4H-02

直線上の Min-sum r-cellular Clustering

小川 航平 † 赤木 俊裕 ‡ 宮田 洋行 § 中野 眞一 ¶
群馬大学 † 群馬大学 ‡ 群馬大学 § 群馬大学 ¶

1 あらまし

様々な制約のもとで, 指定したコストが最適になるように, 施設を配置する問題を施設配置問題という. 様々な施設配置問題が研究されてきた [6, 7]. 本文は, 最近提案された 2 つの施設配置問題を扱う.

距離空間に点集合 P が与えられたとき, これらを, それぞれが r 点以上からなる部分集合 (クラスタ) に, 指定したコストが最小になるように分割する問題を, 一般に, r-gather 問題と呼ぶ [1, 3, 8]. これらの問題は一般に NP 完全である [1, 3]. そこで限られたクラスの入力に対して, この問題を検討したい.

本文は, 直線上に点集合 P が与えられたとき, 2 つの r-gather 問題を解くアルゴリズムを与える.

2 定義

距離空間の点集合 P の, それぞれが r 点以上からなる部分集合 (クラスタ) への分割を r-gather という. このとき, クラスタの最大の半径を max クラスタコストと呼ぶ. ここで, クラスタの半径とは, クラスタ中のすべての点をカバーする最小の円の半径とする. 点集合 P と整数 r が与えられたとき, max クラスタコストが最小の r-gather を求める問題を r-gather clustering 問題と呼ぶ. この問題を直線上の場合で解く $O(n \log n)$ 時間アルゴリズムが知られている [2]. また, 各クラスタについて, クラスタの半径とクラスタ中の点の個数を乗じたものを, 全てのクラスタについて合計したものを cellular clustering コストと呼ぶ [1]. 点集合 P と整数 r が与えられたとき, cellular clustering コストが最小の r-gather を求める問題を, r-cellular clustering 問題とよぶ [1]. この問題は NP 完全であり [定理 3.2, [1]], 3 倍近似アルゴリズムが知られている [定理 3.5[1]]. 一方, P が直線上の点集合であったとしても, これを解くアルゴリズムは知られていない.

3 アルゴリズム

文献 [4] は, ある施設配置問題を, 最短路問題に帰着して解く方法を提案している. 本文は, P が直線上の点集合であるとき, r-cellular clustering 問題を, 最短路問題に帰着して解く方法を提案する. $P = \{p_1, p_2, ..., p_n\}$ は水平直線上の点集合とし, $p_1, p_2, ..., p_n$ は左から右への順でソートされているとする.

補題 3.1 cellular clustering コストが最小の r-gather 中の, $s < t < u < v$ なる 4 点 p_s, p_t, p_u, p_v と, 2 つのクラスタ C_l と C_r について, $p_s, p_u \in C_l$, かつ, $p_t, p_v \in C_r$ となることはないとしてよい.

証明 3.1 そのような 4 点の組の個数が最小であるような, cellular clustering コストが最小の r-gather を考えよう.

Min-sum r-cellular Clustering on the Line
†Kohei Ogawa · Gunma Univercity
‡Toshihiro Akagi· Gunma Univercity
§Hiroyuki Miyata · Gunma Univercity
¶Shin-ichi Nakano· Gunma Univercity

このとき, $p_s, p_t \in C_l$, かつ, $p_u, p_v \in C_r$ のように 2 点 p_t と p_u を入れ替えることにより, コストが同じで, 4 点の個数がより少ない r-gather があることになり, 矛盾である. □

補題 3.2 cellular clustering コストが最小の r-gather における の各クラスタは, 高々$2r - 1$ 個の点をもつとしてよい.

証明 3.2 $2r$ 個以上の点を持つクラスタはさらに分割することにより, コストが同じか, または少ない r-gather となることによる. □

水平直線上の点集合を $P = \{p_1, p_2, ..., p_n\}$ とする. このとき, グラフ $G = (V, E)$ を次のように定義する. 図 1 参照. まず, $V = \{v_0, v_1, ..., v_n\}$ とする. 直感的には, $i = 1, 2, ..., n-1$ なる各点 $v_i \in V$ は区間 $[p_i, p_{i+1}]$ に対応する. 特に v_0 は区間 $[-\infty, p_1]$ に対応し, v_n は区間 $[p_n, +\infty]$ に対応する. また, $i < j$ かつ $i + r \leq j \leq i + 2r - 1$ となる各 2 点 $v_i, v_j \in V$ を, 辺 (v_i, v_j) で結び, これらの集合を E とする. $|E| \leq r|V|$ である. 辺 (v_i, v_j) の長さは, $(j-i)\frac{|P_j - P_{i+1}|}{2}$ とする. 次の補題が言える.

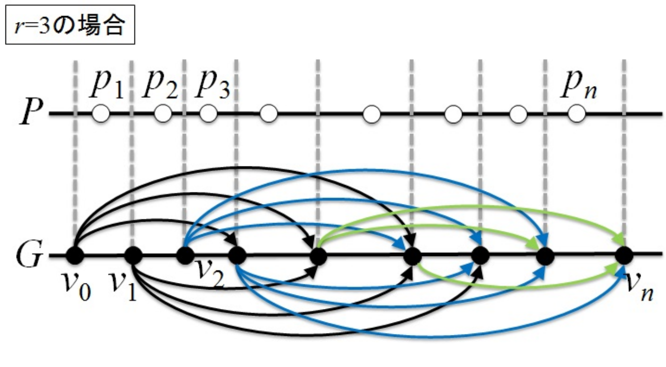

図 1: 補助グラフ G の例

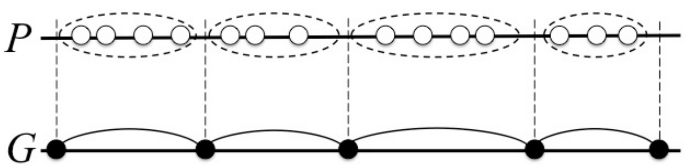

図 2: クラスタの集合とパスの対応

補題 3.3 G の各辺のコストを $O(1)$ 時間で計算できる.

G 中の v_0 から v_n へのパスを P の r-gather に次のように対応させる (図 2 参照). パス中に辺 (v_i, v_j) があるとき, 点集合 $\{p_{i+1}, p_{i+2}, ..., p_j\}$ がひとつのクラスタに

対応させる．すると, G 中の v_0 から v_n への最短路が, P の r-cellular clustering 問題の解に対応する．

G は有向グラフ, かつ, サイクルを持たないので, v_0 から v_n への最短路は $\Theta(|V|+|E|)$ 時間で計算できる [p.655, [5]].

定理 3.1 直線上の点集合 P が与えられたとき, r-cellular clustering 問題を $O(r|P|)$ 時間で解くアルゴリズムがある．このアルゴリズムは, この問題をとくはじめての多項式時間アルゴリズムである．

4 r-gather sum 問題

本章は, もうひとつの r-gather 問題を定義し, これを解くアルゴリズムを与える．

r-gather の各クラスタについて, クラスタ中の各点からのコストの合計が最小の点をそのクラスタの施設点とよぶ．クラスタ中の各点からクラスタの施設点へのコストを合計したものを, そのクラスタの sum コストとし, これらをすべてのクラスタについて合計したものを, その r-gather の sum コストとよぶ．

sum コストが最小の r-gather を求める問題を r-gather sum 問題と呼ぶ．(より一般的な問題が文献 [8] で扱われている)． 3 章と同様に, グラフ $G=(V,E)$ を定める．ただし, 各辺 (v_i,v_j) のコストは, クラスタ $P_{ij}=\{p_{i+1},p_{i+2},...,p_j\}$ の sum コストとする．これは施設点を, $p_{i+1},p_{i+2},...,p_j$ の座標の中央値に置いたときの, クラスタの各点から施設点へのコストの合計であることが次の補題でいえる．

補題 4.1 施設点は $p_{i+1},p_{i+2},...,p_j$ の座標の中央値として良い．

証明 4.1 施設点が, 中央値でないとして矛盾を導く．2 つの場合に分けて考えよう． $j-i$ が奇数のとき, 中央値の方に施設を移動すると, コストが減ることになり矛盾である． $j-i$ が偶数のとき, もし, 施設点が $p_{\frac{i+j}{2}}$ と $p_{\frac{i+j}{2}+1}$ の間のどこにあっても sum コストは同じである．このとき, sum コストが最小となる． □

$O(n)$ 時間の前処理ののち, G の各辺のコストは $O(1)$ 時間で計算できることを説明しよう． 前処理として, 各 i について $A_i=\sum_{j=i,i+1,...,n}|p_n-p_j|$ と $B_i=\sum_{j=i,i+1,...,n}|p_j-p_1|$ を計算しておく．これには $O(n)$ 時間かかる．

辺 (v_i,v_j) のコストについて考えよう．2 つの場合に分け考える．

まず, $j-i$ が奇数の時を考えよう．この辺 (v_i,v_j) はクラスタ $\{p_{i+1},p_{i+2},...,p_j\}$ に対応する．このクラスタの点の座標の中央値は, $x=(i+j+1)/2$ としたとき p_x の座標である．このとき, このクラスタの sum コストは

$$
\begin{aligned}
&|p_x-p_{i+1}|+|p_x-p_{i+2}|+\cdots+|p_x-p_{x-1}|\\
&\quad+|p_x-p_j|+|p_x-p_{j-1}|+\cdots+|p_x-p_{x+1}|\\
&=|p_n-p_{i+1}|+|p_n-p_{i+2}|+\cdots+|p_n-p_{x-1}|\\
&\quad-(x-i-1)|p_n-p_x|\\
&+|p_1-p_j|+|p_1-p_{j-1}|+\cdots+|p_1-p_{x+1}|\\
&\quad-(j-x)|p_x-p_1|\\
&=A_{i+1}-A_x-(x-i-1)|p_n-p_x|\\
&\quad+B_j-B_x-(j-x)|p_x-p_1|
\end{aligned}
$$

よって, 最後の式は O(1) 時間で計算できるので、辺のコストは O(1) で計算できる．

$j-1$ が偶数の場合も, $x=(i+j)/2$ としたとき, p_x と p_{x+1} の中点を中央値として扱うことにより, 同様に $O(1)$ 時間で計算できる．また, G の v_0 から v_n への最短路は $\Theta(|V|+|E|)$ 時間で計算できる [p.655, [5]].

定理 4.1 直線上の点集合 P が与えられたとき, r-gather sum 問題を $O(r|P|)$ 時間で解くアルゴリズムがある．

5 まとめ

r-cellular clustering 問題と直線上の r-gather sum 問題を, それぞれ最短路問題に帰着して解くアルゴリズムを設計した．これらのアルゴリズムは非常に簡単であり, かつ, 高速である．計算時間は $O(r|P|)$ 時間である．

参考文献

[1] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, A. Zhu, Achieving Anonymity via Clustering, ACM Trans. Algorithms, Vol. 6, Article 49 (2010).

[2] T. Akagi, S. Nakano, On r-gathering on the Line, Proc. of FAW, Lecture Notes in Computer Science, Vol. 9130, pp. 25-32 (2015).

[3] A. Armon, On min-max r-gatherings, Theoretical Computer Science, 412, pp.573-582 (2011).

[4] D. Z. Chen, H. Wang, New Algorithms for Facility Location Problems on the Real Line, Algorithmica, Vol. 69, Issue 2, pp. 370-383 (2014).

[5] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, Introduction to Algorithms, Third Edition, The MIT Press (2009).

[6] Z. Drezner, Facility Location : A Survey of Applications and Methods, Springer (1995).

[7] Z. Drezner and H. W. Hamacher, Facility Location : Applications and Theory, Springer (2004).

[8] S. Guha, A. Meyerson, K. Munagala, Hierarchical Placement and Network Design Problem, Proc. of FOCS 2000, pp.603-612 (2000).