Correlation Clustering: Maximizing Agreements via Semidefinite Programming

Chaitanya Swamy[*]

Abstract
We consider the Correlation Clustering problem introduced in [2]. Given a graph $G = (V, E)$ where each edge is labeled either "+" (similar) or "−" (different), we want to cluster the nodes so that the + edges lie within the clusters and the − edges lie between clusters. Specifically, we want to maximize *agreements* — the number of + edges within clusters and − edges between clusters. This problem is NP-Hard [2]. We give a 0.7666-approximation algorithm for maximizing agreements on any graph even when the edges have non-negative weights (along with labels) and we want to maximize the weight of agreements. These were posed as open problems in [2]. Previously the only results known were a trivial 0.5-approximation for arbitrary edge weighted graphs, and a PTAS with unit edge weights when $|E| = \Omega(|V|^2)$. Somewhat surprisingly, our algorithm always produces a clustering with *at most 6 clusters*. As a corollary we get a 0.7666-approximation algorithm for the k-clustering variant of the problem where we may create *at most k* clusters. A major component of this algorithm is a simple, easy-to-analyze algorithm that by itself achieves an approximation ratio of 0.75, opening at most 4 clusters.

Applications and Related Work. Bansal, Blum & Chawla [2] introduced the correlation clustering problem and motivated it by a document clustering application. We have a corpus of documents and each node in G represents a document in the corpus. An edge (u, v) with a + label means that the documents corresponding to nodes u, v are similar and a − label means that they are different. The goal is to cluster the documents so that similar documents (+ edges) lie in the same cluster and dissimilar documents (− edges) lie in different clusters. Correlation clustering can also be viewed as an agnostic learning problem. Each edge (u, v) is an "example" and we want to represent the target function f using a hypothesis class of vertex clusters. Our result implies that *any function f has a representation using a hypothesis of at most 6 clusters that is close to the best possible representation of f by a hypothesis from the class of all possible vertex clusterings.*

There is a trivial 0.5-approximation algorithm for maximizing agreements; putting all vertices in one big cluster, OR placing every vertex in a separate cluster, agrees with at least half the edge labels. Bansal et al. give an algorithm to approximate agreements with an *additive* error of $\epsilon|V|^2$, obtaining a PTAS when $|E| = \Omega(|V|^2)$. An equivalent optimization problem is to minimize *disagreements* — the number of − edges within clusters and + edges between clusters. In [2] a constant-factor approximation algorithm is given for minimizing disagreements on a complete unweighted graph. Recently, [3], [4] and [5] independently gave an $O(\log|V|)$-approx. algorithm for minimizing disagreements on weighted graphs. Independent of our work, [3] gives a 0.7664-approx. algorithm for maximizing agreements and shows that the problem is APX-Hard.

The k-clustering variant is a special case of the MAX-2-LIN-MOD-k problem [1]. Here we have a set of weighted linear equations and inequations of the form $y_i - y_j \equiv z \pmod{k}$, $y_i - y_j \not\equiv z \pmod{k}$. We want to set $y_i \in \{0, \ldots, k-1\}$ to maximize the total weight of the satisfied (in)equations. We get a 0.7666-approximation for the special case when we only have equations $y_i - y_j \equiv 0 \pmod{k}$ and inequations $y_i - y_j \not\equiv 0 \pmod{k}$. In contrast, for the general case only a $\left(\frac{1}{k} + \epsilon(k)\right)$-approximation is known due to Andersson, Engebretsen and Håstad [1].

Our Techniques. We consider a semidefinite programming relaxation of the problem and round its optimal solution. We consider two rounding procedures. The first one extends the Goemans-Williamson random-hyperplane rounding procedure for MAX CUT [7]. We choose 2 hyperplanes producing a solution with at most 4 clusters where the weight of agreements is, in expectation, at least 0.75 times the optimal solution value. The second procedure is based on the rounding scheme of [6]. We show that by randomly choosing one of these we obtain a clustering in which the total weight of agreements is at least 0.7666 times the optimal solution value. Thus choosing the better of the two rounding procedures gives a 0.7666-approximation algorithm.

1 Problem Description

We consider a somewhat more general version of correlation clustering to maximize agreements. Each edge e has two weights $w_{\mathsf{in}}(e), w_{\mathsf{out}}(e) \geq 0$. An edge e contributes $w_{\mathsf{in}}(e)$ to the total agreement weight if it lies within a cluster and $w_{\mathsf{out}}(e)$ otherwise. The problem is to find a clustering that maximizes $\sum_{e \text{ within cluster}} w_{\mathsf{in}}(e) + \sum_{e \text{ not in cluster}} w_{\mathsf{out}}(e)$. The weights $w_{\mathsf{in}}(e), w_{\mathsf{out}}(e)$ can be viewed as confidence estimates of whether e should be labeled + or − respectively, thus giving a soft labeling. The correlation clustering problem considered in [2] is a special case obtained by setting $w_{\mathsf{in}}(e) = 1$ if e is labeled + and 0 otherwise, $w_{\mathsf{out}}(e) = 1 - w_{\mathsf{in}}(e)$.

[*] swamy@cs.cornell.edu. Cornell University, Ithaca, NY 14853. Research supported by NSF grant CCR-9912422.

1.1 A Semidefinite Program.

Let $e_i \in \mathbb{R}^n$ be the vector with 1 in the i^{th} coordinate and 0s everywhere else. We can formulate the problem as the following mathematical program: maximize $\left\{\sum_{e=(u,v)}\bigl(w_{\mathsf{in}}(e)(x_u \cdot x_v) + w_{\mathsf{out}}(e)(1 - x_u \cdot x_v)\bigr) : x_v \in \{e_1, \dots, e_n\} \text{ for every } v \in V\right\}$. Vector e_i represents a possible cluster i. For any clustering, if we set $x_v = e_i$ for every vertex v assigned to cluster i, $i = 1, \dots, k$, the objective function value becomes the weight of agreements in the clustering. We relax the constraints $x_v \in \{e_1, \dots, e_n\}$ to get a semidefinite program.

$$\begin{aligned} \max \sum_{e=(u,v)} &\Bigl(w_{\mathsf{in}}(e)(x_u \cdot x_v) + w_{\mathsf{out}}(e)(1 - x_u \cdot x_v)\Bigr) && \text{(SP)} \\ \text{s.t.} \qquad x_v \cdot x_v &= 1 \qquad \text{for all } v \\ x_u \cdot x_v &\geq 0 \qquad \text{for all } u, v,\ u \neq v && (1.1) \end{aligned}$$

Our formulation resembles the MAX k-CUT relaxation in [6] but they relax a mathematical program involving k vectors $\{a_i\}$ s.t. $a_i \cdot a_i = 1$, $a_i \cdot a_j = \frac{-1}{k-1}$ for $i \neq j$.

2 The Algorithm

We solve (SP) and round the optimal solution. We consider two rounding procedures. Due to space limitations we only describe one of these which by itself gives an approximation ratio of 0.75, and sketch the improvements.

We extend the Goemans-Williamson rounding for MAX CUT by choosing multiple hyperplanes. Let $\{x_v \in \mathbb{R}^n\}$ be the optimal solution to (SP). While rounding we need to ensure that *both*, the probability that edge e lies inside a cluster, and the probability that e lies between clusters, are comparable to the coefficients of $w_{\mathsf{in}}(e)$ and $w_{\mathsf{out}}(e)$ respectively in the objective function. Choosing too many random hyperplanes rapidly decreases the probability of the former, and with too few hyperplanes, e.g., 1, the probability of the latter decreases to 0.5 times the coefficient of $w_{\mathsf{out}}(e)$. We choose 2 hyperplanes passing through the origin independently at random with normals distributed uniformly in the unit sphere. Let q_1, q_2 be the normals to the hyperplanes. These partition the vertices into 4 sets, some possibly empty, based on $x_v \cdot q_i$. Let $R_{s_1,s_2} = \{v : (-1)^{s_i} x_v \cdot q_i \geq 0,\ i = 1, 2\}$ where $s_i \in \{0, 1\}$. Each such non-empty set defines a cluster.

Analysis. Let $p_{\mathsf{in}}(\theta)$, $p_{\mathsf{out}}(\theta) = 1 - p_{\mathsf{in}}(\theta)$ denote the probabilities that nodes u and v with $x_u \cdot x_v = \cos\theta$ lie in the same cluster or different clusters respectively.

LEMMA 2.1. $p_{\mathsf{in}}(\theta) = (1 - \theta/\pi)^2$.

LEMMA 2.2. *For any* $\theta \in \left[0, \frac{\pi}{2}\right]$, $p_{\mathsf{in}}(\theta) \geq 0.75\cos\theta$ *and* $p_{\mathsf{out}}(\theta) \geq 0.75(1 - \cos\theta)$.

Proof. Let $f(\theta) = \frac{p_{\mathsf{in}}(\theta)}{\cos\theta}$ and $g(\theta) = \frac{p_{\mathsf{out}}(\theta)}{(1-\cos\theta)}$. $f(\theta)$ is minimized at the unique point $\vartheta \in \left[0, \frac{\pi}{2}\right]$ s.t. $\tan\theta = \frac{2}{\pi-\theta}$. So $\vartheta < 0.68288$ and $f(\theta) \geq \frac{(1-\vartheta/\pi)^2}{\cos\vartheta} = \frac{2(\pi-\vartheta)}{\pi^2 \sin\vartheta} > 0.7895$ for $\theta \in \left[0, \frac{\pi}{2}\right]$. $\frac{dg}{d\theta} = g(\theta)\bigl(\frac{1}{\theta} - \frac{1}{2\pi-\theta} - \cot(\theta/2)\bigr) \leq 0$ for $\theta \in \left[0, \frac{\pi}{2}\right]$ since $\cos\theta \geq 1 - \frac{\theta^2}{2}$, $\sin\theta \leq \theta$ for $\theta \in \left[0, \frac{\pi}{2}\right]$. So, $g(\theta) \geq g(\frac{\pi}{2}) = 0.75$ for $\theta \in \left[0, \frac{\pi}{2}\right]$.

THEOREM 2.1. *The above rounding procedure delivers a solution of expected value at least* $0.75 \cdot OPT$.

Proof. Let $\mathcal{C}$ be the clustering obtained by rounding. Let X_e be the contribution of edge $e = (u, v)$ to $\mathcal{C}$ and $\theta = \arccos(x_u \cdot x_v)$. $\mathrm{E}\bigl[X_e\bigr] = w_{\mathsf{in}}(e)p_{\mathsf{in}}(\theta) + w_{\mathsf{out}}(e)p_{\mathsf{out}}(\theta) \geq 0.75\bigl(w_{\mathsf{in}}(e)(x_u \cdot x_v) + w_{\mathsf{out}}(e)(1 - x_u \cdot x_v)\bigr)$ by Lemma 2.2, so $\mathrm{E}\bigl[\mathcal{C}\bigr] = \sum_e \mathrm{E}\bigl[X_e\bigr] \geq 0.75 \cdot OPT$.

2.1 Improvements.

For the other rounding procedure we adapt a rounding scheme in [6]. We choose 6 random vectors $r_1, \dots, r_6 \in \mathbb{R}^n$ whose coordinates have the standard normal distribution. Each r_i defines a (possibly empty) cluster $C_i = \{v : v \cdot r_i = \max_j v \cdot r_j\}$ in our clustering. The analysis of this algorithm is however significantly more involved. Randomly choosing this scheme or the 2-hyperplane rounding algorithm gives a 0.7666-approximation algorithm that produces at most 6 clusters. So this also works for the k-clustering variant when $k \geq 6$. For $k \leq 5$ we use a relaxation where (1.1) is replaced by $x_u \cdot x_v \geq \frac{-1}{k-1}$ and the objective function is $\max \sum_{e=(u,v)}\bigl(w_{\mathsf{in}}(e)\frac{1+(k-1)(x_u \cdot x_v)}{k} + w_{\mathsf{out}}(e)\frac{(k-1)(1-x_u \cdot x_v)}{k}\bigr)$, and round this by choosing either 1 or 2 hyperplanes. This achieves a ratio of 0.77.

THEOREM 2.2. *There is a 0.7666-approximation algorithm for maximizing agreements. This also gives a 0.7666-approx. algorithm for the k-clustering variant.*

Acknowledgments. I thank David Shmoys for giving helpful suggestions. I also thank Dotan Emanuel, Amos Fiat and Moses Charikar for sharing their results.

References

[1] G. Andersson, l. Engebretsen, and J. Håstad. A new approach to use semidefinite programming with applications to linear equations mod p. *J. Algorithms*, 39:162–204, 2001.

[2] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Proc. 43rd IEEE FOCS*, 238–247, 2002.

[3] M. Charikar, V. Guruswami and A. Wirth. Clustering with qualitative information. To appear in *Proc. 44th IEEE FOCS*, 2003.

[4] E. Demaine and N. Immorlica. Correlation clustering with partial information. *Proc. 6th APPROX*, 1–13, 2003.

[5] D. Emanuel and A. Fiat. Correlation clustering — minimizing disagreements on arbitrary weighted graphs. *Proc. 11th ESA*, 208–220, 2003.

[6] A. Frieze and M. Jerrum. Improved approximation algorithms for MAX k-CUT and MAX-BISECTION. *Algorithmica*, 18:67–81, 1997.

[7] M. Goemans and D. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *JACM*, 42:1115-1145, 1995.