Theoretical Computer Science 558 (2014) 51–61

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Data stability in clustering: A closer look

Shalev Ben-David [a], Lev Reyzin [b,*]

[a] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[b] Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607, USA

ARTICLE INFO

Available online 17 September 2014

Keywords:
Clustering
Data stability
Resilience assumptions

ABSTRACT

We consider the model introduced by Bilu and Linial (2010) [13], who study problems for which the optimal clustering does not change when distances are perturbed. They show that even when a problem is NP-hard, it is sometimes possible to obtain efficient algorithms for instances resilient to certain multiplicative perturbations, e.g. on the order of $O(\sqrt{n})$ for max-cut clustering. Awasthi et al. (2012) [6] consider center-based objectives, and Balcan and Liang (2012) [9] analyze the k-median and min-sum objectives, giving efficient algorithms for instances resilient to certain constant multiplicative perturbations. Here, we are motivated by the question of to what extent these assumptions can be relaxed while allowing for efficient algorithms. We show there is little room to improve these results by giving NP-hardness lower bounds for both the k-median and min-sum objectives. On the other hand, we show that constant multiplicative resilience parameters can be so strong as to make the clustering problem trivial, leaving only a narrow range of resilience parameters for which clustering is interesting. We also consider a model of additive perturbations and give a correspondence between additive and multiplicative notions of stability. Our results provide a close examination of the consequences of assuming stability in data.

1. Introduction

Clustering is one of the most widely-used techniques in statistical data analysis. The need to partition, or cluster, data into meaningful categories naturally arises in virtually every domain where data is abundant. Unfortunately, most of the natural clustering objectives, including k-median, k-means, and min-sum, are NP-hard to optimize [17,18]. It is, therefore, unsurprising that many of the clustering algorithms used in practice come with few guarantees.

Motivated by overcoming the hardness results, Bilu and Linial [13] consider a **perturbation resilience assumption** that they argue is often implicitly made when choosing a clustering objective: that the optimum clustering to the desired objective Φ is preserved under multiplicative perturbations up to a factor $\alpha > 1$ to the distances between the points. They reason that if the optimum clustering to an objective Φ is not resilient, as in, if small perturbations to the distances can cause the optimum to change, then Φ may have been the wrong objective to be optimizing in the first place. Bilu and Linial [13] show that for max-cut clustering, instances resilient to perturbations of $\alpha = O(\sqrt{n})$ have efficient algorithms for recovering the optimum itself.

* Corresponding author.
E-mail addresses: shalev@mit.edu (S. Ben-David), lreyzin@math.uic.edu (L. Reyzin).

http://dx.doi.org/10.1016/j.tcs.2014.09.025
0304-3975/

Continuing that line of research, Awasthi et al. [6] give a polynomial time algorithm that finds the optimum clustering for instances resilient to multiplicative perturbations of $\alpha = 3$ for center-based[1] clustering objectives when centers must come from the data (we call this the **proper** setting), and $\alpha = 2 + \sqrt{3}$ when the centers do not need to (we call this the **Steiner** setting). Their method relies on a **stability** property implied by perturbation resilience (see Section 2). For the Steiner case, they also prove an NP-hardness lower bound of $\alpha = 3$. Subsequently, Balcan and Liang [9] consider the proper setting and improve the constant past $\alpha = 3$ by giving a new polynomial time algorithm for the k-median objective for $\alpha = 1 + \sqrt{2} \approx 2.4$ stable instances.

1.1. Our results

Our work further delves into the proper setting, for which no lower bounds have previously been shown for the stability property. In Section 3 we show that even in the proper case, where the algorithm is restricted to choosing its centers from the data, for any $\epsilon > 0$, it is NP-hard to optimally cluster $(2 - \epsilon)$-stable instances, both for the k-**median** and **min-sum** objectives (Theorems 5 and 7). To prove this for the min-sum objective, we define a new notion of stability that is implied by perturbation resilience, a notion that may be of independent interest.

Then in Section 4, we look at the implications of assuming resilience or stability in the data, even for a constant perturbation parameter α. We show that for even fairly small constants, the data begins to have very strong structural properties, as to make the clustering task fairly trivial. When α exceeds $2 + \sqrt{3}$, the data begins to show what is called **strict separation**, where each point is closer to points in its own cluster than to points in other clusters (Theorem 8).

Finally, in Section 5, we look at whether the picture can be improved for clustering data that is stable under additive, rather than multiplicative, perturbations. One hope would be that **additive stability** is a more useful assumption, where a polynomial time algorithm for ϵ-stable instances might be possible. Unfortunately, this is not the case. We consider a natural additive model and show that severe lower bounds hold for the additive notion as well (Theorems 13 and 17). On the positive side, we show via reductions that algorithms for multiplicatively stable data also work for additively stable data for a different but related parameter.

Our results demonstrate that on the one hand, it is hard to improve the algorithms to work for low stability constants, and that on the other hand, higher stability constants can be quite strong, to the point of trivializing the problem. Furthermore, switching from a multiplicative to an additive stability assumption does not help to circumvent the hardness results, and perhaps makes matters worse. These results, taken together, narrow the range of interesting parameters for theoretical study and highlight the strong role that the choice of constant plays in stability assumptions.

One thing to note that there is some difference between the very related resilience and stability properties (see Section 2), stability being weaker and more general [6]. Some of our results apply to both notions, and some only to stability. This still leaves open the possibility of devising polynomial-time algorithms that, for a much smaller α, work on all the α-perturbation resilient instances, but not on all α-stable ones.

1.2. Previous work

We examine previous work on stability, both as a data dependent assumption in clustering and in other settings.

1.2.1. Stability as a data assumption in clustering

The classical approach in theoretical computer science to dealing with the worst-case NP-hardness of clustering has been to develop efficient approximation algorithms for the various clustering objectives [2,3,10,14,19,15], and significant efforts have been exerted to improve approximation ratios and to prove lower bounds. In particular, for metric k-median, the best known guarantee is a $(3 + \epsilon)$-approximation [3], and the best known lower bound is $(1 + 1/e)$-hardness of approximation [17,18]. For metric min-sum, the best known result is an $O(\text{polylog}(n))$-approximation to the optimum [10].

In contrast, a more recent direction of research has been to characterize under what conditions we can find a desirable clustering efficiently. Perturbation resilience/stability are such conditions, but they are related to other stability notions in clustering. Ostrovsky et al. [23] demonstrate the effectiveness of Lloyd-type algorithms [21] on instances with the stability property that the cost of the optimal k-means solution is small compared to the cost of the optimal $(k-1)$-means solution, and their guarantees have later been improved by Awasthi et al. [5].

In a different line of work, Balcan et al. [8] consider what stability properties of a similarity function, with respect to the ground truth clustering, are sufficient to cluster well. In a related direction, Balcan et al. [7] argue that, for a given objective Φ, approximation algorithms are most useful when the clusterings they produce are structurally close to the optimum originally sought in choosing to optimize Φ in the first place. They then show that, for many objectives, if one makes this assumption explicit – that all c-approximations to the objective yield a clustering that is ϵ-close to the optimum – then one can recover an ϵ-close clustering in polynomial time, even for values of c below the hardness of approximation constant. The assumptions and algorithms of Balcan et al. [7] have subsequently been carefully analyzed by Schalekamp et al. [24].

[1] For center-based clustering objectives, the clustering is defined by a choice of centers, and the objective is a function of the distances of the points to their closest center.