A PAC-Bayesian Analysis of Co-clustering, Graph Clustering, and Pairwise Clustering

Yevgeny Seldin
Max Planck Institute for Biological Cybernetics
Tübingen, Germany
seldin@tuebingen.mpg.de

Abstract

We review briefly the PAC-Bayesian analysis of co-clustering (Seldin and Tishby, 2008, 2009, 2010), which provided generalization guarantees and regularization terms absent in the preceding formulations of this problem and achieved state-of-the-art prediction results in MovieLens collaborative filtering task. Inspired by this analysis we formulate weighted graph clustering[1] as a prediction problem: given a subset of edge weights we analyze the ability of graph clustering to predict the remaining edge weights. This formulation enables practical and theoretical comparison of different approaches to graph clustering as well as comparison of graph clustering with other possible ways to model the graph. Following the lines of (Seldin and Tishby, 2010) we derive PAC-Bayesian generalization bounds for graph clustering. The bounds show that graph clustering should optimize a trade-off between empirical data fit and the mutual information that clusters preserve on the graph nodes. A similar trade-off derived from information-theoretic considerations was already shown to produce state-of-the-art results in practice (Slonim et al., 2005; Yom-Tov and Slonim, 2009). This paper supports the empirical evidence by providing a better theoretical foundation, suggesting formal generalization guarantees, and offering a more accurate way to deal with finite sample issues.

1 Introduction

Co-clustering and graph clustering are important tools for analysis of social data with wide applications in recommender systems, social networks analysis, etc. As a result a multitude of different approaches to co-clustering and graph clustering were developed - see (Banerjee et al., 2007; Schaeffer, 2007; Yom-Tov and Slonim, 2009) and references therein. Comparing the different approaches is usually a painful task, mainly because the goal of each of these clustering methods is usually formulated in terms of the solution: most clustering methods start by defining some objective functional and then minimizing it. But for a given problem how can we choose whether to apply a graph cut method, spectral clustering, or an information-theoretic approach?

In this paper we formulate weighted graph clustering as a prediction problem[2]. Given a subset of edge weights we analyze the ability of graph clustering to predict the remaining edge weights. The philosophy behind this formulation is that if a model (not necessarily cluster-based) is able to predict with high precision all edge weights of a graph given a small subset of edge weights then it is a good model of the graph. The advantage of this formulation of graph modeling is that it is independent of a specific way chosen to model the graph and can be used to compare any two solutions, either

[1] Pairwise clustering is equivalent to clustering of a weighted graph, where edge weights correspond to pairwise distances. Hence, from this point on, we restrict the discussion to graph clustering.

[2] Unweighted graphs can be modeled by setting the weight of present edges as 1 and absent edges as 0.

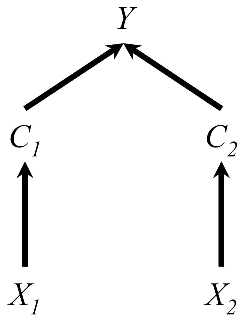

Figure 1: **Illustration of a graphical model corresponding to equation** (1) **for the case of** $d = 2$**.**

by comparison of generalization bounds or by cross-validation. The generalization bounds or cross-validation also address the finite-sample nature of the graph clustering problem and provide a clear criterion for model order selection.

The formulation and analysis of graph clustering presented here are based on the analysis of co-clustering suggested in (Seldin and Tishby, 2008, 2009, 2010), which we briefly review in the next section. Then we adapt the analysis to the graph clustering problem.

2 Review of PAC-Bayesian Analysis of Co-clustering

Co-clustering is a widely used method for analysis of data in the form of a matrix by simultaneous clustering of rows and columns of the matrix (Banerjee et al., 2007). In (Seldin et al., 2007; Seldin and Tishby, 2009) it was pointed out that there are actually two different classes of problems that are solved with co-clustering that should be analyzed separately. The first class of problems, which we are concerned with in this paper, are discriminative prediction tasks from which typical representative is collaborative filtering (Herlocker et al., 2004). The second class of problems, which is not considered in this paper, are those of estimation of a joint probability distribution in co-occurrence data analysis, such as the analysis of words-documents co-occurrence matrices in text mining, see (Seldin and Tishby, 2010) for further details. In collaborative filtering one is given a matrix of viewers by movies with ratings given by the viewers to the movies. The matrix is usually sparse and the task is to predict the missing entries. We assume that there is an unknown probability distribution $p(X_1, X_2, Y)$ over the triplets of viewer X_1, movie X_2, and rating Y. The goal is to build a discriminative predictor $q(Y|X_1, X_2)$ that given a viewer and movie pair will predict the expected rating Y. A natural form of evaluation of such predictors, no matter whether they are based on co-clustering or not, is to evaluate the expected loss $\mathbb{E}_{p(X_1,X_2,Y)}\mathbb{E}_{q(Y'|X_1,X_2)}l(Y, Y')$, where $l(Y, Y')$ is an externally provided loss function for predicting Y' instead of Y.

2.1 PAC-Bayesian Analysis of Discriminative Prediction with Co-clustering

Let $\mathcal{X}_1 \times .. \times \mathcal{X}_d \times \mathcal{Y}$ be a $(d+1)$-dimensional product space and assume that each $\mathcal{X}_i$ is categorical and its cardinality $|X_i|$ is fixed and known. We also assume that $\mathcal{Y}$ is finite with cardinality $|Y|$ and that the loss function $l(Y, Y')$ is bounded. In the collaborative filtering example $\mathcal{X}_1$ is the space of viewers, $\mathcal{X}_2$ is the space of movies, $d = 2$, and $\mathcal{Y}$ is the space of the ratings (e.g., on a five-star scale). The loss $l(Y, Y')$ can be, for example, an absolute loss $l(Y, Y') = |Y - Y'|$ or a quadratic loss $l(Y, Y') = (Y - Y')^2$.

We assume there exists an unknown probability distribution $p(X_1, .., X_d, Y)$ over $\mathcal{X}_1 \times .. \times \mathcal{X}_d \times \mathcal{Y}$ and that we are given an i.i.d. sample of size N from it. We use $\hat{p}(X_1, .., X_d, Y)$ to denote the empirical frequencies of $(d + 1)$-tuples $\langle X_1, .., X_d, Y\rangle$ in the sample. We consider the following form of discriminative predictors:

$$q(Y|X_1, .., X_d) = \sum_{C_1,..,C_d} q(Y|C_1, .., C_d) \prod_{i=1}^{d} q(C_i|X_i). \tag{1}$$

The hidden variables $C_1, .., C_d$ represent a clustering of $X_1, .., X_d$. The hidden variable C_i accepts values in $\{1, .., |C_i|\}$, where $|C_i|$ is the number of clusters used along dimension i. The free parameters of the model (1) are the conditional probability distributions $q(C_i|X_i)$ which represent the probability of assigning X_i to cluster C_i and the conditional probability $q(Y|C_1, .., C_d)$ which represents the probability of assigning label Y to cell $\langle C_1, .., C_d\rangle$ in the cluster product space. We denote the free parameters collectively by $\mathcal{Q} = \big\{\{q(C_i|X_i)\}_{i=1}^d, q(Y|C_1, .., C_d)\big\}$. Factorization (1) corresponds to the graphical model in Figure 1. We define:

$$L(\mathcal{Q}) = \mathbb{E}_{p(X_1,..,X_d,Y)}\mathbb{E}_{q(Y'|X_1,..,X_d)}l(Y, Y') \tag{2}$$

and

$$\hat{L}(\mathcal{Q}) = \mathbb{E}_{\hat{p}(X_1,..,X_d,Y)}\mathbb{E}_{q(Y'|X_1,..,X_d)}l(Y, Y'), \tag{3}$$

where $q(Y|X_1, .., X_d)$ is defined by (1). We define the mutual information $\bar{I}(X_i; C_i)$ corresponding to the joint distribution $\bar{q}(X_i; C_i) = \frac{1}{|X_i|}q(C_i|X_i)$ defined by $q(C_i|X_i)$ and a *uniform* distribution over X_i as:

$$\bar{I}(X_i; C_i) = \frac{1}{|X_i|}\sum_{x_i \in \mathcal{X}_i}\sum_{c_i=1}^{|C_i|} q(c_i|x_i)\ln\frac{q(c_i|x_i)}{\bar{q}(c_i)}, \tag{4}$$

where $\bar{q}(c_i) = \frac{1}{|X_i|}\sum_{x_i} q(c_i|x_i)$ is the marginal distribution over C_i. Finally, we denote the KL-divergence between two Bernoulli distributions with biases $\hat{L}(\mathcal{Q})$ and $L(\mathcal{Q})$ by

$$kl(\hat{L}(\mathcal{Q})\|L(\mathcal{Q})) = \hat{L}(\mathcal{Q})\ln\frac{\hat{L}(\mathcal{Q})}{L(\mathcal{Q})} + (1 - \hat{L}(\mathcal{Q}))\ln\frac{1 - \hat{L}(\mathcal{Q})}{1 - L(\mathcal{Q})}. \tag{5}$$

The following generalization bound for discriminative prediction with co-clustering was proved in (Seldin and Tishby, 2010).

Theorem 1. *For any probability measure $p(X_1, .., X_d, Y)$ over $\mathcal{X}_1 \times .. \times \mathcal{X}_d \times \mathcal{Y}$ and for any loss function l bounded by 1, with a probability of at least $1 - \delta$ over a selection of an i.i.d. sample S of size N according to p, for all randomized classifiers $\mathcal{Q} = \big\{\{q(C_i|X_i)\}_{i=1}^d, q(Y|C_1, .., C_d)\big\}$:*

$$kl(\hat{L}(\mathcal{Q})\|L(\mathcal{Q})) \le \frac{\sum_{i=1}^d \big(|X_i|\bar{I}(X_i; C_i) + |C_i|\ln|X_i|\big) + \Big(\prod_{i=1}^d |C_i|\Big)\ln|Y| + \frac{1}{2}\ln(4N) - \ln\delta}{N}. \tag{6}$$

In practice Seldin and Tishby (2010) replace the bound (6) with a parameterized trade-off

$$\mathcal{F}(\mathcal{Q}) = \beta N\hat{L}(\mathcal{Q}) + \sum_{i=1}^d n_i\bar{I}(X_i; C_i) \tag{7}$$

and suggest an alternating projection algorithm for finding a local minimum of $\mathcal{F}(\mathcal{Q})$ (for a fixed β). The value of β can be set by substitution of $\hat{L}(\mathcal{Q})$ and $\bar{I}(X_i; C_i)$ back into (6) or via cross-validation. Although the bound is shown to be reasonably tight, cross-validation is still better in practice. This algorithm achieves state-of-the-art performance on the MovieLens collaborative filtering dataset. Below we adapt this algorithm to solving graph clustering problems.

3 Formulation and Analysis of Graph Clustering

3.1 Graph Clustering as a Prediction Problem

Assume that $\mathcal{X}$ is a space of $|X|$ nodes and denote by w_{ij} the weight of an edge connecting nodes i and j.[3] We assume that the weights w_{ij} are generated according to an unknown probability distribution $p(W|X_1, X_2)$, where $X_1, X_2 \in \mathcal{X}$ are the edge endpoints. We further assume that we know $\mathcal{X}$ and are given a sample of size N from $p(X_1, X_2, W)$. The goal is to build a regression function $q(W|X_1, X_2)$ that will minimize the expected prediction error of the edge weights

$$\mathbb{E}_{p(X_1,X_2,W)}\mathbb{E}_{q(W'|X_1,X_2)}l(W, W') \tag{8}$$

for some externally given loss function $l(W, W')$ for approximating W with W'. Note that (8) does not assume any specific form of $q(W|X_1, X_2)$ and enables comparison of all possible approaches to this problem.

[3] All the results can be straightforwardly extended to hyper-graphs.

3.2 PAC-Bayesian Analysis of Graph Clustering

In this work we analyze the generalization abilities of $q(W|X_1, X_2)$ based on clustering:

$$q(W|X_1, X_2) = \sum_{C_1, C_2} q(W|C_1, C_2) q(C_1|X_1) q(C_2|X_2). \tag{9}$$

One can immediately see the relation between (9) and (1). The only difference is that in (9) the nodes X_1, X_2 belong to the same space of nodes $\mathcal{X}$ and the conditional distribution $q(C|X)$ is shared for mapping the endpoints of an edge. Let $\hat{p}(x_1, x_2, w)$ be the empirical distribution over edge weights, the empirical loss of a prediction strategy $\mathcal{Q} = \{q(C|X), q(W|C_1, C_2)\}$ corresponding to (9) can then be written as:

$$\begin{aligned}\hat{L}(\mathcal{Q}) &= \sum_{x_1, x_2, w} \hat{p}(x_1, x_2, w) \sum_{w'} q(w'|x_1, x_2) l(y, y') \\ &= \sum_{x_1, x_2, w} \hat{p}(x_1, x_2, w) \sum_{w', c_1, c_2} q(w'|c_1, c_2) q(c_1|x_1) q(c_2|x_2) l(w, w') \\ &= \sum_{w, w'} l(w, w') \sum_{c_1, c_2} q(w'|c_1, c_2) \sum_{x_1, x_2} q(c_1|x_1) \hat{p}(x_1, x_2, w) q(c_2|x_2). \end{aligned} \tag{10}$$

The following generalization bound for graph clustering can be proved by a minor adaptation of the proof of theorem 1.

Theorem 2. *For any probability measure $p(X_1, X_2, W)$ over the space of nodes and edge weights $\mathcal{X} \times \mathcal{X} \times \mathcal{W}$ and for any loss function l bounded by 1, with a probability of at least $1 - \delta$ over a selection of an i.i.d. sample S of size N according to p, for all graph clustering models defined by $\mathcal{Q} = \{q(C|X), q(W|C_1, C_2)\}$:*

$$kl(\hat{L}(\mathcal{Q}) \| L(\mathcal{Q})) \leq \frac{|X| \bar{I}(X; C) + |C| \ln |X| + |C|^2 \ln |W| + \frac{1}{2} \ln(4N) - \ln \delta}{N}, \tag{11}$$

where $|C|$ is the number of node clusters and $|W|$ is the number of different edge weights[4].

3.3 An Algorithm for Graph Clustering

Following the lines of (Seldin and Tishby, 2010) we replace (11) with a parameterized trade-off:

$$\mathcal{G}(\mathcal{Q}) = \beta N \hat{L}(\mathcal{Q}) + |X| \bar{I}(X; C) \tag{12}$$

and adapt Algorithm 1 from (Seldin and Tishby, 2010) to find local minima of (12) - see Algorithm 1 box below. The partial derivative $\frac{\partial \hat{L}(\mathcal{Q})}{\partial q(C|X)}$ in the algorithm accepts the form:

$$\frac{\partial \hat{L}(\mathcal{Q})}{\partial q(C|X)} = 2 \sum_{w, w'} l(w, w') \sum_{x_2, c_2} q(w'|C, c_2) \hat{p}(X, x_2, w) q(c_2|x_2). \tag{13}$$

4 Related Work

The regularization of pairwise clustering by mutual information $\bar{I}(X; C)$ was already applied by Slonim et al. (2005). In their work they maximized a parameterized trade-off $\langle s \rangle - T\bar{I}(X; C)$, where $\langle s \rangle$ measured average pairwise similarities within a cluster[5]. Their algorithm demonstrated superior results in cluster coherence compared to 18 other clustering methods. The regularization by mutual information was motivated by information-theoretic considerations inspired by the rate distortion theory (Cover and Thomas, 1991). Namely, the authors drew a parallel between $\langle s \rangle$ and distortion and $\bar{I}(X; C)$ and compression rate of the clustering algorithm. Further, Yom-Tov and Slonim (2009) showed that the algorithm performs fairly well even when it is presented only a subset of pairwise relations.

[4] The limitation of working with a fixed set of allowed edge weights can be resolved by weight quantization, or likely in a more elegant way in future work.

[5] The loss $L(\mathcal{Q})$ is slightly more general than $\langle s \rangle$ since it also considers edges between the clusters.

Algorithm 1 Algorithm for alternating projection minimization of $\mathcal{G}(\mathcal{Q}) = \beta N \hat{L}(\mathcal{Q}) + |X| \bar{I}(X;C)$.

```
Input: $\hat{p}(x_1, x_2, w)$, $N$, $|X|$, $|C|$, $l(w, w')$, $|W|$, $\beta$.
Initialize: $q_0(C|X)$ and $q_0(W|C_1, C_2)$ randomly.
$t \leftarrow 0$
$\bar{q}_0(c) \leftarrow \frac{1}{|X|} \sum_x q_0(c|x)$
repeat
    $q_{t+1}(c|x) \leftarrow \bar{q}_t(c) e^{-\beta N \frac{\partial \hat{L}(\mathcal{Q}_t)}{\partial q(c|x)}}$
    $Z_{t+1}(x) \leftarrow \sum_c q_{t+1}(c|x)$
    $q_{t+1}(c|x) \leftarrow \frac{q_{t+1}(c|x)}{Z_{t+1}(x)}$
    $\bar{q}_{t+1}(c) \leftarrow \frac{1}{|X|} \sum_x q_{t+1}(c|x)$
    $w^*_{t+1}(c_1, c_2) \leftarrow \arg\min_{w'} \sum_w l(w, w') \sum_{x_1, x_2} q_{t+1}(c_1|x_1) \hat{p}(x_1, x_2, w) q_{t+1}(c_2|x_2)$
    $q_{t+1}(w|c_1, c_2) \leftarrow \delta[w, w^*_{t+1}(c_1, c_2)]$
    $t \leftarrow t + 1$
until convergence
return $q_t(C|X), q_t(W|C_1, C_2)$ from the last iteration.
```

5 Discussion

We have formulated graph clustering as a prediction problem. This formulation enables direct comparison of graph clustering with any other approach to modeling the graph. By applying PAC-Bayesian analysis we have shown that graph clustering should optimize a trade-off between empirical fit of the observed graph and the mutual information that clusters preserve on the graph nodes. Prior work of Slonim et al. (2005) and Yom-Tov and Slonim (2009) underscores practical benefits of such regularization. Our formulation suggests a better founded and accurate way of dealing with the finite sample nature of the graph clustering problem and tuning the trade-off parameter β between model fit and model complexity in (12). It also suggests formal guarantees on the approximation quality. A more detailed experimental evaluation of the tightness of the analysis will be presented in future work.

References

Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Srujana Merugu, and Dhamendra Modha. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 8, 2007.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

Jonathan Herlocker, Joseph Konstan, Loren Terveen, and John Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1), 2004.

Satu Elisa Schaeffer. Graph clustering. *Computer Science Review*, 2007.

Yevgeny Seldin and Naftali Tishby. Multi-classification by categorical features via clustering. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

Yevgeny Seldin and Naftali Tishby. PAC-Bayesian generalization bound for density estimation with application to co-clustering. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

Yevgeny Seldin and Naftali Tishby. A PAC-Bayesian approach to unsupervised learning with application to co-clustering analysis. *Journal of Machine Learning Research*, 2010. Submitted. Preprint available at http://www.kyb.mpg.de/∼seldin.

Yevgeny Seldin, Noam Slonim, and Naftali Tishby. Information bottleneck for non co-occurrence data. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2007.

Noam Slonim, Gurinder Singh Atwal, Gasper Tracik, and William Bialek. Information-based clustering. *Proceedings of the National Academy of Science*, 102(51), 2005.

Elad Yom-Tov and Noam Slonim. Parallel pairwise clustering. In *SIAM International Conference on Data Mining (SDM)*, 2009.