Abstract

In this work, we study a wide range of *constrained* clustering problems in offline and streaming settings. We study these problems corresponding to three clustering objectives: k-median, k-means, and k-supplier. The (unconstrained) k-median problem is defined as follows. We are given a set of clients C in a metric space $\mathcal{X}$, with distance function $d(.,.)$. We are also given a set of feasible facility locations $L \subseteq \mathcal{X}$. The goal is to open a set $F \subseteq L$ of k facilities that minimizes the objective function: cost$(F, C) \equiv \sum_{j \in C} d(F, j)$, where $d(F, j)$ is the distance of client j to the closest facility in F. The k-means problem is defined in similar manner by replacing the distances with squared distances in the cost function, i.e., cost$(F, C) \equiv \sum_{j \in C} d(F, j)^2$. On the other hand, the k-supplier objective is defined as: cost$(F, C) \equiv \max_{j \in C} \{d(F, j)\}$. Furthermore, for $L = C$, the k-supplier problem is known as the k-center problem.

In many applications, there are additional constraints imposed on the clusters. For example, to balance the load among the facilities in resource allocation problems, a capacity u is imposed on every cluster. That is, no more than u clients can be assigned to any facility/cluster. This problem is known as the *capacitated* clustering problem. Likewise, various other applications have different constraints, which give rise to different *constrained* versions of the problem. In the past, the constrained versions of clustering problems were studied separately as independent problems. Recently, Ding and Xu [72] gave a unified framework for these problems that they called the *constrained clustering* framework. They proposed this framework in the context

of the k-median and k-means objectives in the continuous Euclidean space where $L = \mathbb{R}^p$ (p-dimensional Euclidean space) and C is a finite subset of $\mathbb{R}^p$. In this work, we extend this framework to the k-supplier objective and general metric spaces. The unified framework allows us to obtain results simultaneously for the following constrained versions of the problem: r-gather, r-capacity, balanced, chromatic, fault-tolerant, strongly private, ℓ-diversity, and fair clustering problems. We also study the *outlier* versions of these problems. In the outlier version, a clustering is obtained over at least $|C| - m$ clients instead of the entire client set.

For the constrained k-supplier and k-center problems, we obtain the following results:

(1) We give 3 and 2 approximation algorithms for the constrained k-supplier and k-center problems, respectively, with FPT (fixed-parameter tractable) running time $k^{O(k)} \cdot n^{O(1)}$, where $n = |C \cup L|$. Moreover, we note that the obtained approximation guarantees are tight. That is, for any constant $\varepsilon > 0$, no algorithm can achieve $(3-\varepsilon)$ and $(2-\varepsilon)$ approximation guarantees for the constrained k-supplier and k-center problems, respectively, in FPT time parameterized by k, assuming FPT $\neq$ W[2].

(2) For the outlier versions of the constrained k-supplier and k-center problems, we give 3 and 2 approximation guarantees with FPT running time $(k+m)^{O(k)} \cdot n^{O(1)}$, where $n = |C \cup L|$ and m is the number of outliers. Moreover, we note that the obtained approximation guarantees are tight. That is, for any constant $\varepsilon > 0$, no algorithm can achieve $(3 - \varepsilon)$ and $(2 - \varepsilon)$ approximation guarantees for the constrained k-supplier and k-center problems, respectively, in FPT time parameterized by k and m, assuming FPT $\neq$ W[2].

For the constrained k-median and k-means problems, we obtain the following results:

(3) We give $(3 + \varepsilon)$ and $(9 + \varepsilon)$ approximation algorithms for the constrained k-median and k-means problems, respectively, with FPT running time $(k/\varepsilon)^{O(k)} \cdot n^{O(1)}$, where

$n = |C \cup L|$. For the outlier version of the constrained k-median and k-means problems, we give $(3+\varepsilon)$ and $(9+\varepsilon)$ approximation algorithms, respectively, with FPT running time $\left((k+m)/\varepsilon\right)^{O(k)} \cdot n^{O(1)}$, where $n = |C \cup L|$ and m is the number of outliers.

(4) We also study the problems when $C \subseteq L$, i.e., a facility can be opened at a client location as well. For this special case, we design $(2+\varepsilon)$ and $(4+\varepsilon)$-approximation algorithms for the constrained k-median and k-means problems, respectively, with FPT running time $(k/\varepsilon)^{O(k)} \cdot n^{O(1)}$, where $n = |L|$. For the outlier version, we obtain the same approximation guarantees with FPT running time $\left((k+m)/\varepsilon\right)^{O(k)} \cdot n^{O(1)}$, where $n = |L|$ and m is the number of outliers. Note that the case $C \subseteq L$ subsumes the case $C = L$. Therefore, this result also holds for the case when $C = L$.

(5) We show that the analysis of our algorithm is tight. That is, there are instances for which our algorithm does not provide better than $(3-\delta)$ and $(9-\delta)$ approximation guarantee corresponding to k-median and k-means objectives, respectively, for any arbitrarily small constant $\delta > 0$. Similarly, the analysis of our algorithm is tight for the special case $C \subseteq L$.

(6) Our algorithms are based on a simple sampling-based approach. This approach allows us to convert these algorithms to constant-pass log-space streaming algorithms.

(7) We also study the constrained k-median/means problem in continuous Euclidean space where $L = \mathbb{R}^p$ and C is a finite subset of $\mathbb{R}^p$. We design $(1+\varepsilon)$-approximation algorithm for the outlier version of these problems with FPT running time $O\left(np \cdot ((k+m)/\varepsilon)^{O(k/\varepsilon^{O(1)})}\right)$, where $n = |C|$ and m is the number of outliers. We also convert these algorithms to constant-pass log-space streaming algorithms.

We also study the *socially fair k-median/k-means problem*, which is a generalization of the k-supplier and k-median/means problems. The problem is defined as follows. We are given a set of clients C in a metric space $\mathcal{X}$ with a distance function $d(.,.)$. There are ℓ groups:

$C_1, \ldots, C_\ell \subseteq C$. We are also given a set L of feasible centers in $\mathcal{X}$. The goal in the socially fair k-median problem is to find a set $F \subseteq L$ of k centers that minimizes the maximum average cost over all the groups. That is, find F that minimizes the objective function: fair-cost$(F, C) \equiv \max_j \left\{ \sum_{x \in C_j} d(F, x)/|C_j| \right\}$, where $d(F, x)$ is the distance of x to the closest center in F. The socially fair k-means problem is defined similarly by using squared distances, i.e., $d^2(.,.)$ instead of $d(.,.)$. We obtain the following results for this problem:

(8) We design $(3+\varepsilon)$ and $(9+\varepsilon)$ approximation algorithms for the socially fair k-median and k-means problems, respectively, in FPT time $f(k, \varepsilon) \cdot n^{O(1)}$, where $f(k, \varepsilon) = (k/\varepsilon)^{O(k)}$ and $n = |C \cup L|$.

(9) Furthermore, these approximation guarantees are tight; that is, for any constant $\varepsilon > 0$, no algorithm can achieve $(3 - \varepsilon)$ and $(9 - \varepsilon)$ approximation guarantees for the socially fair k-median and k-means problems in FPT time parametrized by k, assuming FPT $\neq$ W[2].

Lastly, we give hardness of approximation result for the k-median problem in the continuous Euclidean space where $L = \mathbb{R}^p$ and C is a finite subset of $\mathbb{R}^p$. This solves an open problem posed explicitly in the work of Awasthi *et al.* [19]. More precisely, we obtain the following result:

(10) There exists a constant $\varepsilon > 0$ such that the Euclidean k-median problem in $O(\log k)$ dimensional space cannot be approximated to a factor better than $(1 + \varepsilon)$, assuming the Unique Games Conjecture.

Furthermore, we study the hardness of approximation for the Euclidean k-means/k-median problems in the *bi-criteria setting*. In the bi-criteria setting, algorithms are allowed to output βk centers (for some constant $\beta > 1$), and the approximation ratio is computed with respect to the optimal k-means/k-median cost. We show the following results:

(11) For any constant $1 < \beta < 1.015$, there exists a constant $\varepsilon > 0$ such that there is no $(1 + \varepsilon)$ bi-criteria approximation algorithm for the Euclidean k-median problem in $O(\log k)$ dimensional space assuming the Unique Games Conjecture.

(12) For any constant $1 < \beta < 1.28$, there exists a constant $\varepsilon > 0$ such that there is no $(1 + \varepsilon)$ bi-criteria approximation algorithm for the Euclidean k-means problem in $O(\log k)$ dimensional space assuming the Unique Games Conjecture.