (DP) allows the number of clusters to be learned from data [7]. Further, we can relax the assumption that the number and composition of row clusters is independent of the column object. For example, some movies might have the same distribution of ratings by all patrons; other movies might be rated differently by persons with different genre preferences. Figure 2 compares a regular grid partition with a hierarchical tree-style partitioning generated by the Mondrian Process (MP) [8]. We present experiments showing that the additional parsimony of the MP provides consistent performance improvements over a nonparametric grid partition model with independent DP priors on row and column clusters. We also show improvement of nonparametric co-clustering ensembles over individual co-clusterings.

References

[1] Ben Taskar. Eran Segal and Daphne Koller. Probabilistic Classication and Clustering in Relational Data *Proceedings of the International Joint Conference on Artificial Intelligence*, 2001.

[2] Inderit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha, Information-Theoretic Co-Clustering. *ACM SIGKDD international conference on Knowledge discovery and data mining*, 8998, 2003.

[3] Hanhuai Shan and Arindam Banerjee. Bayesian Co-clustering. *IEEE International Conference on Data Mining*, 2008.

[4] M. Mahdi Shafiei, and Evangelos E. Milios. Latent Dirichlet Co-Clustering. *International Conference on Data Mining*, 542551, 2006.

[5] William H. Jeffereys and James O. Berger, Ockham's Razor and Bayesian Statistics. *American Scientist* 80: 6472. Preprint available at http://quasar.as.utexas.edu/papers/ockham.pdf.

[6] Pu Wang, Carlotta Domeniconi, and Kathryn B. Laskey. Latent Dirichlet Bayesian Co-Clustering, in *Proceedings of the European Conference on Machine Learning and Principles and Practise of Knowledge Discovery in Databases*, Bled, Slovenia, September 7-11, 2009.

[7] Pu Wang, Carlotta Domeniconi, and Kathryn B. Laskey, Nonparametric Bayesian Co-clustering Ensembles, *Workshop on Nonparametric Bayes*, held in conjunction with NIPS, Whistler, BC, Canada, December 11-12, 2009.

[8] Daniel M. Roy and Yih W. Teh. The Mondrian Process. In *Advances in Neural Information Processing Systems (NIPS)*, volume 21, 2008.

Topic: learning algorithm
Preference: oral