Better Guarantees for Sparsest Cut Clustering

Maria-Florina Balcan

Microsoft Research, New England
One Memorial Drive, Cambridge, MA
mabalcan@microsoft.com

1 Introduction

The field of approximation algorithms for clustering is a very active one and a large number of algorithms have been developed for clustering objectives such as k-median, min-sum, and sparsest cut clustering. For most of these objectives, the approximation guarantees do not match the known hardness results, and much effort is spent on obtaining tighter approximation guarantees [1, 4, 5, 8, 6, 9, 10].

However, for many practical clustering problems such as clustering proteins by function, or clustering images by subject, there is some unknown correct "target" clustering; in such cases the pairwise information is merely based on some heuristics and the real goal is to achieve low error on the data. In these settings, the implicit hope is that approximately optimizing objective functions such as those mentioned above will in fact produce a clustering of low error, i.e., a clustering that is close pointwise to the truth. Formally, for a set of n data points the error of a clustering the error of a clustering $\mathcal{C}' = \{C'_1, ..., C'_k\}$ with respect to target clustering $\mathcal{C} = \{C_1, ..., C_k\}$ is the fraction of points on which $\mathcal{C}$ and $\mathcal{C}'$ disagree under the optimal matching of clusters in $\mathcal{C}$ to clusters in $\mathcal{C}'$, i.e.

$$err(\mathcal{C}) = \min_{\sigma \in S_k} \frac{1}{n} \sum_{i=1}^{k} |C_i - C'_{\sigma(i)}|,$$

where S_k is the set of bijections $\sigma : [k] \to [k]$.

Mathematically, the implicit assumption made when using a c-approximation algorithm for objective ϕ in order to get clusterings of low error, is that the clustering error of any c-approximation to ϕ on the data set is bounded by some ϵ. We will refer to this assumed property as the (c, ϵ) property for ϕ. From this perspective, a natural motivation for improving a c_2-approximation to a c_1-approximation (for $c_1 < c_2$) is that perhaps the data satisfies the (c_1, ϵ) property, but not the (c_2, ϵ) property. This is well justified, but perhaps one can do even better by using the (c, ϵ) property explicitly.

In recent work, Balcan, Blum, and Gupta [2] have shown that if we make this implicit assumption explicit, then one can get accurate clusterings even in cases where getting a good approximation to these objective functions is provably NP-hard. More specifically, [2] show that for any $c > 1$, if we assume that any c-approximation to the k-median or the k-means clustering objectives is ϵ-close to the target, then one can produce clusterings that are $O(\epsilon)$-close to the target, *even for values c for which obtaining a c-approximation is NP-hard.* Balcan and Braverman [3] have recently shown similar results for the min-sum objective [3]. This is perhaps even more interesting since this problem has a large gap between the best known approximation guarantees and the best known inapproximability results [3].

2 The Open Question

This line of work shows how for a clustering objective such as k-means or min-sum one can get much better results than those obtained so far in the approximation algorithms literature by wisely using implications of assumptions that were already being made implicitly. This note poses as an open question obtaining similar results for other natural classes of commonly used and studied clustering objective functions for which the best known approximation ratios are quite large. In particular, clustering techniques based on the sparsest cut objective and variations [1] have been extremely popular, so it would be interesting to analyze them in this framework.

Since it is cleanest to think about it, we will state the question for the case of two clusters, i.e., $k = 2$. We are given a weighted graph $G = (V, E)$ with positive edge weights w_e for every edge e (for learning motivated clustering applications these represent similarities between the two endpoints of the edge e). Given a cut or a partition of the graph into two pieces or clusters $(S, \bar{S})$, $S \cup \bar{S} = V$, we define the sparsity of the cut $(S, \bar{S})$ as

$$\phi(S) = \frac{w(S, \bar{S})}{|S| \cdot |\bar{S}|}.$$

The objective of sparsest cut problem [1, 9] is to find a cut $(S^*, \bar{S^*})$ which minimizes $\phi(S)$.

This objective and its variations have been long studied. Leighton and Rao [9] designed the first interesting approximation algorithm by giving an $O(\log n)$-approximation algorithm for this problem. They used a linear programming relaxation of the problem based on multicommodity flows. Later in a seminal paper Arora, Vazirani, and Rao [1] have developed a semidefinite programming relaxation based algorithm which provides an $O(\sqrt{\log n})$ approximation for the (uniform) sparsest cut problem. This $O(\sqrt{\log n})$ approximation is the best factor known for this problem. However, perhaps one can cluster as well *as if* one could approximate this objective to a constant factor. In particular, it would be interesting to see if one could cluster well in the Balcan et. al framework [2] under the (c, ϵ) property. That is, under the assumption that all the c-approximations to the sparsest cut objective are ϵ-close to the target, can one find a clustering that is $O(\epsilon)$-close to the target, even though no c-approximation to sparsest cut is known? Given that best known approximation for sparsest cut is $O(\sqrt{\log n})$ [1], a result for any constant c would be interesting in this framework.

Other interesting objectives to consider in this framework are the balanced cut objective [1] or the bicriteria objective of [7].

References

[1] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, 2004.

[2] M.-F. Balcan, A. Blum, and A. Gupta. Approximate clustering without the approximation. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2009.

[3] M.-F. Balcan and M. Braverman. Finding low error clusterings. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[4] Y. Bartal, M. Charikar, and D. Raz. Approximating min-sum k-clustering in metric spaces. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing*, 2001.

[5] M. Charikar, S. Guha, E. Tardos, and D. B. Shmoy. A constant-factor approximation algorithm for the k-median problem. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, 1999.

[6] K. Jain, M. Mahdian, and A. Saberi. A new greedy approach for facility location problems. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, 2002.

[7] R. Kannan, S. Vempala, and A. Vetta. On clusterings: good, bad and spectral. *JACM*, 51(3):497–515, 2004.

[8] A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, 2004.

[9] F. Leighton and S. Rao. Multicommodity max-flow mincut theorems and their use in designing approximation algorithms. *Journal of the ACM*, 46(6):787 – 832, 1999.

[10] L.J. Schulman. Clustering for edge-cost minimization. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, 2000.