Fair $k$-Center Clustering for Data Summarization
Matthäus Kleindessner, Pranjal Awasthi, Jamie Morgenstern

A linear-time approximation algorithm

We need a slightly more general form of the standard (unfair) $k$-center problem. Algorithm $1$ can be easily adapted to yield a 2-approximation algorithm for it.

**Algorithm 2** $k$-center with given centers $C_0 \subseteq S_1$

\[
\minimize_{C = \{c_1, \ldots, c_k\} \subseteq S\text{ with }d(c, S) = \min_{c \in S} d(c, S)} \max_{c \in C} d(s, C)
\]

Our algorithm is a recursive algorithm with respect to the number of groups $m$. It is easiest to understand in the case of two groups.

**Two groups ($S = S_1 \cup S_2$)**

1) Run Algorithm $1$ on $S$ with $k = k_1 + k_2$. Let $\hat{C}$ be its output and $\hat{k}_1 = |\hat{C} \cap S_1|$, $i \in \{1, 2\}$. The cost of $\hat{C}$ is at most twice the optimal cost.

2) If $\hat{k}_1 = k_1$ and $\hat{k}_2 = k_2$, we are done. Assume that $\hat{k}_1 > k_1$ and $\hat{k}_2 < k_2$.

As long as $\hat{k}_1 > k_1$, there is a center in $\hat{C} \cap S_1$ that has an element of $S_1$ in its cluster, replace such a center with an element of $S_1$ in its cluster. Due to the triangle inequality, the cost of $\hat{C}$ is at most doubled. Once we have made all such available swaps, the remaining clusters with a center in $S_1$ are entirely contained within $S_1$. Denote the union of these clusters by $S' \subseteq S_1$.

3) Apply Algorithm $2$ to $S'$ with $k = k_1$ and the centers in $\hat{C} \cap S_1$ as given centers $C_0$. Denote its output by $\hat{C}$. Recall that $\hat{C} \subseteq S_1$.

4) We can show that the cost of $\hat{C} \cup (\hat{C} \cap S_1)$ is at most five times the optimal cost.

Return $\hat{C} \cup (\hat{C} \cap S_1)$ arbitrary elements of $S_2$ as final output.

**Example:**

1) $\hat{C}_1 = S_1$ and $\hat{C}_2 = S_2$

2) $\hat{C}_1 = S_1$ and $\hat{C}_2 = S_2$

3) $\hat{k}_1 = 4$, $\hat{k}_2 = 0$ and $\hat{k}_2 = 1$, $\hat{k}_1 = 0$.

4) $\hat{k}_1 = 3$, $\hat{k}_2 = 1$

Our algorithm is a linear-time $5$-approximation algorithm when $m = 2$.

**Theorem 1** Our algorithm is a $5$-approximation algorithm for the fair $k$-center problem with $m = 2$ groups, not a $(5 - \varepsilon)$-approximation algorithm for any $\varepsilon > 0$. It has running time $O(k|S|)$, assuming $d$ can be evaluated in $O(1)$.

**Arbitrary number of groups:**

The main idea is the same as for the case $m = 2$.

1) Run Algorithm $1$ on $S$ with $k = \sum_{i = 1}^{m} k_i$.

2) Exchange centers for elements in their clusters such that the number of centers from $S_i$ comes closer to $k_i$.

3) Run Algorithm $2$ on a subset $S' \subseteq S \setminus S_i$, where the number of centers from $S_i$ is less than or equal to $k_i$ (we may consider $S_i$ to have been “resolved”).

The difficulty with this idea comes from the exchanging process:

By constructing a graph on the set of groups and computing all shortest paths on it we can overcome this difficulty. Our resulting algorithm has the following guarantee:

**Theorem 2** Our algorithm is a $(3 \cdot 2^m - 1)$-approximation algorithm for the fair $k$-center problem with $m$ groups, but not a $(8 - \varepsilon)$-approximation algorithm. It has running time $O(k|S|\log^2|S|)$, assuming $d$ can be evaluated in $O(1)$.

**What is the exact approximation factor of our algorithm for $m > 2$?**

**Experiments**

![Approximation factor of our algorithm (Alg. 4) and the algorithm by Chen et al. (2016) (M.C.) in various scenarios.](image)

The cost of the output (left) and running time (right) of our algorithm (Alg. 4) and the algorithm by Chen et al. (2016) (M.C.) as a function of $|S| = n$. The images of the medical doctors were found on [https://pixnio.com](https://pixnio.com) and [https://commons.wikimedia.org](https://commons.wikimedia.org) and are in the public domain.

**References**
