Approximation Algorithms for Stochastic Clustering (NeurIPS 2018 Poster)

David G. Harris[1], Shi Li[2], Thomas Pensyl[1], Aravind Srinivasan[3] and Khoa Trinh[1]

[1] Department of Computer Science, University of Maryland [2] Department of Computer Science & Engineering, University at Buffalo

[3] Department of Computer Science & Institute for Advanced Computer Studies, University of Maryland

Traditional Clustering Problems

- **Input**: a set C of clients
- a set F of potential centers
- a metric d over $F \cup C$

Remark: F and d may be implicitly given. In the k-means problem, F is the set of all points in the Euclidean space, and d is the ℓ_2 distance.

- **Output**: a subset $S \subseteq F$ of k centers so as to minimize some aggregating function of the vector $\big(d(j,S)\big)_{j \in C}$, where $d(j,S) \triangleq \min_{i \in S} d(j,i)$.
  - k-supplier-center: minimize $\max_{j \in C} d(j,S)$
  - k-median: minimize $\sum_{j \in C} d(j,S)$
  - k-means: minimize $\sum_{j \in C} d^2(j,S)$

The Motivational Question

Can we achieve better per-client guarantees by outputting a distribution over $\binom{F}{k} \triangleq \{S \subseteq F : |S| = k\}$?

A Motivational Example

- $k+1$ separated clusters with inter-cluster distances ≈ 1.
- a single set $S \in \binom{F}{k}$: $\max_{j \in C} d(j,S) \approx 1$
- a dist. π over $\binom{F}{k}$: $\max_{j \in C} \mathbb{E}_{S \sim \pi} d(j,S) \approx \frac{1}{k+1}$

Stochastic Clustering

- **Input**: C, F, d as before
- **Output**: A random $S \in \binom{F}{k}$ so that each client j has a good guarantee stochastically. Measurements of QoS include:
  - Covering Probability: $\Pr_S\big[d(j,S) \le r_j\big]$, where r_j is pre-specified.
  - Expected Service Cost: $\mathbb{E}_S[d(j,S)]$.

Applications

- Provider can periodically change service centers.

spring	summer	fall	winter	spring	summer	$\cdots$
S_1	S_2	S_3	S_4	S_1	S_2	$\cdots$

- Algorithmic Fairness: An e-commerce aggregator gives a sample of products to a randomly chosen set of k influencers, and each user hopes it will be similar to an influencer with high probability.

Problem 1: Chance k-Coverage

- **Input**: C, F, d as before
  $r_j \ge 0, p_j \in [0,1]$ for every $j \in C$
- Instance is feasible: $\exists$ dist. π^* over $\binom{F}{k}$ s.t.
  $$\Pr_{S \sim \pi^*}\big[d(j,S) \le r_j\big] \ge p_j, \forall j \in C.$$
- **Output**: A random S, following some dist. π.

- **Def.**: Algorithm is an (α, β)-approximation if
  $$\Pr_{S \sim \pi}\big[d(j,S) \le \alpha r_j\big] \ge \beta p_j, \forall j \in C.$$

Our Results for CkC

Cases	general	general	$p_j \equiv p$	$r_j \equiv r$
(α, β)	$(1, 1-1/e)$	$(9, 1)$	$(3, 1)$	$(3, 1)$

Tool 1: Greedy Clustering

Algorithm 1 GreedyCluster($r \in \mathbb{R}^C_{\ge 0}, w \in \mathbb{R}^C$)

1: $C^* \leftarrow \emptyset, C' \leftarrow C$
2: **while** $C' \ne \emptyset$ **do**
3: $j \leftarrow$ client in C' with the smallest w_j
4: $C^* \leftarrow C^* \cup \{j\}$
5: $C' \leftarrow C' \setminus \{j' : B(j, r_j) \cap B(j', r_{j'}) \ne \emptyset\}$
6: **return** C^*

- r_j defines the radius of the ball around j
- w defines the order in which we consider C

- **Lemma 1.** (1A) For every two distinct clients $j, j' \in C^*$, we have $B(j, r_j) \cap B(j', r_{j'}) = \emptyset$. (1B) For every $j' \in C$, $\exists j \in C^*$ with $w_j \le w_{j'}$ and $B(j, r_j) \cap B(j', r_{j'}) \ne \emptyset$.

Tool 2: Dependent Rounding

Lemma 2. There exists a poly-time algorithm DepRound(y) which takes as input $y \in [0,1]^n$ and outputs a random set $Y \subseteq [n]$ such that
(2A) $\Pr[i \in Y] = y_i, \forall i \in [n]$,
(2B) $\lfloor y([n]) \rfloor \le |Y| \le \lceil y([n]) \rceil$,
(2C) $\Pr[Y \cap S = \emptyset] \le \Pi_{i \in S}(1 - y_i), \forall S \subseteq [n]$.

LP Relaxation for CkC

- We can find a feasible solution y to LP:

$$\begin{aligned} y(B(j, r_j)) &\ge p_j & \forall j \in C \quad (1)\\ y(F) &\le k & (2)\\ y_i &\in [0,1] & \forall i \in F \quad (3)\end{aligned}$$

- $B(j, r) \triangleq \{i \in F : d(i,j) \le r\}, \forall j \in C, r \ge 0$.
- $y(S) \triangleq \sum_{i \in S} y_i, \forall S \subseteq F$.

$(1, 1-1/e)$-Approx. for CkC

1: solve LP (1)-(3) to obtain y
2: **return** $S \leftarrow$ DepRound(y)

- (2B) and (2) $\Rightarrow |S| \le \lceil y(F) \rceil \le k$.
  (2C) $\Rightarrow \Pr\left[d(j,S) \le r_j\right] = \Pr\left[S \cap B(j, r_j) \ne \emptyset\right]$
  $$\ge 1 - \prod_{i \in B(j, r_j)} (1 - y_j)$$
  $$\ge 1 - e^{y(B(j,r_j))} \ge 1 - 1/e.$$

$(3, 1)$-Approx. When $r_j \equiv r$ / $p_j \equiv p$

1: solve LP (1)-(3) to obtain y
2: $C^* \leftarrow$ GreedyCluster(r, $-p$ / r)
3: $V^* \leftarrow$ DepRound($p_{|C^*}$) $\quad \triangleright$ $p_{|C^*}$: p restricted to C^*
4: **return** {nearest $i \in F$ to $j : j \in V^*$}

- By (2B), (1), (1A) and (2), we have
  $$|V^*| \le \Big\lceil \sum_{j \in C^*} p_j \Big\rceil \le \Big\lceil \sum_{j \in C^*} y(B(j, r_j)) \Big\rceil \le \lceil y(F) \rceil \le k.$$

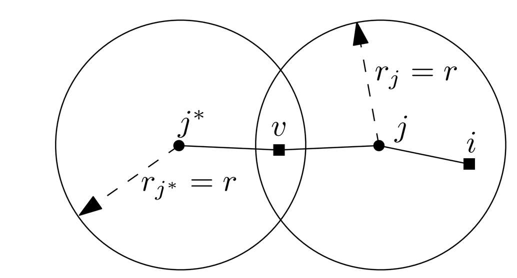

Figure 1: Analysis for covering probability of j^*.

- Case $r_j \equiv r$: Fix $j^* \in C$ and we analyze its covering probability. See Figure 1.
  - (1B) $\Rightarrow \exists j \in C^* : B(j, r_j) \cap B(j^*, r_{j^*}) \ne \emptyset$ & $p_j \ge p_{j^*}$.
  - Let i be nearest center to j, $v \in B(j, r_j) \cap B(j^*, r_{j^*})$.
  - $d(i, j^*) \le 3r_{j^*}$, via 3 hops $j^* - v - j - i$ of length $\le r_{j^*}$.
  - $\Pr[d(j,S) \le 3r_{j^*}] \ge \Pr[i \in S] \ge \Pr[j \in V^*] \overset{(1A)}{=} p_j \ge p_{j^*}$.
- Case $p_j \equiv p$: we have $p_j = p_{j^*}$ and $r_j \le r_{j^*}$; other parts of analysis are the same.

$(9, 1)$-Approx. for CkC

- Use the iterative rounding framework of [1].

Problem 2: Approximate $\mathbb{E}[d(j,S)]$

- **Input**: C, F, d as before, t_j for every $j \in C$
- Instance is feasible: $\exists$ dist. π^* over $\binom{F}{k}$ s.t.
  $$\mathop{\mathbb{E}}_{S \sim \pi^*} d(j,S) \le t_j, \forall j \in C.$$
- **Output**: A random S, following some dist. π s.t. $$\mathop{\mathbb{E}}_{S \sim \pi} d(j,S) \le \beta t_j, \forall j \in C.$$

Our Result for Problem 2

- Using 0-sum-game and Multiplicative Weight Update: α-approx. for k-median implies $(\alpha+\epsilon)$-approx. for Problem 2.

References

[1] R. Krishnaswamy, S. Li, and S. Sandeep. Constant approximation for k-median and k-means with outliers via iterative rounding. In *Proceedings of the 50th annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 646–659, 2018.