Clustering with fair center representation

Parameterized approximation algorithms and heuristics

Suhas Thejaswi[1] · Ameet Gadekar[1]
Bruno Ordozgoiti[2] · Michał Osadnik[1]
[1]Aalto University · [2]Queen Mary University of London

Diversity aware clustering FPT algorithm

- The problem of finding representatives among a set of individuals can be considered as a clustering problem.
- In certain scenarios, it may be adequate to consider additional requirements to ensure that some groups are adequately represented using cardinality requirements.
- We introduced this problem as the *diversity-aware k-median* problem in our earlier work [2].

1. Find feasible constraint patterns (brute-force enumeration).
2. Create an instance of a k-Med-k-PM problem for each set of facility types satisfying constraints.
3. Reduce the number of clients via coresets [1].
4. Guess *leaders* from a set of clients in coreset and guess distances of the closest facility in the optimal solution.
5. Making use of recent developments in submodular maximization subject to matroid constraint we obtain a $(1 + \frac{2}{\epsilon} + \epsilon)$ approximation.

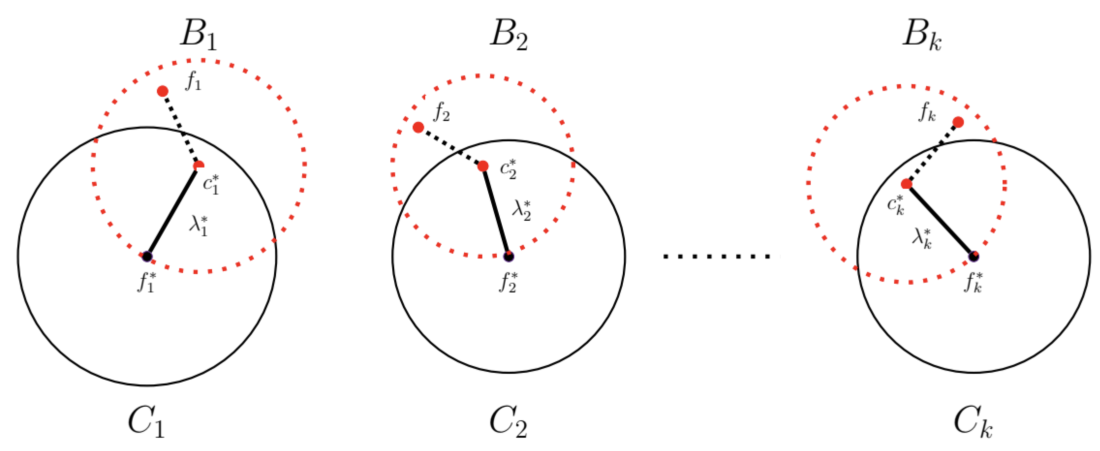

Figure 1: An illustration of facility selection for the FPT algorithm for solving k-Med-k-PM instance.

Problem formulation

Div-k-Med instance $((U,d), F, C, \mathscr{G}, \vec{r}, k)$.

Input:

- metric space (U,d)
- set $C \subseteq U$ of clients
- set $F \subseteq U$ of facilities
- a collection of facilities called **groups**, $\mathscr{G} = \{G_1, \ldots, G_t\}$
- vector of requirements $\vec{r} = (r[1], \ldots, r[t])$
- $k \leq |F|$

Output: subset of facilities $S \subseteq F$, satisfying:

- $|S \cap G_i| \geq r[i]$
- $|S| \leq k$
- clustering cost $\sum_{c \in C} d(c, S)$ is minimized.

Main result

Theorem 1 *For every $\epsilon > 0$, there exists a randomized $(1 + \frac{2}{e} + \epsilon)$-approximation algorithm for Div-k-Med with running time $f(k,t,\epsilon) \cdot \textsf{poly}(|U|)$, where $f(k,t,\epsilon) = \mathscr{O}\left(\left(\frac{2^t k^3 \log^2 k}{\epsilon^2 \log(1+\epsilon)}\right)^k\right)$. Furthermore, the approximation ratio is tight for any FPT algorithm w.r.t (k,t), assuming Gap-ETH. For Div-k-Means, with the same running time, we obtain a $(1+\frac{8}{e}+\epsilon)$-approximation, which is tight assuming Gap-ETH.*

Algorithmic results for Div-k-Med

Approx.	Approx. factor	Runtime method
$(3+\epsilon, 2k)$	LS + LP	$\mathscr{O}^*(2^{tk})$
$(3+\epsilon, 2k)$	LS + DP	$\mathscr{O}^*(kt2^t(r+1)^t)$
$(1+\frac{2}{e}+\epsilon, k)$	FPT(k,t,ϵ)	$\mathscr{O}^*\left(\left(\frac{2^t k^3 \log^2 k}{\epsilon^2 \log(1+\epsilon)}\right)^k\right)$

Note that the running times do not include the time needed for the submodular maximization due to the variety of techniques applicable.

With same running time bounds we obtain a $(1 + \frac{8}{e} + \epsilon)$-approximation for the Div-k-Means problem.

Bicriteria

The FPT approximation algorithms are theoretically the best possible, however, they are not practical. For bicriteria approximation, we first use an approximation algorithm for k-Median/k-Means. Then, if required we add facilities to satisfy the feasibility constraints (requirements vector) by solving the feasibility problem.

Strategies

- *Exhaustive Search* is the same strategy as in the previous algorithm that is terminating upon finding any feasible solution.
- *Dynamic Program* has lower theoretical running time and memory complexity but does not perform well when instances have multiple feasible solutions.
- *Linear Program* has good practical performance, however, this is a heuristic and it will not ensure finding a feasible solution always.

Additional results

Theorem 2 *For every $\epsilon > 0$, there exists a randomized $(3+\epsilon)$-approximation algorithm that outputs at most $2k$ facilities for the Div-k-Med problem in time $\mathscr{O}(2^t(r+1)^t \cdot \textsf{poly}(|U|, 1/\epsilon))$.*

Lemma 1 *Given an instance $I = ((U,d), F, C, \mathscr{G}, \vec{r}, k)$ of the Div-k-Med problem, we can enumerate all the k-multisets with feasible constraint pattern in time $\mathscr{O}(2^{tk} t|U|)$.*

(k-Med-k-PM) instance:

- metric space (U,d)
- set $C \subseteq U$ of clients
- set $F \subseteq U$ of facilities
- a partition of **groups**, $\mathscr{G} = \{G_1, \ldots, G_k\}$

Output: subset $S \subseteq F$ of facilities containing at most one facility from each group G_i and minimize clustering cost.

Experiments

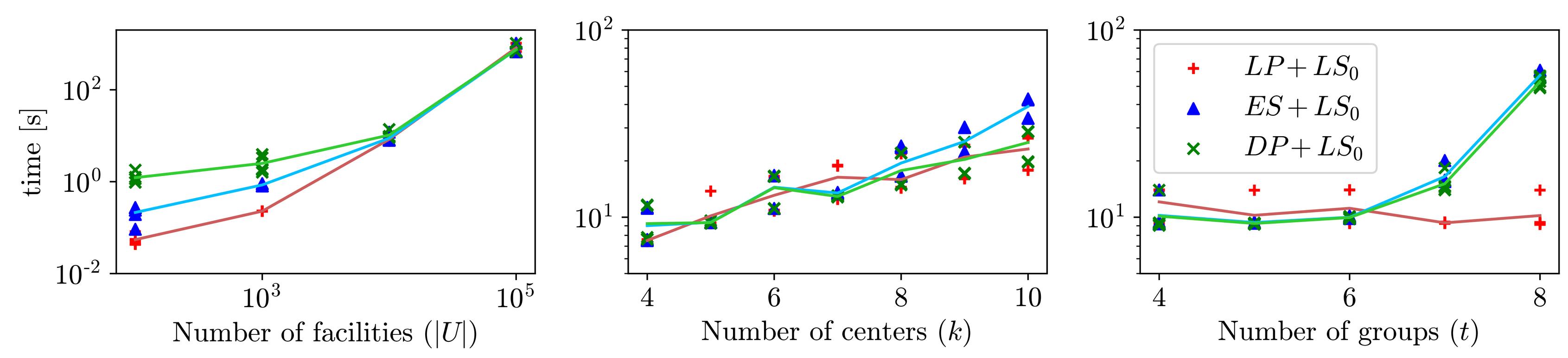

Figure 2: Scalability of bicriteria algorithms for Div-k-Med.

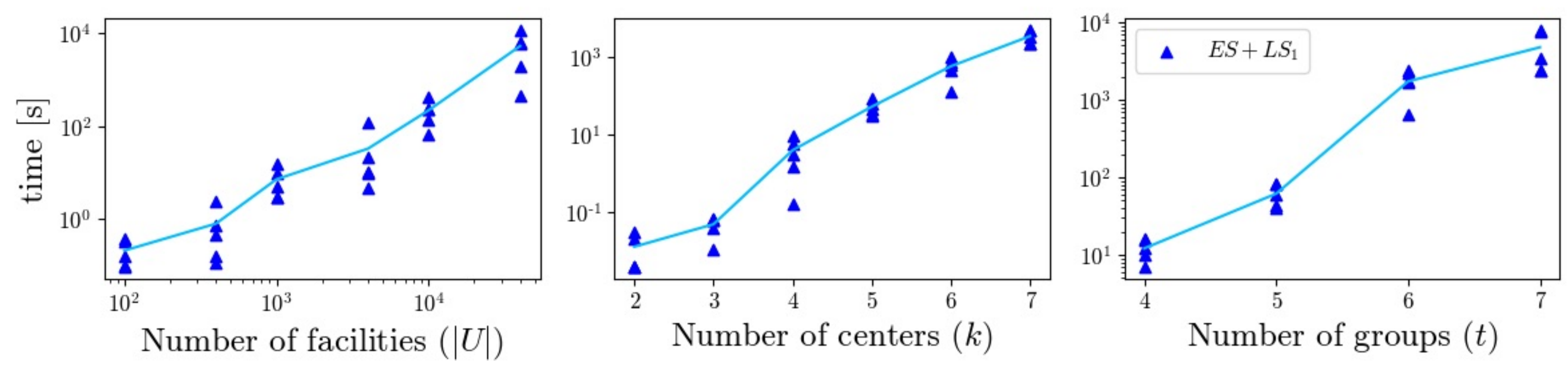

Figure 3: Scalability of $\textsf{ES} + \textsf{LS}_1$ algorithm for Div-k-Med.

References

[1] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *STOC*, page 569–578. ACM, 2011.
[2] S. Thejaswi, B. Ordozgoiti, and A. Gionis. Diversity-aware k-median: Clustering with fair center representation. In *ECML-PKDD*, pages 1–16. Springer, 2021.