Characterizing α-Center Stable Clusterings

Zachary Frenette

August 12, 2016

1 Introduction

In recent years, there has been a lot of research done to try and understand why clustering appears "easy" in practice, even though it is NP-hard for many common objective functions. To address this mismatch between the perspective of a practitioner and a theorist, several different notions have been proposed. A common approach used to overcome the hardness of computational problems is to settle for an approximate solution as opposed to an optimal one. Another approach often used is to design an algorithm that solves the problem efficiently on some subset of "easy" instances. Both approaches have been extensively studied when it comes to clustering, but in this project, we will focus on the latter approach.

In a clustering instance, we are given a set X of n points in a finite metric space, a distance function $d : X \times X \to \mathbb{R}^+$, an objective function $\mathcal{O}$, and an integer $k \geq 1$. The goal is to partition X into a set of k clusters $\mathcal{C} = \{C_1, \dots C_k\}$ such that $\mathcal{O}(\mathcal{C}, d)$ is minimized. In this project, we will restrict our discussion to center-based clusterings. That is, solutions that can be defined by k points $\{c_1, \dots, c_k\}$ called centers such that every point is assigned to its closest center. Moreover, if the centers are allowed to be arbitrary points in the metric space, then we call them Steiner points.

When it comes to defining a set of "easy" clustering instances, many ideas have been explored, and the work of Ben-David provides an extensive survey of such ideas [4]. For our purposes, we will be interested in a notion called α-center stability. Introduced by Awasthi et al., an instance (X, d) is α-center stable with respect to $\mathcal{O}$ if, in any optimal clustering, points are closer by a factor of α to their own center than to any other center [1]. Formally, this property can be stated as follows.

Definition 1. *Given a set X of n points in a finite metric space, a distance function $d : X \times X \to \mathbb{R}^+$, an objective function $\mathcal{O}$, an integer $k \geq 1$, and a parameter $\alpha \geq 1$, we say that (X, d) is α-center stable with respect to $\mathcal{O}$ if, for any optimal k-clustering $\mathcal{C}^* = \{C_1^*, \dots, C_k^*\}$ with centers $\{c_1^*, \dots, c_k^*\}$, we have $\alpha d(p, c_i^*) < d(p, c_j^*)$ for all $1 \leq i \leq k$, $j \neq i$, and every $p \in C_i^*$.*

The remaining sections of this project are organized as follows. In section 2, we provide a brief overview of the known results regarding α-center stability. In section 3, we characterize the set of α-center stable clusterings, independent of any objective function. Finally, we state some concluding remarks and mention avenues for future work in section 4.

Table 1: Summary of Previous Results

Objective Function	**Steiner Points**	**Efficiency**	α	**Ref.**
center-based	Y	polynomial	$2+\sqrt{3}$	[1]
k-median	Y	NP-hard	$3-\epsilon$	
center-based	N	polynomial	$1+\sqrt{2}$	[3]
center-based	N	polynomial	2	[2]
k-median	N	NP-hard	$2-\epsilon$	[5]

2 Previous Work

There are several known results regarding α-center stability. If Steiner points are allowed, Awasthi, Blum and Sheffet showed that an optimal clustering can be found in polynomial time for instances that are $(2+\sqrt{3})$-center stable [1]. Furthermore, they show that for the k-median objective function, the problem of finding an optimal clustering in $(3-\epsilon)$-center stable instances is NP-hard.

On the other hand, when there are no Steiner points, Balcan and Liang showed that for $(1+\sqrt{2})$-center stable instances, an optimal clustering can be found in polynomial time [3]. These results were later improved by Bakshi and Chepurko, where they show that an optimal clustering can be found in polynomial time for instances that are 2-center stable [2]. Lastly, Ben-David and Reyzin showed that for the k-median objective function, the problem of finding an optimal clustering in $(2-\epsilon)$-center stable instances is NP-hard [5].

These results are summarized in Table 1.

3 Stability Without an Objective Function

In this section, we explore the notion of α-center stability, independent of any objective function. This is motivated by the fact that, in practice, users might prefer a "nice" and "meaningful" clustering rather than an optimal one. The following definition is analogous to Definition 1, without the dependence on an objective function. As a result, α-center stability becomes a property of a clustering rather than a property of the data.

Definition 2. *Given a set X of n points in a finite metric space, a distance function $d : X \times X \to \mathbb{R}^+$, an integer $k \geq 1$, and a parameter $\alpha \geq 1$, a k-clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ with centers $\{c_1, \ldots, c_k\}$ is α-center stable if, for all $1 \leq i \leq k$ and every $p \in C_i$, $\alpha d(p, c_i) < d(p, c_j)$ for all $j \neq i$.*

In our attempt to characterize the set of α-center stable clusterings, we will make use of a property called strict separation [5]. Informally, this property states that points are closer to other points in their own cluster than to points in any other cluster.

Definition 3. *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be a k-clustering of (X, d). We say that $\mathcal{C}$ satisfies a property called strict separation if for every $p, p' \in C_i$ and every $q \in C_j$ such that $i \neq j$, we have $d(p, p') < d(p, q)$.*

Furthermore, Ben-David and Reyzin showed a relationship between strict separation and α-center stability [5]. In particular, when α become large enough, an α-center stable clustering will automatically satisfy strict separation.

Theorem 1. *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be an α-center stable k-clustering for (X, d) with $\alpha > 2 + \sqrt{3}$. Then $\mathcal{C}$ satisfies the strict separation property.*

Proof. Let $p, p' \in C_i$ and let $q \in C_j$, for some $i \neq j$. The proof will follow from repeated application of the triangle inequality and of Definition 2. Suppose that p and p' have center c_i and that q has center c_j. First, we have

$$d(p, p') \leq d(p, c_i) + d(p', c_i) < \frac{1}{\alpha} d(p, c_j) + d(p', c_i), \tag{1}$$

where the first inequality follows from the triangle inequality, and the second inequality follows from the definition of α-center stability. Also, observe that $d(p', c_i) < \frac{1}{a} d(p', c_j)$. Thus, subtracting $\frac{1}{a} d(p', c_i)$ from both sides gives

$$\begin{aligned} (1 - \frac{1}{\alpha}) d(p', c_i) &< \frac{1}{a}(d(p', c_j) - d(p', c_i)) \\ &\leq \frac{1}{a} d(c_i, c_j). \end{aligned}$$

Therefore, it follows that

$$\begin{aligned} d(p', c_i) &< \frac{1}{\alpha - 1} d(c_i, c_j) \\ &\leq \frac{1}{\alpha - 1} d(p, c_i) + \frac{1}{\alpha - 1} d(p, c_j) \\ &< \frac{1}{\alpha(\alpha - 1)} d(p, c_j) + \frac{1}{\alpha - 1} d(p, c_j) \\ &= \frac{\alpha + 1}{\alpha(\alpha - 1)} d(p, c_j). \end{aligned}$$

By plugging this result into Equation 1, we get

$$\begin{aligned} d(p, p') &< \frac{1}{\alpha} d(p, c_j) + \frac{\alpha + 1}{\alpha(\alpha - 1)} d(p, c_j) \\ &= \frac{2}{\alpha - 1} d(p, c_j). \end{aligned} \tag{2}$$

We now obtain a bound on $d(p, c_j)$ in terms of $d(p, q)$. In particular, we have

$$\begin{aligned} d(p, c_j) &\leq d(p, q) + d(q, c_j) \\ &< d(p, q) + \frac{1}{\alpha} d(q, c_i) \\ &\leq d(p, q) + \frac{1}{\alpha}(d(p, q) + d(p, c_i)) \\ &= \frac{\alpha + 1}{\alpha} d(p, q) + \frac{1}{\alpha} d(p, c_i) \\ &< \frac{\alpha + 1}{\alpha} d(p, q) + \frac{1}{\alpha^2} d(p, c_j). \end{aligned}$$

Subtracting $\frac{1}{\alpha^2}$ on both sides gives $\frac{\alpha^2-1}{\alpha^2}d(p,c_j) < \frac{\alpha+1}{\alpha}d(p,q)$, or equivalently, $d(p,c_j) < \frac{\alpha}{\alpha-1}d(p,q)$. By plugging this into Equation 2, we conclude that

$$\begin{aligned} d(p,p') &< \frac{2}{\alpha-1}d(p,c_j) \\ &< (\frac{2}{\alpha-1})(\frac{\alpha}{\alpha-1})d(p,q) \\ &= \frac{2\alpha}{(\alpha-1)^2}d(p,q). \end{aligned}$$

Hence $d(p,p') < d(p,q)$ if we pick α such that $2\alpha < (\alpha-1)^2$. Solving this quadratic equation gives $\alpha > 2+\sqrt{3}$, as required. □

Though we won't be able to characterize the set of α-center stable clusterings for arbitrary values of α, we will be able to do so when $\alpha > 2+\sqrt{3}$. We begin by showing that, for any value of α, there exist instances that have many α-center stable clusterings. We first consider a set of points $X = \cup_{i=1}^{4} X_i$ such that the following conditions hold:

- For every $1 \le i \le 4$ and every $p,q \in X_i$, $d(p,q) = 1-\epsilon$.
- For any $p \in X_1$ and $q \in X_2$, $d(p,q) = \alpha$. Likewise, for any $p \in X_3$ and $q \in X_4$, $d(p,q) = \alpha$.
- If $1 \le i \le 2$ and $3 \le j \le 4$, then for any $p \in X_i$ and $q \in X_j$, $d(p,q) = \alpha^2 + \epsilon$.

For this instance, there are two α-center stable clusterings. In particular, it can be verified that $\mathcal{C} = \{X_1 \cup X_2, X_3, X_4\}$ and $\mathcal{C}' = \{X_1, X_2, X_3 \cup X_4\}$ are both valid α-center stable 3-clusterings of (X,d). Furthermore, by taking m independent copies of X, one can create an instance having exponentially many α-center stable clusterings.

Although this example is quite artificial, it does provide some useful insight regarding the differences between multiple α-center stable clusterings of the same instance. Observe that X_3 and X_4 are in separate clusters in $\mathcal{C}$, but in the same cluster in $\mathcal{C}'$. In general, one can show that for sufficiently large α, if two points are in distinct clusters in $\mathcal{C}$ but in the same cluster in $\mathcal{C}'$, then the corresponding cluster in $\mathcal{C}'$ must contain both of those clusters from $\mathcal{C}$. This occurs whenever $\alpha > 2+\sqrt{3}$, or more precisely, whenever $\mathcal{C}$ and $\mathcal{C}'$ satisfy strict separation. We formalize this observation with the following lemma.

Lemma 1. *Let $\mathcal{C} = \{C_1,\ldots,C_k\}$ and $\mathcal{C}' = \{C'_1,\ldots,C'_k\}$ be two distinct k-clusterings of (X,d) satisfying the strict separation property. Then, for any two distinct clusters C_i and C_j where $C_i \cap C'_m \neq \emptyset$ and $C_j \cap C'_m \neq \emptyset$, we have $C_i \cup C_j \subseteq C'_m$.*

Proof. Let $p \in C_i$ and $q \in C_j$ such that $p,q \in C'_m$. By strict separation, we have $d(p,p') < d(p,q)$ for every $p' \in C_i$. Likewise, we have $d(q,q') < d(p,q)$ for every $q' \in C_j$. Since p and q are in the same cluster in $\mathcal{C}'$, then C'_m must also include all of the points r such that $d(p,r) < d(p,q)$ or $d(q,r) < d(p,q)$. Hence $C_i \cup C_j \subseteq C'_m$, as required. □

4 Conclusions and Future Work

In this project, we have worked towards characterizing the set α-center stable clusterings, independent of any objective function. This was motivated by the fact that users might prefer a "nice" clustering rather than an optimal one. Furthermore, the goal was to show that finding an arbitrary α-center stable clustering is computationally easier than finding an optimal one, and although we were not able to show this, we believe that our observations can be useful in the pursuit of such algorithms.

In terms for future work, there are still many avenues left to explore. We conclude by listing several interesting open questions:

- What can be said about the set of α-center stable clusterings for smaller values of α?
- Can any of these properties be used to devise efficient algorithms that find α-center stable clusterings?
- Can other notions "clusterable" data provide more promising results?

References

[1] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Information Processing Letters*, 112(12):49 – 54, 2012.

[2] Ainesh Bakshi and Nadiia Chepurko. Polynomial time algorithm for 2-stable clustering instances. *arXiv preprint arXiv:1607.07431*, 2016.

[3] Maria Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. *SIAM Journal on Computing*, 45(1):102 – 155, 2016.

[4] Shai Ben-David. Computational feasibility of clustering under clusterability assumptions. *arXiv preprint arXiv:1501.00437*, 2015.

[5] Shalev Ben-David and Lev Reyzin. Data stability in clustering: a closer look. *Theoretical Computer Science*, 558:51 – 61, 2014.