SCIENCE CHINA
Information Sciences

• LETTER •

May 2021, Vol. 64 150104:1–150104:2
https://doi.org/10.1007/s11432-020-3066-x

Special Focus on Constraints and Optimization in Artificial Intelligence

An approximation algorithm for *k*-median with priorities

Zhen ZHANG[1], Qilong FENG[1*], Jinhui XU[2] & Jianxin WANG[1]

[1]*School of Computer Science and Engineering, Central South University, Changsha* 410083*, China;*
[2]*Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo* 14200*, USA*

Received 27 April 2020/Accepted 4 August 2020/Published online 31 March 2021

Citation Zhang Z, Feng Q L, Xu J H, et al. An approximation algorithm for *k*-median with priorities. Sci China Inf Sci, 2021, 64(5): 150104, https://doi.org/10.1007/s11432-020-3066-x

Dear editor,

Clustering is a fundamental problem in computer science. This problem is to partition a given set of clients into several clusters such that clients in the same cluster are more similar to each other. In many clustering applications, the clients are different in the levels of services they require. Motivated by such applications, Ravi and Sinha [1] introduced the problem of clustering with priorities, wherein each client is associated with a priority and can only be assigned to a facility opened at the same or higher priorities.

Designing approximation algorithms for clustering with priorities remains an active area of research. For example, several constant factor approximation algorithms are known for the facility location objective [1–3]. Unfortunately, clustering with priorities remains elusive under the extensively studied k-median objective. The main difficulty in dealing with the k-median objective lies in the hard constraint on the number of opened facilities. Kumar and Sabharwal [4] studied the k-median problem with a similar priority constraint. Unfortunately, their algorithm only works for the case where the clients have no more than two different priorities. There has also been work devoted on obtaining approximation algorithms for other clustering problems where the clients require different services, such as clustering with service installation costs [5]. However, these algorithms are not applicable when the priority constraint is imposed. In this study, we study the k-median with priorities problem (k-MP), which can be formally defined as follows.

Definition 1 (k-median with priorities)**.** Given a set $\mathcal{D}$ of clients and a set $\mathcal{F}$ of facilities in a metric space, a set $\mathcal{P} = \{1, 2, \dots, L\}$ of priorities, and an integer $k > 0$, where each client $j \in \mathcal{D}$ is associated with a priority $g(j) \in \mathcal{P}$, each priority $p \in \mathcal{P}$ is associated with a cost $f(p) > 0$ for opening any facility at the priority, and $f(p_1) \geqslant f(p_2)$ for each $p_1, p_2 \in \mathcal{P}$ with $p_1 > p_2$, the goal is to open no more than k facilities and assign each client to a facility opened at the same or higher priorities, such that the total cost (including facility opening cost and the sum of the distance from each client to the corresponding facility) is minimized.

We give a $(6.6743+\epsilon)$-approximation for k-MP. A different deterministic rounding approach is proposed to deal with the priority constraint, which is the crucial step in getting the constant factor approximation ratio.

Theorem 1. Given an instance $\mathcal{I} = (\mathcal{D}, \mathcal{F}, \mathcal{P}, k, g, f)$ of k-MP and a real number $\epsilon > 0$, there is a $(6.6743 + \epsilon)$-approximation algorithm for the problem. The running time of the algorithm is polynomial in $|\mathcal{D}|$, $|\mathcal{F}|$, $|\mathcal{P}|$, and k.

Our algorithm. Let $\mathcal{I} = (\mathcal{D}, \mathcal{F}, \mathcal{P}, k, g, f)$ be an instance of k-MP. For each $i, j \in \mathcal{F} \cup \mathcal{D}$, let $c(i, j)$ denote the distance for i to j. For each $j \in \mathcal{D}$ and $\mathcal{F}' \subseteq \mathcal{F}$, define $c(j, \mathcal{F}') = \min_{i\in\mathcal{F}'} c(j, i)$. For each $p \in \mathcal{P}$, define $\mathcal{D}_p = \{j \in \mathcal{D} : g(j) = p\}$. Without loss of generalization, we can assume that $\mathcal{D}_p \neq \emptyset$ for each $p \in \mathcal{P}$. We formalize $\mathcal{I}$ as an integer programming and relax the integrality constraints to get the following linear programming (LP):

$$\min \sum_{i\in\mathcal{F}, j\in\mathcal{D}} c(j,i)x_{ij} + \sum_{i\in\mathcal{F}, p\in\mathcal{P}} f(p)y_{ip} \qquad \text{LP1}$$

$$\text{s.t.} \ \sum_{i\in\mathcal{F}} x_{ij} = 1, \forall j \in \mathcal{D}, \qquad (1)$$

$$x_{ij} \leqslant \sum_{p\in\mathcal{P} \cap p\geqslant g(j)} y_{ip}, \ \forall i \in \mathcal{F}, \ j \in \mathcal{D}, \qquad (2)$$

$$\sum_{i\in\mathcal{F}, p\in\mathcal{P}} y_{ip} \leqslant k, \qquad (3)$$

$$x_{ij}, y_{ip} \geqslant 0, \ \forall i \in \mathcal{F}, \ j \in \mathcal{D}, \ p \in \mathcal{P}. \qquad (4)$$

LP1 have a variable x_{ij} for each $i \in \mathcal{F}$ and $j \in \mathcal{D}$ indicating whether j is assigned to i, and a variable y_{ip} for each $i \in \mathcal{F}$ and $p \in \mathcal{P}$ indicating whether i is opened at p. Constraints (1) and (2) say that each client must be assigned to a facility opened at the same or higher priorities, and constraint (3) says that at most k facilities can be opened. Let OPT denote the cost of an optimal integer solution to LP1.

We consider the following relaxation of LP1 for $\gamma \geqslant 0$, where constraint (3) is removed and the penalty for violating the constraint is added to the objective function:

$$\min \sum_{i\in\mathcal{F}, j\in\mathcal{D}} c(j,i)x_{ij} + \sum_{i\in\mathcal{F}, p\in\mathcal{P}} (f(p)+\gamma)y_{ip} - \gamma k \quad \text{LP2}^{\gamma}$$

[*] Corresponding author (email: csufeng@csu.edu.cn)

$$\text{s.t.}\quad (1), (2), \text{and } (4).$$

Given a solution $\mathcal{G}' = (x', y')$ to LP2$^\gamma$, define $C(\mathcal{G}') = \sum_{i\in\mathcal{F}, j\in\mathcal{D}} c(j,i)x'_{ij}$ and $F(\mathcal{G}') = \sum_{i\in\mathcal{F}, p\in\mathcal{P}} f(p)y'_{ip}$ for brevity. Let $V(\mathcal{G}') = F(\mathcal{G}') + C(\mathcal{G}')$. We can get integer solutions to LP2$^\gamma$ that have the following guarantee.

Lemma 1. Given an instance $\mathcal{I} = (\mathcal{D}, \mathcal{F}, \mathcal{P}, k, g, f)$ of k-MP and a real number $\epsilon > 0$, we can find in polynomial time either a 3-approximation solution to $\mathcal{I}$, or two integer solutions $\mathcal{G}_1 = (x^1, y^1)$ and $\mathcal{G}_2 = (x^2, y^2)$ to LP2$^\gamma$ such that $\sum_{i\in\mathcal{F}, p\in\mathcal{P}} y^1_{ip} = k_1 < k$, $\sum_{i\in\mathcal{F}, p\in\mathcal{P}} y^2_{ip} = k_2 > k$, and $aV(\mathcal{G}_1) + bV(\mathcal{G}_2) \leqslant \big(3 + O(\epsilon)\big)\mathrm{OPT}$, where $a = \frac{k_2 - k}{k_2 - k_1}$ and $b = 1 - a$ (see Appendix A for the proof).

We find two integer solutions $\mathcal{G}_1 = \{x^1, y^1\}$ and $\mathcal{G}_2 = \{x^2, y^2\}$ using Lemma 1, where $\sum_{i\in\mathcal{F}, p\in\mathcal{P}} y^1_{ip} = k_1 < k$ and $\sum_{i\in\mathcal{F}, p\in\mathcal{P}} y^2_{ip} = k_2 > k$. Let $V(\mathcal{G}_f) = aV(\mathcal{G}_1) + bV(\mathcal{G}_2)$ for brevity. Let $\mathcal{H}^1 = \{i \in \mathcal{F} : \sum_{p\in\mathcal{P}} y^1_{ip} \geqslant 1\}$ and $\mathcal{H}^2 = \{i \in \mathcal{F} : \sum_{p\in\mathcal{P}} y^2_{ip} \geqslant 1\}$. We have $|\mathcal{H}^1| \leqslant k_1$ and $|\mathcal{H}^2| \leqslant k_2$. For each $i \in \mathcal{H}^1$, let $p_1(i) = \max_{p\in\mathcal{P}\,\cap\, y^1_{ip}=1} p$. Similarly, let $p_2(i) = \max_{p\in\mathcal{P}\,\cap\, y^2_{ip}=1} p$ for each $i \in \mathcal{H}^2$. For each $j \in \mathcal{D}$, let $i_1(j)$ and $i_2(j)$ denote its nearest facilities from $\{i \in \mathcal{H}^1 : p_1(i) \geqslant g(j)\}$ and $\{i \in \mathcal{H}^2 : p_2(i) \geqslant g(j)\}$, respectively. Let $c_1(j) = c(j, i_1(j))$ and $c_2(j) = c(j, i_2(j))$. Let $\phi(i) = \{j \in \mathcal{D} : i_2(j) = i\}$ for each $i \in \mathcal{H}^2$ and $\phi(\mathcal{A}) = \bigcup_{i\in\mathcal{A}} \phi(i)$ for each $\mathcal{A} \subseteq \mathcal{H}^2$.

Proposition 1. For any $0 < \tau < 1$, if $V(\mathcal{G}_1) \geqslant \frac{1}{\tau}V(\mathcal{G}_f)$, then we have $a < \tau$ and $V(\mathcal{G}_2) < V(\mathcal{G}_f)$ (see Appendix B for the proof).

Given a real number $0 < \tau < 1$, we break the analysis into the following two cases: (1) $V(\mathcal{G}_1) < \frac{1}{\tau}V(\mathcal{G}_f)$, and (2) $V(\mathcal{G}_1) \geqslant \frac{1}{\tau}V(\mathcal{G}_f)$. For case (1), we have $V(\mathcal{G}_1) < \frac{1}{\tau}\big(3+O(\epsilon)\big)\mathrm{OPT}$ by Lemma 1. Recall that $\mathcal{G}_1$ is a feasible solution to LP1. Thus, $\mathcal{G}_1$ yields a $\frac{1}{\tau}\big(3+O(\epsilon)\big)$-approximation for the problem for case (1).

We can assume that $|\mathcal{H}^2| > k$ for case (2). Otherwise we can convert $\mathcal{G}_2$ to a feasible solution to LP1 by retaining only the highest priority opened at each facility. Using Lemma 1 and Proposition 1, we know that $\mathcal{G}_2$ is a $\big(3+O(\epsilon)\big)$-approximation solution to the problem. For each $i' \in \mathcal{H}^2$, let $\varphi(i')$ denote the nearest facility to i' from $\mathcal{H}^1$. Let $\mathcal{L}_i = \{i' \in \mathcal{H}^2 : \varphi(i') = i\}$ for each $i \in \mathcal{H}^1$. By triangle inequality and the definition of φ, for each $j \in \mathcal{D}$, we have $c(i_2(j), \varphi(i_2(j))) \leqslant c(i_2(j), i_1(j)) \leqslant c_2(j) + c_1(j)$, and thus

$$c(j, \varphi(i_2(j))) \leqslant 2c_2(j) + c_1(j). \tag{5}$$

Our idea is to consider $\mathcal{G}_2$ as the solution to $\mathcal{I}$ initially, and then reduce the number of opened facilities to k. Let $\Psi_1(i) = \sum_{j\in\phi(\mathcal{L}_i)} \big(c_2(j) + c_1(j)\big)$ for each $i \in \mathcal{H}^1$ and $\Psi_2(i') = \sum_{j\in\phi(i')} \big(c_2(j)+c_1(j)\big)$ for each $i' \in \mathcal{H}^2$. Consider a facility $i \in \mathcal{H}^1$. If we open i at priority $\max_{i'\in\mathcal{L}_i} p_2(i')$, close all the facilities from $\mathcal{L}_i$, and reassign each $j \in \phi(\mathcal{L}_i)$ to i, then the number of opened facilities can be reduced by $|\mathcal{L}_i| - 1$, and the cost of the solution is increased by no more than $\Psi_1(i)$ owing to inequality (5). We consider the following LP that minimizes the increased cost and reduces the number of opened facilities to k.

$$\begin{aligned} \min \quad & \sum_{i\in\mathcal{H}^1} z_i \Psi_1(i) && \text{LP3} \\ \text{s.t.} \quad & \sum_{i\in\mathcal{H}^1} z_i(|\mathcal{L}_i| - 1) = |\mathcal{H}^2| - k, && (6) \\ & 0 \leqslant z_i \leqslant 1, \quad \forall i \in \mathcal{H}^1. && (7) \end{aligned}$$

LP3 associates a variable z_i with each $i \in \mathcal{H}^1$. $z_i = 1$ indicates that we close all the facilities from $\mathcal{L}_i$, open i at priority $\max_{i'\in\mathcal{L}_i} p_2(i')$, and reassign each $j \in \phi(\mathcal{L}_i)$ to i. Constraint (6) enforces that the number of opened facilities should be reduced to k.

Proposition 2. We can find in polynomial time an optimal solution to LP3 that has at most one fractional variable, which is associated with a facility $i \in \mathcal{H}^1$ such that $|\mathcal{L}_i| > 1$ (see Appendix C for the proof).

Let z^* be the solution to LP3 given by Proposition 2. Let $\mathcal{O}_0 = \{i \in \mathcal{H}^1 : z^*_i = 0\}$ and $\mathcal{O}_1 = \{i \in \mathcal{H}^1 : z^*_i = 1\}$. We construct a solution to k-MP as follows. For each $i \in \mathcal{O}_0$ and $i' \in \mathcal{L}_i$, open i' at priority $p_2(i')$, and assign each $j \in \phi(i')$ to i'. For each $i \in \mathcal{O}_1$, open i at priority $\max_{i'\in\mathcal{L}_i} p_2(i')$, and assign each $j \in \phi(\mathcal{L}_i)$ to i. If z^* has a fractional variable, then let $t \in \mathcal{H}^1$ denote the facility associated with the fractional variable. Let $\mathcal{L}^\dagger = \arg\min_{\mathcal{L}\subseteq\mathcal{L}_t\,\cap\,|\mathcal{L}|=\lceil z^*_t|\mathcal{L}_t|\rceil} \sum_{i'\in\mathcal{L}} \Psi_2(i')$. We open t at priority $\max_{i'\in\mathcal{L}^\dagger} p_2(i')$ and assign each $j \in \phi(\mathcal{L}^\dagger)$ to t. For each $i' \in \mathcal{L}_t\backslash\mathcal{L}^\dagger$, open i' at priority $p_2(i')$, and assign each $j \in \phi(i')$ to i'. It may be the case that $\mathcal{H}^1 \cap \mathcal{H}^2 \neq \emptyset$ and $t \in \mathcal{L}_t\backslash\mathcal{L}^\dagger$. For this case, t may be opened at two different priorities, and we can close the lower one without adjusting the assignment of the clients from $\phi(\mathcal{L}_t)$. If z^* does not have a fractional variable, then let $\mathcal{L}_t = \mathcal{L}^\dagger = \emptyset$. Let $\mathcal{H}'$ denote the set of facilities opened in the solution. The following result implies that our solution is a feasible one to k-MP.

Lemma 2. $|\mathcal{H}'| \leqslant k$ (see Appendix D for the proof).

Let Φ denote the cost of our solution for k-MP. We are able to show that Φ is near to $V(\mathcal{G}_f)$.

Lemma 3. For any $0 < \tau < 1$, if $V(\mathcal{G}_1) \geqslant \frac{1}{\tau}V(\mathcal{G}_f)$, then $\Phi < \max\{2, \frac{3+\tau}{2-\tau}\}V(\mathcal{G}_f)$ (see Appendix E for the proof).

Using Lemmas 1 and 3, we know that our solution is a $(\max\{6, \frac{9+3\tau}{2-\tau}\} + O(\epsilon))$-approximation solution to k-MP for case (2). Recall that $\mathcal{G}_1$ is a $\frac{1}{\tau}\big(3 + O(\epsilon)\big)$-approximation solution to k-MP for case (1). Putting together, we obtain a $(\max\{6, \frac{9+3\tau}{2-\tau}, \frac{3}{\tau}\} + O(\frac{\epsilon}{\tau}))$-approximation solution for the problem. Let $\tau = \sqrt{6} - 2$, and then the approximation ratio is upper-bounded by $6.6743 + O(\epsilon)$.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61872450, 61672536, 61828205, 61802441, 71631008), Hunan Provincial Key Lab on Bioinformatics, and Hunan Provincial Science and Technology Program (Grant No. 2018WK4001).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Ravi R, Sinha A. Multicommodity facility location. In: Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms, 2004. 342–349

2 Li G D, Wang Z, Wu C C. Approximation algorithms for the stochastic priority facility location problem. Optimization, 2013, 62: 919–928

3 Wang F M, Xu D C, Wu C C. Approximation algorithms for the priority facility location problem with penalties. J Syst Sci Complex, 2015, 28: 1102–1114

4 Kumar A, Sabharwal Y. The priority k-median problem. In: Proceedings of the 27th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, 2007. 71–83

5 Shmoys D B, Swamy C, Levi R. Facility location with service installation costs. In: Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms, 2004. 1088–1097