

# $k$ -median 问题反向贪心随机算法

王守强

(山东交通学院信息科学与电气工程学院 济南 250023)

**摘 要**  $k$ -median 问题的近似算法研究一直是计算机科学工作者关注的焦点。基于均衡限制条件,利用反向贪心策略,给出求解该问题的随机近似算法。证明该算法以较大的概率满足其近似性能比的期望值为  $(3+O(\ln(\ln(k)/\alpha)))$ 。该算法的时间复杂度为  $O(\lceil \frac{k}{\alpha} \ln(k) \rceil^2 (n+m))$ ,其中  $n$  和  $m$  分别代表设施集合以及客户点集的大小。最后,通过计算机实验验证了  $k$ -median 问题的反向贪心算法的实际计算效果。

**关键词**  $k$ -median, 随机算法, 反向贪心, 近似性能比

**中图分类号** TP301 **文献标识码** A

## Randomized Reverse Greedy Algorithm for $k$ -median Problem

WANG Shou-qiang

(Department of Information Engineering, Shandong Jiaotong University, Jinan 250023, China)

**Abstract** Research on the approximated algorithms for  $k$ -median problem has been a focus of computer scientists. Based on the balanced parameter, this paper presented a randomized algorithm for  $k$ -median problem by means of the reverse greedy. The new algorithm's expected approximate ratio was proved to be  $(3+O(\ln(\ln(k)/\alpha)))$  with high probability. The running time of the new algorithm is  $O(\lceil \frac{k}{\alpha} \ln(k) \rceil^2 (n+m))$ , where  $n$  and  $m$  represent the number of clients and facilities for the given instance. Finally, computer verification was used to study the real computational effect of the algorithm.

**Keywords**  $k$ -median, Randomized algorithm, Reverse greedy, Approximation ratio

## 1 引言

给定设施集合  $F$  和客户集合  $C$ ,  $d(f_i, c_j) \in \mathbb{Z}^+$  表示设施  $f_i \in F$  为客户  $c_j \in C$  提供服务的费用。欲在  $F$  中选择至多  $k$  个设施为  $C$  中客户服务,并使总的服务费用最小,即所谓  $k$  中心问题,简称  $k$ -median。该问题形式化描述为:

实例:设施点集  $F$  和客户点集  $C$ , 正整数  $k \in \mathbb{Z}^+$ ,  $k \leq |F|$ ,  $\forall f_i \in F, c_j \in C, d(f_i, c_j)$  表示设施  $f_i$  为客户  $c_j$  服务的费用。

目标:求  $M \subseteq F$ , 满足  $|M| \leq k$ , 使  $\sum_{c_j \in C} \min\{d(f_i, c_j) \mid f_i \in M\}$  最小。

将设施与客户均看作空间里的点,  $F$  中的点称为设施点;  $C$  中的点称为客户点。设施  $f_i$  为客户  $c_j$  提供服务的费用  $d(f_i, c_j)$  可表示为设施点与客户点之间的距离。若任意两点之间的距离对称且满足三角不等式,则该问题又称为公制空间(metric 空间)  $k$ -median 问题,该问题为 NP-Hard 问题。1982 年, Hochbaum 等首次从求解算法角度考虑了  $k$ -median<sup>[1,2]</sup>。目前所见求解该问题的近似算法技术可归纳为线性规划<sup>[3]</sup>、过滤技术<sup>[4,5]</sup>、原始对偶方法<sup>[6]</sup>、贪心与放大技术<sup>[7]</sup>、局部搜索算法<sup>[8,9]</sup>、贪心技术<sup>[10,11]</sup>、反向贪心技术<sup>[12]</sup>以及各种算法设计技术的混合应用。2001 年, Arya 等人<sup>[8]</sup>设计了

$k$ -median 问题  $3+2/p$  ( $p \geq 1$ ) 的局部搜索算法。2005 年, 潘锐、朱大铭、马绍汉等人<sup>[9]</sup>证明了  $k$ -median 问题的一个不可近似到的一个常数界  $1+2/e$ , 除非  $NP \subseteq DTIME(n^{O(\log \log n)})$ , 给出在最大连接费用至多是最小连接费用  $\omega$  倍的情况下近似度为  $(1+\omega)/2$  的局部搜索算法。目前,关于  $k$ -median 问题小于 3 的常数近似算法未出现相关文献报道。

尽管算法工作者提出许多近似性能较好的算法,但其实现技术都比较复杂,时间复杂度高。如 Arya<sup>[8]</sup>等给出的局部搜索算法的近似性能比为  $3+2/p$  (注:  $p$  为每次迭代时交换设施点的个数),但算法需要执行多次迭代,每次迭代的时间复杂度为  $O(n^p)$ ,总的迭代次数不确定,当  $p$  较大时,算法显得不太实用。又如拉格朗日松弛算法可以获得较好的结果,但由于这类算法的求解过程涉及大规模线性规划的计算,面对大型问题实例时,利用普通 PC 运行该类算法完成相应计算任务通常极为困难。在各种实际工程环境中,如何设计出满足实际应用需求的各种性能出色的算法,引起部分算法工作者探索研究。2006 年, Chrobak 等人<sup>[12]</sup>提出用反向贪心算法求解  $k$ -median 问题。算法的时间复杂度为  $O(n^3)$ ,近似性能界于  $\Omega(\ln(n)/\ln(\ln(n)))$  与  $O(\ln(n))$  之间。2008 年, 肖进杰提出一个求解该问题的贪心算法,并证明该算法的近似度

为  $O(\ln(n/k))^{[10,11]}$ 。

给定问题实例  $I$ , 如果要求所选择的每个设施点至少服务  $\alpha n/k$  个客户点数, 其中  $\alpha$  为给定的参数并且满足  $0 < \alpha \leq 1$ , 我们将该类问题称为均衡限制的  $k$ -median 问题,  $\alpha$  称为均衡限制参数。该问题也是一个有实用价值的原问题变种, 例如在传感网结点布局中选择汇聚结点位置时, 要求在该位置上的汇聚结点需至少能够连接一定数量的传感结点。目前, 国内未出现关于该问题的近似算法相关研究文献。基于 metric 空间, 本文提出求解该问题的一个快速的反向贪心随机算法, 证明了该随机算法以较高的概率满足: 算法近似性能比的期望值为  $(3 + O(\ln(\ln(k)/\alpha)))$ 。算法基本思想描述为: 设最优解将客户点集  $C$  划分为  $k$  个子集  $C_1^*, C_2^*, \dots, C_k^*$ 。从  $C$  中随机选取  $O(\frac{k}{\alpha} \ln(k))$  个点, 记  $S$  为这些取样点的集合, 证明  $S$  包含每个子集  $C_i^*$  ( $1 \leq i \leq k$ ) 中至少一个点的概率不小于  $1/2$ ; 基于取样子集  $S$ , 从设施集合  $F$  中找出与  $S$  中每个点最近的设施点, 记这些设施点的子集为  $F_i$  ( $|F_i| \leq |S|$ )。证明了  $F_i$  中存在至多  $k$  个设施, 以这些设施服务  $C$ , 服务费用的近似性能比的期望值至多为 3; 以  $F_i$  作为初始设施点集, 通过对  $F_i$  执行反向贪心求出  $k$  个设施。

本文第 1 节给出了问题的描述及关于该问题的研究进展; 第 2 节介绍了从给定设施点集中求出一个候选设施子集的随机算法; 第 3 节介绍了对候选设施点集进行反向贪心求解  $k$ -median 问题的算法; 第 4 节介绍了实验结果。

## 2 候选设施点集

给定设施集合  $F$ 、客户点集  $C$  以及均衡限制参数  $\alpha$ , 记  $M = \{f_1^*, f_2^*, \dots, f_k^*\} \subseteq F$  为给定实例最优解所对应的设施集合,  $M$  将  $C$  划分为  $k$  个子集  $C_1^*, C_2^*, \dots, C_k^*$ , 每个子集  $C_i^*$  ( $1 \leq i \leq k$ ) 的大小均满足  $|C_i^*| \geq \alpha \times n/k$  ( $0 < \alpha \leq 1$ )。我们首先证明: 从  $C$  中随机选取一个较小的样本点集  $S$ ,  $S$  以较高的概率满足: 它包含每个  $C_i^*$  中至少一个点, 即  $|S \cap C_i^*| \geq 1$  ( $1 \leq i \leq k$ )。

**定理 1** 如果从  $C$  中均匀地随机选取  $(2 + \sqrt{3}) \frac{k}{\alpha} \ln(2k)$  个点, 设这些取样点的集合为  $S$ , 那么, 对于每个最优划分子集  $C_i^*$  ( $|C_i^*| \geq \alpha \times n/k$ ),  $|S \cap C_i^*| \geq 1$  均成立的概率至少为  $1/2$ , 即  $\Pr[|S \cap C_i^*| \geq 1] \geq 1/2$ 。

证明: 不失一般性, 设  $|C_1^*| \leq |C_2^*| \leq \dots \leq |C_k^*|$ ,  $n_i^* = |C_i^*|$ , 记  $S_i = S \cap C_i^*$ ,  $n_i = |S_i|$ 。由于  $C_i^*$  中每个点被选取的概率为  $\frac{n_i^*}{n}$ , 因此  $n_i$  的期望值为  $E(n_i) = |S| \frac{n_i^*}{n}$ , 根据 Chernoff Bounds 不等式得:

$$\Pr[n_i < \lambda |S| \frac{n_i^*}{n}] < e^{-(\frac{1-\lambda}{2})^2 |S| \frac{n_i^*}{n}} \quad (0 < \lambda < 1) \quad (1)$$

定义事件  $A_i$  为  $n_i$  小于  $(|S| \frac{n_i^*}{n})$  的事件, 则:

$$\Pr[\exists i n_i < \lambda |S| \frac{n_i^*}{n}] = \Pr[\bigcup_{i=1}^k A_i] \leq \sum_{i=1}^k \Pr(A_i) \quad (2)$$

由于  $|C_i^*| \geq \alpha \times n/k$  ( $0 < \alpha \leq 1$ ), 因此  $\frac{n_i^*}{n} \geq \frac{\alpha}{k}$ 。选取  $\lambda = (2 - \sqrt{3})$ ,  $|S| = (2 + \sqrt{3}) \frac{k}{\alpha} \ln(2k)$ , 将它们代入式 (1) 可得:

$$\Pr[\exists i n_i < \lambda |S| \frac{n_i^*}{n}] < \frac{1}{2}。因此:$$

$$\Pr[\forall i n_i \geq \lambda |S| \frac{n_i^*}{n}] = 1 - \Pr[\exists i n_i < \lambda |S| \frac{n_i^*}{n}] \geq \frac{1}{2}$$

进一步, 将选取的  $\lambda$  值以及  $|S|$  值代入该结论可得:  $n_i \geq \lambda |S| \frac{n_i^*}{n} \geq \ln(2k)$  成立的概率至少为  $\frac{1}{2}$ 。显然当  $k > 1$  时,  $n_i > 1$ , 即  $S$  包含每个最优子集  $C_i^*$  中至少一个点的概率大于或等于  $\frac{1}{2}$ 。

定理 1 表明, 如果已知每个最优划分子集  $C_i^*$  的大小至少为  $\alpha \times n/k$  个点, 只需从客户点集  $C$  中随机选取  $(2 + \sqrt{3}) \frac{k}{\alpha} \ln(2k)$  个点, 由这些点组成的集合记为  $S$ , 则  $S$  包含每个子集  $C_i^*$  ( $1 \leq i \leq k$ ) 中至少一个点的概率不小于  $1/2$ 。

针对上述所求的样本集  $S$ , 对每个客户点  $x_i \in S$  ( $1 \leq i \leq |S|$ ), 记  $f_i \in F$  为距离  $x_i$  最近的设施点, 由所有设施点  $f_i$  构成的集合记为  $R$ , 显然  $|R| \leq |S|$ 。对于给定的实例  $I$ , 设  $cost(R)$  为设施子集  $R \subseteq F$  服务客户点集  $C$  的费用,  $OPT(I)$  为实例  $I$  的最优解值。下面证明在  $R$  中存在至多  $k$  个设施, 以这些设施服务客户点集  $C$ , 服务费用的近似性能比的期望值至多为 3。在给出该定理的证明之前, 首先给出下述引理。

**引理 1** 给定问题实例  $I$ , 设  $OPT(I)$  最优解中服务  $C_i^*$  的设施为  $f_i^*$ 。如果从  $C_i^*$  中均匀地随机选取一个点  $x$ , 则  $x$  与  $f_i^*$  距离的期望值为:

$$E(d(f_i^*, x)) = (\sum_{x \in C_i^*} d(x, f_i^*)) / |C_i^*|$$

证明: 由于  $C_i^*$  中每个点被选取的概率为  $1/|C_i^*|$ , 因此该结论显然成立。

**定理 2** 如果样本点集  $S$  满足:  $\forall i |S \cap C_i^*| \geq 1$  ( $1 \leq i \leq k$ ), 在由  $S$  所确定的设施集合  $R$  中, 则存在一设施子集  $R_k \subseteq R$  ( $|R_k| \leq k$ ), 其服务费用的期望值满足:  $E(cost(R_k)) \leq 3OPT(I)$ 。

证明: 记  $S$  中  $k$  个客户点  $x_1, x_2, \dots, x_k$  满足:  $x_i \in S \cap C_i^*$  ( $1 \leq i \leq k$ ), 即  $x_i$  是最优划分子集  $C_i^*$  中的一个点。对每个客户点  $x_i$ , 设施集合  $F$  中与  $x_i$  最近的设施记为  $f_i$ , 由所有  $f_i$  构成  $R$  的子集记为  $R_k$ , 显然  $|R_k| \leq k$ 。  $R_k$  服务客户点集  $C$  的费用记为  $cost(R_k)$ 。

对客户点集  $C$  中的所有点, 考虑一种特殊的设施分配策略, 该策略用设施  $f_i$  服务  $C_i^*$ , 即用设施  $f_i$  代替  $f_i^*$  服务最优划分子集  $C_i^*$  中的所有客户点。该特殊分配产生的服务费用记为  $cost(R_k')$ 。

$$\begin{aligned} cost(R_k') &= \sum_{i=1}^k \sum_{y \in C_i^*} d(y, f_i) \\ &\leq \sum_{i=1}^k \sum_{y \in C_i^*} (d(y, x_i) + d(x_i, f_i)) \\ &\leq \sum_{i=1}^k \sum_{y \in C_i^*} (d(y, f_i^*) + d(x_i, f_i^*) + d(x_i, f_i)) \\ &\leq \sum_{i=1}^k \sum_{y \in C_i^*} (d(y, f_i^*) + 2d(x_i, f_i^*)) \\ &= \sum_{i=1}^k \sum_{y \in C_i^*} (d(y, f_i^*) + 2 \sum_{i=1}^k |C_i| d(x_i, f_i^*)) \\ &= 3 \sum_{i=1}^k \sum_{y \in C_i^*} d(y, f_i^*) \quad (\text{根据引理 1}) \\ &= 3OPT(I) \end{aligned}$$

$cost(R_k)$  所对应的分配策略是将  $C$  中每个客户点  $y$  分配给离  $y$  最近的设施, 由此可得  $cost(R_k) \leq cost(R_k')$ , 所以结论

成立。

如果样本点集  $S$  满足  $\forall i |S \cap C_i^*| \geq 1$ , 从  $R$  中选取不同的  $k$  个点, 计算这  $k$  个点服务客户点集  $C$  的费用, 可求出一个可行解。枚举  $R$  中所有可能的  $k$  个点, 取费用最小的可行解作为算法的最终解。定理 2 表明: 该策略所求解的近似性能比的期望值至多为 3。根据定理 1 有, 该结论成立的概率至少为  $1/2$ 。

根据定理 1,  $|S| = (2 + \sqrt{3}) \frac{k}{\alpha} \ln(2k)$ , 显然  $|R| \leq |S| = (2 + \sqrt{3}) \frac{k}{\alpha} \ln(2k)$ 。从  $R$  中枚举所有可行解的个数至多为  $C_{|R|}^k$ 。由于  $C_{|R|}^k \leq \left(\frac{|R|e}{k}\right)^k$  [13], 将  $|R|$  值代入该式可得枚举子集个数为  $O\left(\left(\frac{(2 + \sqrt{3})e \ln(k)}{\alpha}\right)^k\right)$ , 因此, 当  $k$  较大时, 采用枚举  $R$  中所有可能  $k$  个点的策略, 其时间复杂度较高、实用性较差。

### 3 反向贪心随机算法

给定设施集合  $F$  以及客户点集  $C$ , 从  $C$  中均匀地随机选取部分点, 记  $S$  为这些点的集合。对于每个客户点  $x_i \in S (1 \leq i \leq |S|)$ , 在  $F$  中找出离  $x_i$  最近的设施点  $f_i$ , 设  $R$  为所有设施点  $f_i$  构成的集合。仍记  $\text{cost}(R)$  为设施集合  $R$  服务  $C$  的费用。算法的基本思想为: 从  $R$  中寻找设施  $r$ , 使之满足  $\text{cost}(R \setminus \{r\}) - \text{cost}(R)$  的值最小, 执行  $R = R - \{r\}$ 。重复这种删除过程直到  $|R| = k$  为止。该算法详细描述如下。

#### 算法 1 $k$ -median 问题反向贪心随机算法

输入: 设施集合  $F$ , 客户集合  $C$ , 均衡限制参数  $\alpha$

输出: 服务费用

1. 从客户集合  $C$  中均匀地随机选取  $(2 + \sqrt{3}) \frac{k}{\alpha} \ln(2k)$  个点, 记为  $S$ 。
2. 对于  $S$  中每一个点  $x_i$ , 计算  $F$  中与  $x_i$  最近的点  $f_i$ , 构成设施集合  $R$ 。
3. 计算  $R$  服务  $C$  的费用  $\text{cost}(R)$
4. While  $|R| > k$  时执行 5-6 步
5. 对每个  $r \in R$ , 计算  $\text{cost}(R \setminus \{r\})$
6. 选择满足  $\min(\text{cost}(R \setminus \{r\}) - \text{cost}(R))$  的设施  $r_0$   
 $R = R \setminus \{r_0\}, \text{cost}(R) = \text{cost}(R \setminus \{r_0\})$
7. 计算  $R$  服务  $C$  的费用

算法第 2 步从设施集合  $F$  中寻找离  $S$  中每个点最近的设施, 时间复杂度为  $O(|S| \times |F|)$ 。算法第 3 步计算  $\text{cost}(R)$  时, 时间复杂度为  $O(|C| \times |R|)$ 。每执行一次循环, 算法第 5 步对每个设施  $r \in R$ , 计算  $\text{cost}(R \setminus \{r\})$  时, 需要将设施  $r$  所服务的每个客户点重新分配到  $R \setminus \{r\}$  中的某个设施, 因此算法第 5 步循环一次的时间复杂度为  $O(|R| \times |C|)$ 。算法共循环  $|R| - k$  步, 每循环一次,  $|R|$  的大小都要减 1, 所以算法第 5 步总的时间复杂度为  $O[|R| + (|R| - 1) + \dots + k] \times |C|$ 。这样算法总的时间复杂度为:  $O(|S| \times |F|) + O[|R| + (|R| - 1) + \dots + k] \times |C|$ 。设  $n = |C|, m = |F|$ , 显然  $|R| \leq |S|$ , 将  $n$  和  $m$  分别代入该式并进行合并, 由此得出上述算法的时间复杂度为  $O\left[\left(\frac{k}{\alpha} \ln(k)\right)^2 (n + m)\right]$ 。

为讨论算法的近似性能, 在给出结论之前, 首先给出相关引理及其证明。

**引理 2**  $1 + 1/2 + 1/3 + \dots + 1/n = \ln(n) + \gamma$  ( $\gamma$  为小于 1 的欧拉常数)。

不失一般性, 设上述算法从  $R$  中依次删除的设施为  $r_0$ ,

$\dots, r_{|R|-k+1}$ 。

#### 引理 3

$$\text{cost}(R) \leq \text{cost}(R \setminus \{r_0\}) \leq \dots \leq \text{cost}(R \setminus \{r_0, \dots, r_{|R|-k+1}\})$$

证明: 根据定义,  $\text{cost}(R)$  表示设施集合  $R$  服务客户点集  $C$  的费用, 此时  $C$  中每一个客户都被其距离最近的设施所服务, 服务费用最小。由算法第 5 步知, 下一步要从这  $|R|$  个设施中选择一个设施  $r_0$ , 它满足  $\text{cost}(R \setminus \{r_0\}) - \text{cost}(R)$  的值最小。设  $r_0$  服务的客户点集为  $N(r_0)$ , 则对于每个客户点  $x \in N(r_0)$ , 此时只能分配给剩下的  $|R| - 1$  个设施, 且  $x$  与新设施的距离肯定大于等于  $x$  与  $r_0$  的距离, 所以  $\text{cost}(R) \leq \text{cost}(R \setminus \{r_0\})$ 。

同理可得:

$$\text{cost}(R \setminus \{r_0\}) \leq \dots \leq \text{cost}(R \setminus \{r_0, r_1, \dots, r_{|R|-k+1}\})$$

因此结论成立。

对于给定实例  $I$ , 定义  $A(I)$  为反向贪心随机算法针对实例  $I$  的目标函数值,  $OPT(I)$  为相应实例  $I$  的最优值, 显然  $A(I) = \text{cost}(R \setminus \{r_0, r_1, \dots, r_{|R|-k+1}\})$ 。设  $M \subseteq F$  为给定实例  $I$  最优解所对应的设施子集, 则  $|M| \leq k$ , 对于  $M$  中每个设施点  $f_i^*$ , 从  $R$  中选择离  $f_i^*$  最近的一个设施点  $f_i'$ , 记  $Q = \{f_1', f_2', \dots, f_{|M|}'\}$ , 则下述不等式成立。

#### 引理 4

$$\sum_{r \in R \setminus Q} [\text{cost}(R \setminus \{r\}) - \text{cost}(R)] \leq \text{cost}(Q) - \text{cost}(R)$$

证明: 设  $x \in C$ , 定义  $d(x, F) = \min_{f \in F} (d(x, f))$ 。对于任意  $r \in R$ , 记  $N(r) = \{x | x \in C \wedge d(x, r) = d(x, R)\}$ 。先分析引理不等式的左端:

$$\begin{aligned} & \sum_{r \in R \setminus Q} [\text{cost}(R \setminus \{r\}) - \text{cost}(R)] \\ &= \sum_{r \in R \setminus Q} \sum_{x \in N(r)} [d(x, R \setminus \{r\}) - d(x, R)] \end{aligned} \quad (3)$$

对于不等式右端:

$$\begin{aligned} \text{cost}(Q) - \text{cost}(R) &= \sum_{x \in C} [d(x, Q) - d(x, R)] \\ &= \sum_{r \in Q} \sum_{x \in N(r)} [d(x, Q) - d(x, R)] + \sum_{r \in R \setminus Q} \sum_{x \in N(r)} [d(x, Q) - d(x, R)] \end{aligned}$$

由于  $r \in Q, Q \subseteq R$ , 以及  $r$  为  $R$  中距离  $x$  最近的设施, 因此  $d(x, Q) = d(x, R) = d(x, r)$ , 则上式前半部分值为 0。由此得出:

$$\text{cost}(Q) - \text{cost}(R) = \sum_{r \in R \setminus Q} \sum_{x \in N(r)} [d(x, Q) - d(x, R)] \quad (4)$$

由于  $Q \subseteq R$ , 因此  $C$  中的任一点  $x$  必满足:  $d(x, Q) \geq d(x, R)$  以及  $d(x, R \setminus Q) \geq d(x, R)$ 。

对于式(3)和式(4)中的同一个设施  $r$ , 由  $r \in R \setminus Q$  以及  $Q \subseteq R$  可得  $Q \subseteq R \setminus \{r\}$ , 所以对于每个设施  $r$  所服务的客户  $x \in N(r)$ ,  $d(x, R \setminus \{r\}) \leq d(x, Q)$  成立。进一步可得,  $d(x, R \setminus \{r\}) - d(x, R) \leq d(x, Q) - d(x, R)$  成立。考虑所有  $x$  并进行相加, 可推出上述不等式成立。

#### 引理 5

$$\text{cost}(Q) - \text{cost}(R) \leq 2\text{cost}(M)$$

证明: 对于客户点集  $C$  中任一点  $x$ , 分别设  $R$  和  $M$  中离  $x$  最近的设施为  $f_{R(x)}$  和  $f_{M(x)}$ , 记  $Q$  中距离  $f_{M(x)}$  最近的设施为  $f_{Q(x)}$ , 根据  $Q$  的定义,  $d(f_{M(x)}, f_{Q(x)}) \leq d(f_{M(x)}, f_{R(x)})$ 。

另一方面:

$$\begin{aligned} d(x, f_{Q(x)}) &\leq d(x, f_{M(x)}) + d(f_{M(x)}, f_{Q(x)}) \\ &\leq d(x, f_{M(x)}) + d(f_{M(x)}, f_{R(x)}) \end{aligned}$$

$$\begin{aligned} &\leq d(x, f_{M(x)}) + d(x, f_{M(x)}) + d(x, f_{R(x)}) \\ &\quad (\text{三角不等式}) \\ &\leq 2d(x, f_{M(x)}) + d(x, f_{R(x)}) \end{aligned}$$

由此可得,  $d(x, f_{Q(x)}) - d(x, f_{R(x)}) \leq 2d(x, f_{M(x)})$ 。

考虑  $C$  中所有  $x$  并对不等式相加可得本引理成立。

**定理 3** 给定问题实例  $I$ , 反向贪心随机算法以较高的概率满足其近似性能比的期望值为  $(3 + O(\ln(\ln(k)/\alpha)))$ , 其中  $\alpha$  为均衡限制参数。

下面利用文献[12]的思想和方法给出该定理的证明。

证明: 定义  $R_j$  为算法执行第 5 步  $R = R \setminus \{r_0\}$  使得  $R$  的大小为  $j$  ( $|R| = j$ ) 时, 设施点的集合。由引理 3, 当  $R$  由  $R_j$  变为  $R_{j-1}$  时, 会导致服务费用增大, 下面分析费用增长量上界。

$$\text{cost}(R_{j-1}) - \text{cost}(R_j) \leq \min_{r \in R_j \setminus Q} (\text{cost}(R_j \setminus \{r\}) - \text{cost}(R_j)) \quad (5)$$

$$\leq \frac{1}{|R_j \setminus Q|} \sum_{r \in R_j \setminus Q} [\text{cost}(R_j \setminus \{r\}) - \text{cost}(R_j)] \quad (6)$$

$$\leq \frac{1}{j - k} \sum_{r \in R_j \setminus Q} [\text{cost}(R_j \setminus \{r\}) - \text{cost}(R_j)] \quad (7)$$

$$\leq \frac{1}{j - k} [\text{cost}(Q) - \text{cost}(R_j)] \quad (\text{根据引理 4})$$

$$\leq \frac{2}{j - k} \text{cost}(M) \quad (\text{根据引理 5})$$

其中, 式(5)是根据算法第 4 步得出的结论; 式(6)是根据  $\text{cost}(R_j \setminus \{r\}) - \text{cost}(R_j)$  的最小值必小于或等于相应的平均值而得出的结论; 由  $|Q| \leq k$  可推出式(7)成立。

整理上式得:

$$\text{cost}(R_{j-1}) - \text{cost}(R_j) \leq \frac{2}{j - k} \text{cost}(M) \quad (8)$$

分别取  $j = |S|, |S| - 1, \dots, k$ , 并对式(8)左右两端相加可得:

$$\begin{aligned} &\text{cost}(R_k) - \text{cost}(R_{|S|}) \\ &\leq \left( \frac{2}{|S| - k} + \frac{2}{|S| - k - 1} + \dots + \frac{2}{k + 1} \right) \text{cost}(M) \end{aligned}$$

由引理 2 可得:

$$\left( \frac{2}{|S| - k} + \frac{2}{|S| - k - 1} + \dots + \frac{2}{k + 1} \right) \text{cost}(M) \leq 2 \ln(|S|/k - 1) \text{cost}(M)$$

又由引理 3 及定理 2 得:

$$E(\text{cost}(R_{|S|})) \leq 3 \text{cost}(M)$$

所以:

$$\begin{aligned} E(\text{cost}(R_k)) &\leq [2 \ln(|S|/k) + 3] \text{cost}(M) \\ &= [2 \ln \frac{(2 + \sqrt{3}) \ln(2k)}{\alpha} - 1 + 3] \text{cost}(M) \end{aligned}$$

由算法结束条件可知,  $A(I) = \text{cost}(R_k)$ , 将其代入上式并进行简化可得:

$$E(A(I)) \leq (3 + O(\ln(\ln(k)/\alpha))) \text{OPT}(I)$$

根据定理 1, 集合  $S$  满足  $\forall i | S \cap C_i^* | \geq 1$  均成立的概率至少为  $1/2$ 。因此, 运算本算法一次, 则至少以  $1/2$  的概率求到近似性能比的期望值为  $(3 + O(\ln(\ln(k)/\alpha))) \text{OPT}(I)$ 。如果重复运算本算法  $\log n$  次, 然后取最好的一次解值, 则至少要以  $1 - \frac{1}{n}$  的概率得出该结论。

## 4 实验结果

选取 ORLIB 数据库中的部分数据集, 测试所提反向贪心

算法的有效性。鉴于算法的随机性, 重复运算该算法多次, 取最小的值作为算法的最终运算结果。表 1 列出算法运算 10 次后针对不同的  $\alpha$  值所求的结果。

表 1 算法近似性能实验数据

| 数据集      | 最优值   | $\alpha=0.3$ |          | $\alpha=0.5$ |          | $\alpha=0.7$ |          |
|----------|-------|--------------|----------|--------------|----------|--------------|----------|
|          |       | Value        | Ratio    | Value        | Ratio    | Value        | Ratio    |
| pmed1    | 5819  | 5819         | 1        | 5856         | 1.006358 | 5894         | 1.012889 |
| pmed6    | 7824  | 7869         | 1.005752 | 7869         | 1.005752 | 7975         | 1.0193   |
| pmed7    | 5631  | 5735         | 1.018469 | 5690         | 1.010478 | 5713         | 1.014562 |
| pmed11   | 7696  | 7754         | 1.007536 | 7750         | 1.007017 | 7873         | 1.022999 |
| pmed12   | 6634  | 6723         | 1.013416 | 6769         | 1.02035  | 6818         | 1.027736 |
| pmed16   | 8162  | 8206         | 1.005391 | 8209         | 1.005758 | 8518         | 1.043617 |
| pmed17   | 6999  | 7121         | 1.017431 | 7090         | 1.013002 | 7253         | 1.036291 |
| pmed21   | 9138  | 9493         | 1.038849 | 9323         | 1.020245 | 9353         | 1.023528 |
| pmed22   | 8579  | 8693         | 1.013288 | 8746         | 1.019466 | 8758         | 1.020865 |
| pmed26   | 9917  | 9938         | 1.002118 | 10029        | 1.011294 | 10426        | 1.051326 |
| pmed27   | 8307  | 8547         | 1.028891 | 8456         | 1.017937 | 8537         | 1.027687 |
| pmed31   | 10086 | 10152        | 1.006544 | 10221        | 1.013385 | 10436        | 1.034702 |
| pmed32   | 9297  | 9432         | 1.014521 | 9463         | 1.017855 | 9699         | 1.04324  |
| pmed35   | 10400 | 10671        | 1.026058 | 10636        | 1.022692 | 10931        | 1.051058 |
| pmed36   | 9934  | 10160        | 1.02275  | 10132        | 1.019932 | 10117        | 1.018422 |
| pmed38   | 11060 | 11146        | 1.007776 | 11357        | 1.026854 | 11583        | 1.047288 |
| pmed39   | 9423  | 9646         | 1.023665 | 9591         | 1.017829 | 9736         | 1.033217 |
| 算法平均近似性能 |       | 1.01485      |          | 1.015071     |          | 1.031101     |          |

针对不同  $\alpha$  值, 反向贪心算法需从客户点集  $C$  中随机选取部分点, 基于取样点计算出候选设施点集, 最后对候选设施点集反向贪心求出  $k$  个中心点。由于算法的随机性,  $C$  中存在部分点可能被重复选取, 导致每次选取时取样点集的大小可能不相等, 表 2 列出算法运算 10 次后, 从客户点集  $C$  中随机选取的平均点数。

表 2 针对不同  $\alpha$  值算法所需取样点数

| 数据集    | 客户数 | k 值 | $\alpha=0.3$ |         | $\alpha=0.5$ |         | $\alpha=0.7$ |         |
|--------|-----|-----|--------------|---------|--------------|---------|--------------|---------|
|        |     |     | 取样数          | 减小比例(%) | 取样数          | 减小比例(%) | 取样数          | 减小比例(%) |
| pmed1  | 100 | 5   | 74           | 26.00   | 57           | 43.00   | 43           | 57.00   |
| pmed6  | 200 | 5   | 96           | 52.00   | 67           | 66.50   | 50           | 75.00   |
| pmed7  | 200 | 10  | 108          | 46.00   | 135          | 32.50   | 111          | 44.50   |
| pmed11 | 300 | 5   | 112          | 62.67   | 74           | 75.33   | 55           | 81.67   |
| pmed12 | 300 | 10  | 212          | 29.33   | 154          | 48.67   | 124          | 58.67   |
| pmed16 | 400 | 5   | 115          | 71.25   | 77           | 80.75   | 55           | 86.25   |
| pmed17 | 400 | 10  | 242          | 39.50   | 171          | 57.25   | 131          | 67.25   |
| pmed21 | 500 | 5   | 120          | 76.00   | 78           | 84.40   | 56           | 88.80   |
| pmed22 | 500 | 10  | 256          | 48.80   | 178          | 64.40   | 133          | 73.40   |
| pmed26 | 600 | 5   | 125          | 79.17   | 78           | 87.00   | 57           | 90.50   |
| pmed27 | 600 | 10  | 276          | 54.00   | 188          | 68.67   | 138          | 77.00   |
| pmed31 | 700 | 5   | 126          | 82.00   | 78           | 88.86   | 56           | 92.00   |
| pmed32 | 700 | 10  | 288          | 58.86   | 193          | 72.43   | 141          | 79.86   |
| pmed35 | 800 | 5   | 128          | 84.00   | 81           | 89.88   | 58           | 92.75   |
| pmed36 | 800 | 10  | 294          | 63.25   | 193          | 75.88   | 143          | 82.13   |
| pmed38 | 900 | 5   | 128          | 85.78   | 79           | 91.22   | 57           | 93.67   |
| pmed39 | 900 | 10  | 304          | 66.22   | 196          | 78.22   | 144          | 84.00   |

表 2 中, 减小比例 = (客户数 - 取样数) / 客户数 \* 100。基于表 2 可知, 算法在确定候选设施点集时, 仅从少量随机选取的客户点集中计算候选设施点集, 然后对候选设施点集进行反向贪心, 找出  $k$  个中心点。与文献[11, 12]相比, 后者以整个设施点集作为候选点集, 执行反向贪心求出  $k$  个设施点。因此, 每执行一次本文反向贪心算法, 花费时间较少, 算法运算效率高。

**结束语** 本文分析了基于取样策略的  $k$ -median 问题的反向贪心算法, 给出  $k$ -median 问题的一个反向贪心随机算法, 证明了该算法的近似性能比的期望值。算法的基本思想

是求一个取样点集  $S$ , 要求  $S$  包含客户点集  $C$  的每个最优划分分子集中至少一个点, 根据  $S$  再进一步确定一个设施子集  $F$ , 利用反向贪心算法从  $F$  中找出  $k$  个点服务  $C$ 。本文在求解取样点集时基于均衡限制参数  $\alpha$ , 算法近似性能比的期望值依赖于  $\alpha$ 。 $\alpha$  越小, 则  $|S|$  越大, 算法近似性能比的期望值越高;  $\alpha$  越大, 则  $|S|$  越小, 算法近似性能比的期望值越小, 因此该算法适用于各划分分子集大小相差不太大的情况。

## 参考文献

- [1] Balinski M L. On finding integer solutions to linear programs [C]//Proceedings of IBM scientific computing symposium on combinatorial problems, 1966;225-248
- [2] Hochbaum D S. Heuristics for the fixed cost median problem [J]. Mathematical Programming, 1982, 22; 148-162
- [3] Lin J H, Vitter J S.  $\epsilon$ -approximation s with minimum constraint violation[C]//Proceedings of stoc'92. 1992;771-782
- [4] Arora S, Raghavan P, Rao S. Approximation schemes for Euclidean  $k$ -median and related problems[C]//Proceedings of stoc' 98. 1998;106-113
- [5] Charikar M, et al. A constant approximation algorithm for the  $k$ -median problem[C]//Proceedings of stoc'1999. Atlanta GA USA, 1999

- [6] Jain K, Vazirani V V. Primal-dual approximation algorithms for metric facility location and  $k$ -median problems[C]//Proceedings of focs'99. 1999;2-13
- [7] Charikar M, Guha S. Improved combinatorial algorithms for the facility location and  $k$ -median problems [C] // Proceedings of focs'99. 1999;1-10
- [8] Arya V, et al. Local search heuristics for  $k$ -median and facility location problems [C] // Proceedings of stoc' 2001. Hersonis-sons, Crete, Greece, 2001
- [9] 潘锐, 朱大铭, 等.  $k$ -median 近似计算复杂度与局部搜索近似算法分析[J]. 软件学报, 2005, 16(3): 392-399
- [10] 肖进杰, 范辉, 等. 贪心算法求解  $k$ -median 问题[J]. 计算机工程与应用, 2006, 42(3): 57-58
- [11] 肖进杰, 谢青松.  $k$ -median 问题贪心近似算法的分析与实验[J]. 计算机工程, 2008, 34(22): 213-215
- [12] Chrobak M, Kenyon C, Young N. The reverse greedy algorithm for the metric  $k$ -median problem [J]. Information Processing Letters, 2006, 97; 68-72
- [13] Motawani R, Raghavan P. Randomized Algorithms. Cambridge University Press[M]. Cambridge, UK, 1995
- [14] Chen K. A constant factor approximation algorithm for  $k$ -median clustering with outliers[C]//Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms. 2008

(上接第 231 页)

1, 3, 4, 6 这 4 种模型都可以适用。

**结束语** 本文系统地分析、比较了 6 种主要基于覆盖的粗糙集模型, 找到了各模型之间的区别和联系。各模型之间要么明确可比, 要么明确不可比。可比的给出了定理证明, 不可比的举出了反例。通过比较多种模型之间关系, 可以加深对这些模型的理解, 使得读者方便地根据不同应用中根据需要选择不同的模型。

今后将在本文工作的基础上, 对于可比较的模型研究其相等的充要条件, 对于不可比较的模型研究其可比必须满足的条件; 也可以从最小覆盖粗糙集模型和最大覆盖粗糙集模型的角度出发, 构造出满足需要的覆盖粗糙集模型。

## 参考文献

- [1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science, 1982(11); 341-356
- [2] 邓维斌, 王国胤, 洪智勇. 基于粗糙集的加权朴素贝叶斯邮件过滤方法[J]. 计算机科学, 2011, 38(2): 218-221
- [3] Qian Y H, Liang J Y, Dang C Y. Converse approximation and rule extraction from decision tables in rough set theory[J]. Computers & Mathematics with Applications, 2008, 55(8): 1754-1765
- [4] Shi Z H, Gong Z T. The further investigation of covering-based rough sets: Uncertainty characterization, similarity measure and generalized models[J]. Information Science, 2010, 180(19): 3745-3763
- [5] Dembczyński K, Greco S, Slowiński R. Rough set approach to multiple criteria classification with imprecise evaluations and assignments[J]. European Journal of Operation Research, 2009, 198(2): 626-636
- [6] Kryszkiewicz M. Rough set approach to incomplete information systems [J]. Information Science, 1998, 112; 39-49

- [7] Stefanowski J, Tsoukias A. On the extension of rough sets under incomplete information [C] // 7<sup>th</sup> International Workshop on New Directions in Rough Sets, Data Mining, and Granular-soft Computing. 1998; 422-450
- [8] Zakowski W. Approximations in the space  $(u, \pi)$  [J]. Demonstration Mathematics, 1983, 16; 761-769
- [9] Samanta P, Chakraborty M K. Covering based approaches to rough sets and implication lattices [C]//RSFD GrC. 2009; 127-134
- [10] Zhu W, Wang F Y. Properties of the First Type of Covering-based Rough Sets [C]//Proceedings of DM Workshop 06 (IC-DM 06). Hong Kong, China, December 2006; 407-411
- [11] Zhu W. Properties of the second type of covering-based rough sets [C]//Workshop Proceedings of GrC&BI 06 (IEEE WI 06). Hong Kong, China, December 2006; 494-497
- [12] Pomykala J A. Approximation operations in approximation space [J]. Bulletin of the Polish Academy of Sciences, 1987, 35 (9/10): 653-662
- [13] Tsang E, Cheng D, Lee J, et al. On the upper approximations of covering generalized rough sets [C]//Proceedings of the 3rd International Conference Machine Learning and Cybernetics. 2004; 4200-4203
- [14] Zhu W, Wang F Y. On three types of covering rough sets[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(8): 1131-1144
- [15] Zhu W, Wang F Y. A new type of covering rough sets [C]//IEEE IS 2006. London, September 2006; 444-449
- [16] Zhu W. Topological approaches to covering rough sets[J]. Information Science, 2007, 177 (6): 1499-1508
- [17] Zhu W. Relationship between generalized rough sets based on binary relation and covering [J]. Information Science, 2009, 179 (1): 210-225