ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Июль—август 2008. Том 15, № 4. 84–91

УДК 519.176

ПРИБЛИЖЁННЫЙ АЛГОРИТМ ДЛЯ ИЕРАРХИЧЕСКОЙ ЗАДАЧИ О НАЗНАЧЕНИЯХ*)

В. В. Шенмайер

Аннотация. Иерархическая задача о назначениях заключается в отыскании иерархической последовательности решений задачи о k-медиане возрастающей мощности. Лучший известный алгоритм для данной задачи в общем метрическом случае имеет относительную оценку точности $20,71$. Рассмотрен случай, когда клиенты и предприятия расположены в точках вещественной прямой, а также случай евклидова пространства. Предлагается алгоритм с точностью, равной 8 в случае вещественной прямой и $8 + 4\sqrt{2}$ (приблизительно $13,66$) — в евклидовом случае.

Ключевые слова: задача о k-медиане, иерархическая кластеризация, задача о последовательности медиан, приближённый алгоритм, точность алгоритма.

Введение

Иерархическая задача о назначениях (hierarchical median problem) является вариантом задачи о k-медиане и отличается от неё тем, что вместо решения для одного наперёд заданного k в ней требуется построить последовательность решений (точных либо приближённых) для всех k, обладающую свойством иерархичности.

Пусть заданы: конечное множество клиентов (потребителей) C; конечное множество предприятий F; расстояние $d(u, f) \geqslant 0$, определённое для каждого клиента u и предприятия f; а также вес $w(u) \geqslant 0$, определённый для каждого клиента u. В метрическом случае клиенты и предприятия являются точками некоторого метрического пространства с соответствующей функцией расстояния.

Назначением будем называть произвольное отображение множества клиентов в некоторое подмножество предприятий, называемых *открытыми*. Стоимость $\mathrm{cost}(\alpha)$ произвольного назначения α равна взвешенной сумме расстояний $\sum\limits_{u \in C} w(u)\, d(u, \alpha(u))$. Заметим, что в задаче о k-медиане

*)Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проект 08–01–00516–а).

требуется найти назначение минимальной стоимости, использующее k открытых предприятий.

Последовательность назначений $\alpha_1, \alpha_2, \ldots, \alpha_n$, где n — мощность множества F, называется *иерархической*, если назначение α_1 назначает всех клиентов на некоторое предприятие f_1 и для каждого $k < n$ назначение α_{k+1} получается из предыдущего переназначением части клиентов, назначенных на одно из предприятий, на новое предприятие $f_{k+1} \notin \{f_1, \ldots, f_k\}$.

Областью обслуживания произвольного предприятия x называется множество клиентов, назначенных на x. Заметим, что любая иерархическая последовательность длины n характеризуется тем, что при любых $k < \ell \leqslant n$

(H1) множество открытых предприятий на шаге k является собственным подмножеством множества открытых предприятий на шаге ℓ;

(H2) область обслуживания любого открытого предприятия x на шаге k равна объединению областей обслуживания некоторых предприятий, включая x, на шаге ℓ.

Иерархическая задача о назначениях заключается в отыскании иерархической последовательности назначений $\alpha_1, \alpha_2, \ldots, \alpha_n$ такой, что стоимость каждого назначения α_k, $k = 1, 2, \ldots, n$, равна стоимости оптимальной k-медианы:

$$\operatorname{cost}(\alpha_k) = \min_{\alpha \in A_k} \operatorname{cost}(\alpha),$$

где A_k — множество назначений, использующих k открытых предприятий.

Поскольку в большинстве случаев искомой последовательности оптимальных назначений, по-видимому, не существует, имеет смысл говорить о приближённых решениях задачи. Говорим, что решение $\alpha_1, \alpha_2, \ldots, \alpha_n$ иерархической задачи о назначениях имеет *точность* (competitive ratio) δ, если стоимость каждого из назначений α_k, $k = 1, 2, \ldots, n$, отличается от стоимости оптимальной k-медианы не более чем в δ раз.

Отметим две близкие задачи: задачу о последовательности медиан (incremental median problem) и иерархическую задачу кластеризации (hierarchical clustering problem). Первая отличается от иерархической задачи о назначениях отсутствием условия (H2), вторая — отсутствием условия (H1). В [1, 2] получены наилучшие алгоритмы для задачи о последовательности медиан, в [3] — наилучший алгоритм для иерархической задачи кластеризации (с более естественной для неё минимаксной целевой функцией).

Первый алгоритм для иерархической задачи о назначениях, имеющий в метрическом случае константную оценку точности, был получен в работе [6]. Наилучший известный алгоритм, решающий данную задачу, имеет оценку точности $20,71$ [4]. Полиномиальный вариант этого алгоритма имеет вдвое худшую оценку [4].

Рассмотрим два частных случая: случай вещественной прямой и случай евклидова векторного пространства. Предлагается алгоритм, решающий иерархическую задачу о назначениях в случае прямой с точностью 8, а в евклидовом случае — с точностью $8+4\sqrt{2}$ (приблизительно $13,66$). В общем метрическом случае алгоритм имеет точность 24.

1. Описание алгоритма

Алгоритм начинает свою работу с нахождения точных решений $\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*$ задачи о k-медиане, $k = 1, 2, \ldots, n$. На следующем шаге определяется множество K, состоящее из непродолжаемой последовательности номеров $i_1, i_2, \ldots, i_t$, в которой $i_1 = 1$ и каждый следующий номер i_{k+1}, $k = 1, 2, \ldots, t-1$, является минимальным номером таким, что

$$\operatorname{cost}(\alpha_{i_{k+1}}^*) \leqslant \operatorname{cost}(\alpha_{i_k}^*) \,/\, 2\,(\Omega + 1),$$

где Ω — константа, зависящая от рассматриваемого случая:

$$\Omega = \begin{cases} 1 & \text{в случае прямой,} \\ \sqrt{2} & \text{в случае евклидова пространства,} \\ 2 & \text{в общем метрическом случае.} \end{cases}$$

Далее находится частичное решение задачи, соответствующее номерам из множества K, т. е. последовательность назначений $\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_t}$, использующих соответственно $i_1, i_2, \ldots, i_t$ открытых предприятий, удовлетворяющая свойствам иерархичности (H1), (H2). Данная последовательность строится рекурсивно в обратном порядке: сначала определяется назначение α_{i_t}, затем назначение $\alpha_{i_{t-1}}$ и т. д. Назначение α_{i_t} полагается равным $\alpha_{i_t}^*$, и при $k = t-1, \ldots, 1$ назначение α_{i_k} определяется следующим образом. Рассмотрим вспомогательное множество клиентов $C_{i_{k+1}}$, состоящее из i_{k+1} клиентов, соответствующих открытым предприятиям назначения $\alpha_{i_{k+1}}$, так что каждый клиент расположен в той же точке, что и соответствующее открытое предприятие, а вес равен сумме весов клиентов из множества C, обслуживаемых данным предприятием. Иными словами, множество $C_{i_{k+1}}$ может быть получено путём перемещения

всех клиентов из множества C в обслуживающие их предприятия при назначении $\alpha_{i_{k+1}}$. Данный переход можно назвать процедурой агломерации. Пусть α'_{i_k} — оптимальное решение задачи об i_k-медиане, в которой множеством клиентов и множеством предприятий является построенное множество $C_{i_{k+1}}$. Назначение α_{i_k} положим равным композиции назначений $\alpha_{i_{k+1}}$ и α'_{i_k}.

Заметим, что при переходе от назначения $\alpha_{i_{k+1}}$ к назначению α_{i_k} свойства иерархичности (H1), (H2) выполнены. Поэтому построенная последовательность действительно является частичным решением иерархической задачи о назначениях.

Полученное частичное решение произвольным образом достраивается до полного путём открытия в произвольном порядке предприятий, ещё закрытых на шаге i_k, $k < t$, но входящих в состав открытых предприятий на шаге i_{k+1}. В качестве областей обслуживания открываемых предприятий берутся их области обслуживания при назначении $\alpha_{i_{k+1}}$. Оставшиеся назначения α_s, $s > i_t$, получаются из назначения α_{i_t} путём последовательного открытия всех оставшихся закрытых предприятий, в качестве областей обслуживания открываемых предприятий берётся пустое множество.

Теорема. *Описанный выше алгоритм имеет оценку точности*

$$4\,\Omega\,(\Omega + 1).$$

Сложность алгоритма определяется сложностью поиска решений n задач о k-медиане, $k = 1, 2, \ldots, n$, что совпадает со сложностью алгоритма из [4].

2. Доказательство оценки

Вначале получим оценку частичного решения $\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_t}$. Из неравенства треугольника следует, что

$$\operatorname{cost}(\alpha_{i_k}) \leqslant \operatorname{cost}(\alpha_{i_{k+1}}) + \operatorname{cost}(\alpha'_{i_k}).$$

Обозначим через $\operatorname{cost}(Y, X)$ стоимость обслуживания клиентов из множества X предприятиями из множества Y, т. е.

$$\operatorname{cost}(Y, X) = \sum_{x \in X} w(x) \min_{y \in Y} d(x, y).$$

Тогда

$$\operatorname{cost}(\alpha'_{i_k}) = \min\{\, \operatorname{cost}(Y,\, C_{i_{k+1}}) \mid Y \subseteq C_{i_{k+1}}\,,\ |Y| = i_k \,\}.$$

Для оценки последнего выражения нам потребуется следующая геометрическая

Лемма. *Пусть α — решение произвольной задачи о p-медиане, в которой множество клиентов C совпадает с множеством предприятий F. Пусть α^* — решение задачи с тем же множеством клиентов C, но с другим множеством предприятий, содержащим множество F в качестве подмножества. Тогда*

$$\operatorname{cost}(\alpha) \leqslant \Omega \operatorname{cost}(\alpha^*). \tag{1}$$

Доказательство леммы будет приведено ниже.

Из леммы следует, что

$$\operatorname{cost}(\alpha'_{i_k}) \leqslant \Omega \min\{ \operatorname{cost}(Y, C_{i_{k+1}}) \mid Y \subseteq C,\ |Y| = i_k \},$$

а в силу неравенства треугольника последнее не превосходит

$$\Omega \ (\min\{ \operatorname{cost}(Y, C) \mid Y \subseteq C,\ |Y| = i_k \} + \operatorname{cost}(\alpha_{i_{k+1}})) \\ = \Omega \, (\operatorname{cost}(\alpha^*_{i_k}) + \operatorname{cost}(\alpha_{i_{k+1}})).$$

Таким образом,

$$\operatorname{cost}(\alpha_{i_k}) \leqslant \Omega \operatorname{cost}(\alpha^*_{i_k}) + (1 + \Omega) \operatorname{cost}(\alpha_{i_{k+1}}).$$

Раскрывая аналогичную оценку для стоимости назначений $\alpha_{i_{k+1}}$, $\alpha_{i_{k+2}}$, …, получим

$$\operatorname{cost}(\alpha_{i_k}) \leqslant \Omega \operatorname{cost}(\alpha^*_{i_k}) + (1 + \Omega) \, \Omega \operatorname{cost}(\alpha^*_{i_{k+1}}) + (1 + \Omega)^2 \, \Omega \operatorname{cost}(\alpha^*_{i_{k+2}}) + \ldots.$$

Согласно выбору номеров из множества K

$$\operatorname{cost}(\alpha_{i_k}) \leqslant \Omega \operatorname{cost}(\alpha^*_{i_k}) \, (1 + 1/2 + 1/4 + \ldots) \leqslant 2 \, \Omega \operatorname{cost}(\alpha^*_{i_k}).$$

Следовательно, частичное решение $\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_t}$ имеет оценку точности $2\,\Omega$.

Оценим стоимость произвольного назначения α_s, $s = 1, 2, \ldots, n$, $s \notin K$. Пусть i_k — максимальный номер из множества K такой, что $i_k < s$. Поскольку целевая функция не возрастает с ростом числа открытых предприятий, с учётом полученной выше оценки частичного решения приходим к неравенствам

$$\operatorname{cost}(\alpha_s) \leqslant \operatorname{cost}(\alpha_{i_k}) \leqslant 2 \, \Omega \operatorname{cost}(\alpha^*_{i_k}).$$

Но согласно построению множества K справедливо неравенство

$$\mathrm{cost}(\alpha_s^*) > \mathrm{cost}(\alpha_{i_k}^*) \,/\, 2\,(\Omega + 1).$$

Следовательно, $\mathrm{cost}(\alpha_s) < 4\,\Omega\,(\Omega + 1)\,\mathrm{cost}(\alpha_s^*)$.

Таким образом, каждое назначение α_s, $s = 1, 2, \ldots, n$, отличается по стоимости от назначения α_s^* не более чем в $4\,\Omega\,(\Omega + 1)$ раз. Теорема доказана.

Доказательство леммы. Поскольку стоимость любого назначения можно разбить на сумму слагаемых, соответствующих областям обслуживания назначения α^*, достаточно доказать утверждение леммы для случая $p = 1$. Пусть a^* — единственное открытое предприятие назначения α^*, a — точка из C, ближайшая к a^*. В случае совпадения точек a и a^* неравенство (1) очевидно, поэтому будем считать, что $a \neq a^*$. Докажем, что назначение α, соответствующее единственному открытому предприятию a, удовлетворяет неравенству (1).

В случае прямой произвольная точка является медианой тогда и только тогда, когда суммарный вес клиентов, расположенных по одну сторону от неё, не превосходит суммарного веса клиентов, расположенных с противоположной стороны и в самой точке [5]. Поэтому точка a является 1-медианой второй задачи так же, как и точка a^*. Следовательно, в случае прямой неравенство (1) выполнено при $\Omega = 1$.

В общем метрическом случае в силу неравенства треугольника получим

$$\sum_{x \in C} w(x)\, d(x, a) \leqslant \sum_{x \in C} w(x)\,(\, d(x, a^*) + d(a, a^*)\,).$$

Согласно выбору точки a последнее не превосходит

$$\sum_{x \in C} w(x)\,(\, d(x, a^*) + d(x, a^*)\,) = 2\,\mathrm{cost}(\alpha^*).$$

Таким образом, в общем случае неравенство (1) выполнено при $\Omega = 2$.

Осталось доказать справедливость неравенства (1) в евклидовом случае. Поскольку преобразования параллельного переноса и поворота не влияют в данном случае на расстояния между точками, а преобразование масштабирования меняет все расстояния пропорционально друг другу, то можно считать, что точка a^* находится в начале координат, а точка a имеет координаты $1, 0, \ldots, 0$.

Согласно неравенству треугольника при переносе произвольного клиента в направлении от точки a^* приращение величины $\mathrm{cost}(\alpha^*)$ не меньше

приращения величины $\mathrm{cost}(\alpha)$. Поэтому достаточно доказать неравенство (1) для самого худшего случая, когда все клиенты расположены на минимальном расстоянии от a^*, т. е. на сфере радиуса 1.

Целевая функция задачи о медиане равна $\sum\limits_{x\in C} w(x)\|x-a\|$. Поскольку расширение множества предприятий не увеличивает стоимость оптимальной медианы, можно считать, что точка a^* является минимумом данной функции не только на множестве предприятий второй задачи, но и на всем евклидовом пространстве. Производная по первой координате в ней равна нулю,

$$\sum_{x\in C} w(x)\frac{x_1}{\|x\|} = \sum_{x\in C} w(x)\,x_1 = 0, \tag{2}$$

где x_i — i-я координата точки x. Вычислим значение целевой функции первой задачи в точке a:

$$\mathrm{cost}(\alpha) = \sum_{x\in C} w(x)\sqrt{(x_1-1)^2+\sum_{i>1} x_i^2} = \sum_{x\in C} w(x)\sqrt{2-2\,x_1}\,.$$

Заметим, что $\sqrt{2-2\,x_1} \leqslant \sqrt{2} - x_1/\sqrt{2}$. Отсюда с учётом равенства (2) получаем

$$\mathrm{cost}(\alpha) \leqslant \sum_{x\in C} \sqrt{2}\,w(x) - \sum_{x\in C} w(x)\frac{x_1}{\sqrt{2}} = \sum_{x\in C} \sqrt{2}\,w(x) = \sqrt{2}\,\mathrm{cost}(\alpha^*).$$

Таким образом, в евклидовом случае неравенство (1) выполнено при $\Omega = \sqrt{2}$. Лемма доказана.

ЛИТЕРАТУРА

1. **Шенмайер В. В.** Приближенный алгоритм для одномерной задачи о последовательности медиан // Дискрет. анализ и исслед. операций. Сер. 1. — 2007. — Т. 14, № 2. — С. 95–101.
2. **Chrobak M., Kenyon C., Noga J., Young N.** Online medians via online bribery // Proc. of the 7th Latin american theoretical informatics symposium. — Berlin, Heidelberg: Springer-Verl., 2006. — P. 311–322. (Lect. Notes Comp. Sci.; Vol. 3887).
3. **Dasgupta S.** Performance guarantees for hierarchical clustering // Proc. of the 15th conference on computational learning theory. — London: Springer-Verl., 2002. — P. 351–363.
4. **Lin G., Nagarajan C., Rajaraman R., Williamson D. P.** A general approach for incremental approximation and hierarchical clustering // Proc. of the 17th ACM-SIAM symposium on discrete algorithms. — New York: ACM Press, 2006. — P. 1147–1156.

5. Mirchandani P., Francis R. (eds.). Discrete location theory. — New York: Wiley, 1990. — 576 p.

6. Plaxton C. G. Approximation algorithms for hierarchical location problems // Proc. of the 35th ACM symposium on theory of computing. — New York: ACM Press, 2003. — P. 40—49.

Шенмайер Владимир Владимирович,
e-mail: shenmaier@mail.ru

Статья поступила
18 декабря 2007 г.

Переработанный вариант —
11 июля 2008 г.