Clustering and Stability

Amit Deshpande

Microsoft Research India

Outline

- Clustering objects by (NP-hard) objectives
- Stable instances
- Are stable instances easy (poly time solvable)? Why?
- Clustering objective, revisited

Clustering

- Given n data points $x_1, x_2, \ldots, x_n$,
  along with a similarity distance $d(x_i, x_j)$,
  and a positive integer k,
  partition the points into disjoint clusters
  so as to maximize
  similar points in the same cluster and
  dissimilar points in different clusters.

- Optimization over all k-partitions of the data

- Many simple objectives are NP-hard even for $k = 2$,
  e.g., maximize the sum of $d(x_i, x_j)$ over pairs not in the same cluster (a.k.a. the MaxCut problem)

Stability and MaxCut

- A given instance is α-perturbation *stable* if the optimal MaxCut does not change even when distances are perturbed within a multiplicative factor $\alpha > 1$.

$$d(x_i, x_j) \leq D(x_i, x_j) \leq \alpha d(x_i, x_j)$$

- Bilu-Linial (2010) show exact poly time algorithm for $\Omega(n)$-stable instances of MaxCut.
- Bilu-Daniely-Linial-Saks (2013) improved this to $\Omega(\sqrt{n})$-stable instances.
- Makarychev-Makarychev-Vijayaraghavan (2014) improved this to $\Omega(\sqrt{\log n} \log \log n)$-stable instances, and showed a matching negative result.

Center-based clustering, k-center/median/means

Given a set X of n data points and an integer $k > 0$, find k centers $c_1, c_2, \ldots, c_k$ that minimize

$$\phi(c_1, c_2, \ldots, c_k) = \left(\sum_{j=1}^{k} \left(\sum_{x \in C_j} d(x, c_j)^p \right)^{q/p} \right)^{1/q},$$

where C_j is the cluster of points that have c_j as their nearest center.

$p = q = 1$ is k-median, $p = q = 2$ is k-means, and $p = q = \infty$ is k-center. All are NP-hard objectives.

For this talk, let's call it discrete k-center/median/means if we optimize only over $C \subseteq X$ or some pre-specified discrete set as part of the input.

Center-based clustering

- Awasthi-Blum-Sheffet (2012) showed exact poly time algorithm for 3-stable instances of any center-based objective such as k-center/median/means.
- Balcan-Haghtalab-White (2016) showed exact poly time algorithm for 2-stable instances of symmetric/asymmetric k-center. No polytime algorithm for $(2-\epsilon)$-stable instances unless NP=RP.
- Balcan-Liang (2016) improved this to $(1+\sqrt{2})$-stable instances of k-center/median/means.
- Angelidakis-Makarychev-Makarychev (2017) improved this to 2-stable instances of k-center/median/means.

Other notions of stability

- Additive perturbation resilience proposed by Ackerman and Ben-David (2009).
- (c, ϵ)-approximation stability by Balcan-Blum-Gupta (2013), i.e., every c-approximation to the optimal cost is ϵ-close (in normalized set difference) to the optimal partition.
- Balcan-Liang (2016) showed that (c, ϵ)-approximation stability implies (c, ϵ)-perturbation resilience.
- ϵ-additive perturbation resilience by Vijayaraghavan et al. (2017), where points move by at most $\epsilon \max_{ij} \|\mu_i - \mu_j\|$.
- Kumar-Kannan (2010), Awasthi-Sheffet (2012), ...

The (Euclidean) k-means problem

Given a set $X \subseteq \mathbb{R}^d$ of n data points and an integer $k > 0$, the k-means clustering objective is to find k centers $c_1, c_2, \ldots, c_k \in \mathbb{R}^d$ that minimize

$$\phi(c_1, c_2, \ldots, c_k) = \sum_{j=1}^{k} \sum_{x \in C_j} \|x - c_j\|^2 ,$$

where C_j is the cluster of points that have c_j as their nearest center.

NP-hard even for $k = 2$ (Aloise et al. and Dasgupta-Freund, 2009) or $d = 2$ (Mahajan et al., 2009).

For this talk, let's call it discrete k-means if we optimize only over $C \subseteq X$ or some pre-specified discrete set as part of the input.

Center-proximity

For $\alpha > 1$, the clustering by $c_1, c_2, \ldots, c_k$ satisfies α-center-proximity if

$$\alpha \; d(x, c_i) \leq d(x, c_j), \quad \text{for } x \in C_i \text{ and } i \neq j.$$

That is, every point is closer by a multiplicative factor of α to its nearest center than to its second nearest center.

Center-proximity

For $\alpha > 1$, the clustering by $c_1, c_2, \ldots, c_k$ satisfies α-center-proximity if

$$\alpha \; d(x, c_i) \leq d(x, c_j), \quad \text{for } x \in C_i \text{ and } i \neq j.$$

That is, every point is closer by a multiplicative factor of α to its nearest center than to its second nearest center.

α-metric-perturbation-resilience implies α-center-proximity.

Center-proximity used as a proxy for metric perturbation-resilience indirectly in most previous results. Balcan-Liang (2016) exploit that the optimal clusters of $(1 + \sqrt{2})$-stable instances are contained in disjoint balls. Vijayaraghavan et al. (2017) use angular separation between optimal clusters for perceptron.

Why center-proximity?

- We do not know how to test if a given clustering instance is stable or perturbation-resilient.
- Underlying ground-truth clustering need not be optimal for our k-center/median/means objective.
- α much larger than 1 is good in theory but impractical.
- For any given $c_1, c_2, \ldots, c_k$ centers, we can easily check if their corresponding clusters satisfy α-center proximity.

Clustering objective, revisited

Given a set $X \subseteq \mathbb{R}^d$ of n data points and an integer $k > 0$ and a parameter $\alpha > 1$, find k centers $c_1, c_2, \ldots, c_k \in \mathbb{R}^d$ that minimize

$$\phi(c_1, c_2, \ldots, c_k) = \sum_{j=1}^{k} \sum_{x \in C_j} \|x - c_j\|^2 \,,$$

where C_j is the cluster of points that have c_j as their nearest center and the clusters satisfy α-center-proximity.

In other words, minimize the cost only over clusterings or partitions that have *additional desirable properties as the ground-truth*.

Our results

Joint work with Anand Louis and Apoorv Vikram Singh (IISc), to appear at AISTATS'19. `https://arxiv.org/abs/1804.10827`

- Exact algorithm to find α-center-proximal (balanced) clustering of the least *k*-means cost, in time exponential in *k* and $1/(\alpha - 1)$ but linear in the number of points *n* and the dimension *d*.
- Similar guarantees for *k*-means with outliers.
- Given any $\alpha > 1$, there exists $\alpha \geq \beta > 1$ and $\epsilon > 0$ such that it is NP-hard to $(1 + \epsilon)$-approximate the minimum of *k*-means objective over β-center-proximal (even balanced) clusterings.

Geometric insight

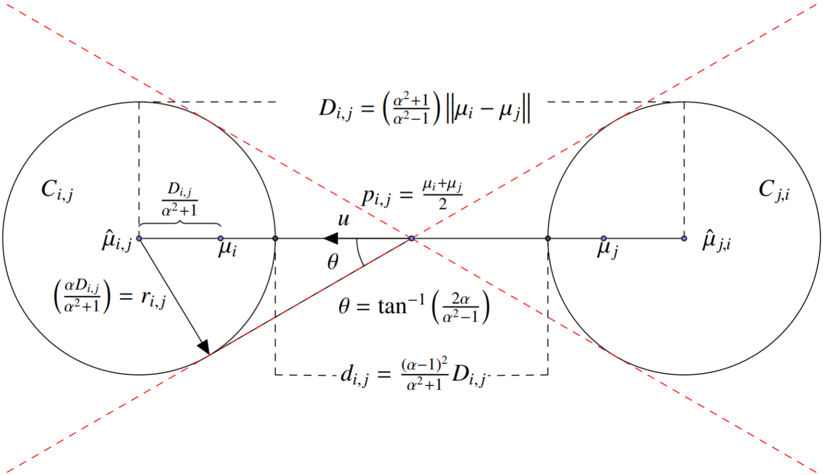

Figure 1: Geometric implication of α-center proximity property.

Related work, open problems

- Friggstad-Khodamoradi-Salavatipour (SODA'2019) show exact (local-search) algorithms for α-stable instances of *k*-means in doubling metrics in poly time.
  Caveat: works for only small or constant *d*.
  `https://arxiv.org/abs/1807.05443`.

- Are instances where most points satisfy α-center-proximity also easy?

- Any other reasonable notions of *stability*?

- How/why do *practical* heuristics work on *practical* instances?

Thank you. Any questions?