Algoritmos de Aproximação para Problemas de Clustering

João Guilherme Alves Santos **Orientadora: Cristina Gomes Fernandes**

Departamento de Ciência da Computação, Instituto de Matemática e Estatística, Universidade de São Paulo

Introdução

Problemas de clustering têm por objetivo agrupar objetos de maneira que objetos no mesmo cluster apresentam mais semelhanças quando comparados a objetos em clusters diferentes. Resumidamente, os problemas que estudamos buscam encontrar uma maneira menos custosa de posicionar instalações para melhor atender um conjunto de clientes. Os clientes estarão em um mesmo cluster se a instalação mais próxima a eles for a mesma. Neste trabalho estudamos vários algoritmos de aproximação e resultados de inaproximabilidade para três problemas de clustering *NP*-difíceis: k-medianas, k-centros e localização de instalações. Esses problemas são muito conhecidos e importantes nas áreas de pesquisa operacional e otimização combinatória.

k-Centros métrico

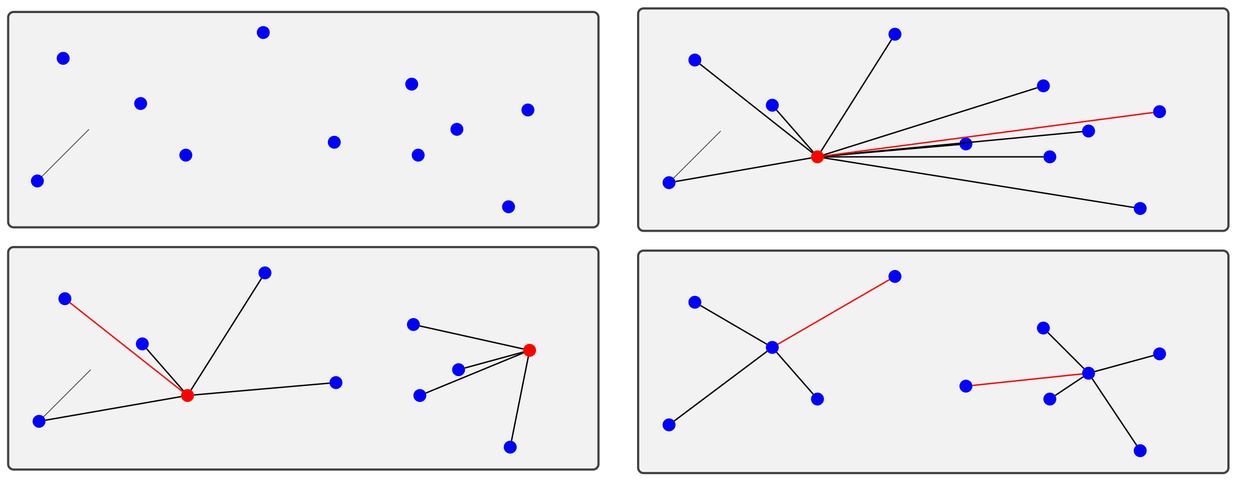
Dado um grafo completo G, um inteiro k e uma métrica $c : V(G) \times V(G) \to \mathbb{R}_+$, encontrar $S \subseteq V(G)$ com $|S| = k$ que minimize $\max_{j \in V(G)} c(j, S)$ em que $c(j, S) \coloneqq \min_{i \in S} c_{ij}$.

Aproximação para o k-centros métrico

O seguinte algoritmo guloso de González [1] é uma 2-aproximação para o k-centros métrico: escolha uma cidade qualquer para construir a primeira instalação e, iterativamente, escolha a cidade mais distante das instalações já escolhidas.

Algorithm Guloso-González(G, c, k)

1: Escolha arbitrariamente $u \in V(G)$.
2: $S \leftarrow \{u\}$
3: **Enquanto** $|S| < k$ **faça**
4: $\quad v \leftarrow \arg\max_{j \in V} c(j, S)$
5: $\quad S \leftarrow S \cup \{v\}$
6: **Devolva** S



k-Medianas métrico

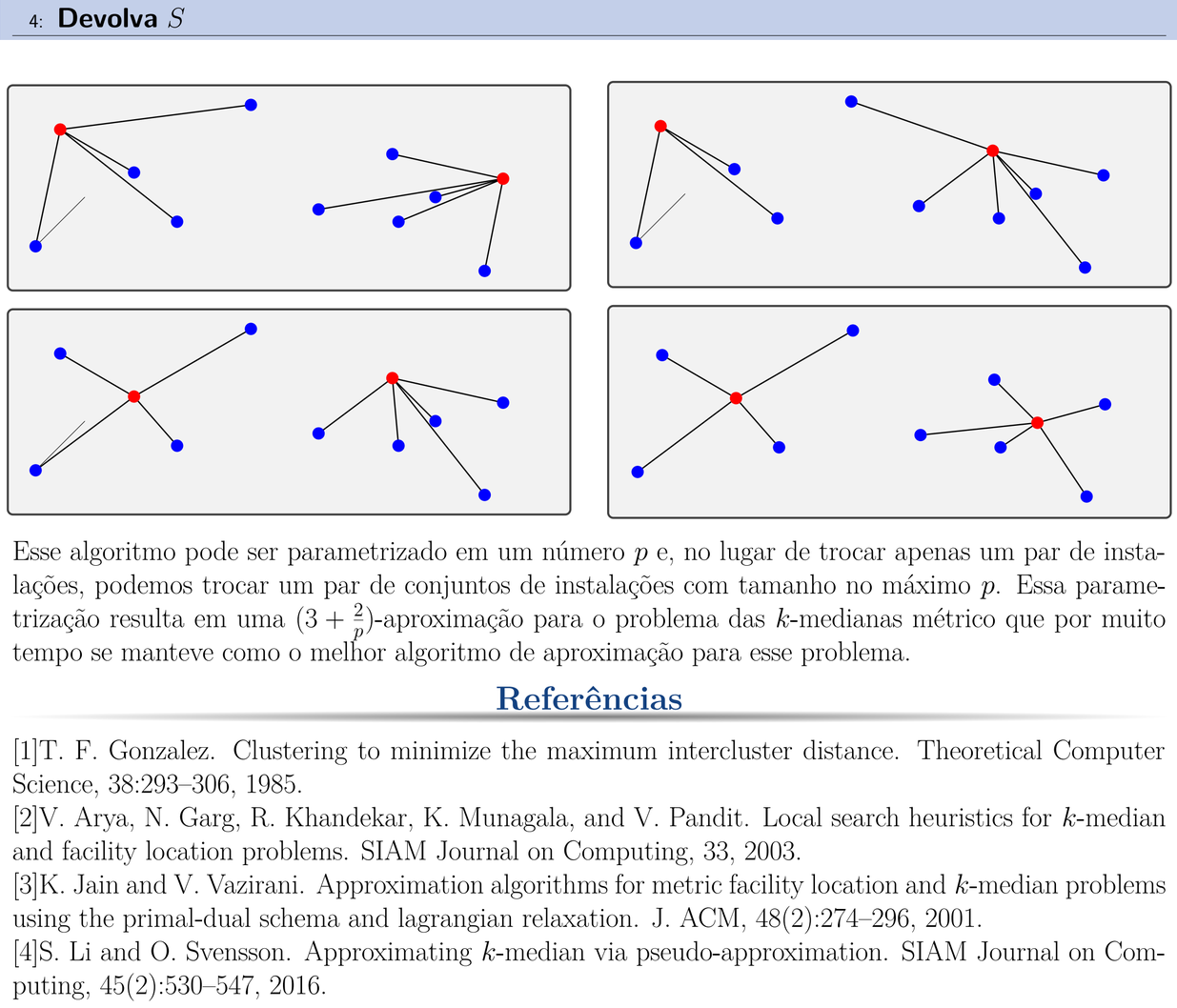
Dado um grafo completo G, um inteiro k e uma métrica $c : V(G) \times V(G) \to \mathbb{R}_+$, encontrar $S \subseteq V(G)$ com $|S| = k$ que minimize $\sum_{j \in V(G)} c(j, S)$ em que $c(j, S) \coloneqq \min_{i \in S} c_{ij}$.

Aproximação para o k-medianas métrico

Para esse problema descreveremos uma 5-aproximação que usa busca local, de Arya et al. [2]. O algoritmo começa com uma solução viável qualquer e vai trocando uma instalação aberta por uma fechada caso a troca melhore o custo, até não haver troca que melhore a solução.

Algorithm BuscaLocal-AGKMMP$(G = (V, E), c, k)$

1: Seja S um conjunto arbitrário com k elementos de V.
2: **Enquanto** existir $u, v \in F$ tal que $u \in S$ e custo$(S - u + v) <$ custo(S) **faça**
3: $\quad S \leftarrow S - u + v$
4: **Devolva** S



Esse algoritmo pode ser parametrizado em um número p e, no lugar de trocar apenas um par de instalações, podemos trocar um par de conjuntos de instalações com tamanho no máximo p. Essa parametrização resulta em uma $(3 + \frac{2}{p})$-aproximação para o problema das k-medianas métrico que por muito tempo se manteve como o melhor algoritmo de aproximação para esse problema.

Localização de instalações métrico

Dado um grafo (D, F)-bipartido completo G, uma métrica $c : F \times D \to \mathbb{R}_+$ e uma função $f : F \to \mathbb{R}_+$, encontrar $S \subseteq F$ que minimize $\sum_{i \in S} f_i + \sum_{j \in D} c(j, S)$ em que $c(j, S) \coloneqq \min_{i \in S} c_{ij}$.

Aproximação para o localização de instalações métrico

Para o problema de localização de instalações métrico descreveremos uma 3-aproximação primal-dual, de Jain e Vazirani [3]. Segue a relaxação do programa inteiro que modela esse problema e o seu dual, onde x_{ij} e y_i indicam se o cliente j está associado à instalação i e se a instalação i está aberta.

$$\begin{array}{lllll} \text{Min} & \sum_{i \in F} f_i y_i + \sum_{i \in F, j \in D} & c_{ij} x_{ij} & & \\ \text{suj. a} & \sum_{i \in F} x_{ij} \geq 1 & \forall j \in D & & \\ & y_i - x_{ij} \geq 0 & \forall i \in F, j \in D & & \\ & x_{ij} \geq 0, y_i \geq 0 & \forall i \in F, j \in D & & \end{array}$$

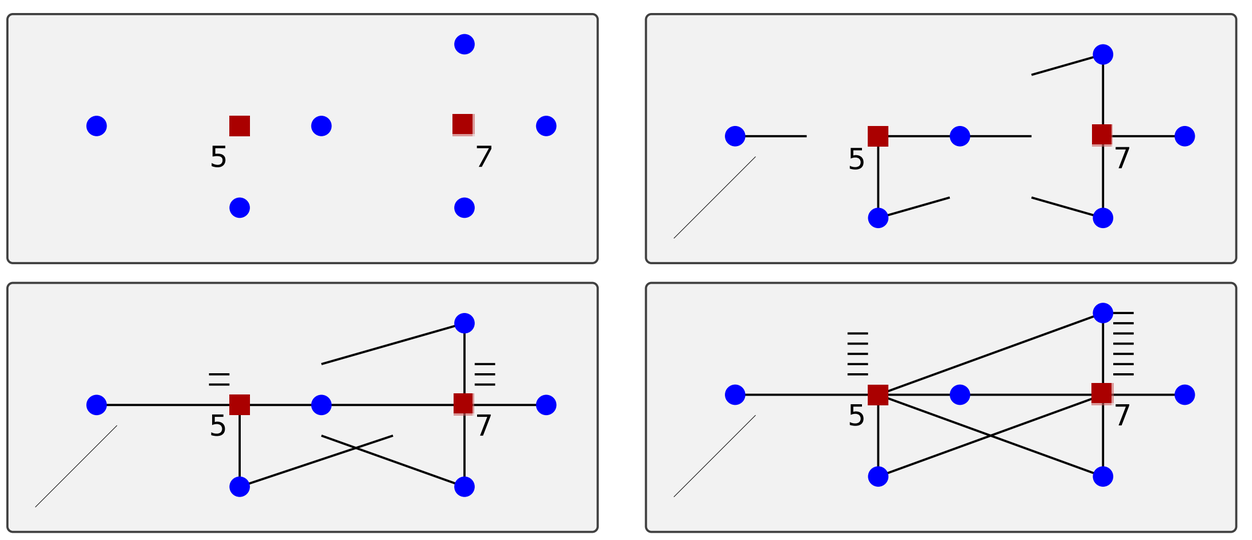
$$\begin{array}{lllll} \text{Max} & \sum_{j \in D} v_j & & & \\ \text{suj. a} & \sum_{j \in D} w_{ij} \leq f_i & \forall i \in F & & (1) \\ & v_j - w_{ij} \leq c_{ij} & \forall i \in F, j \in D & & (2) \\ & w_{ij} \geq 0, v_j \geq 0 & \forall i \in F, j \in D & & \end{array}$$

No dual, v_j representa quanto o cliente j está disposto a pagar para se conectar a alguma facilidade, e w_{ij} é quanto disso ele pagaria para se conectar à facilidade i.
O algoritmo é dividido em duas partes. Na primeira, produzimos uma solução viável (v, w) do dual enquanto construímos um conjunto de instalações momentaneamente abertas. Na segunda, escolhemos algumas instalações desse conjunto para serem propriamente abertas.

Algorithm PrimalDual-JV$(G = (D \cup F, E), c, f)$

1: $v \leftarrow 0$; $\quad w \leftarrow 0$; $\quad C \leftarrow D$; $\quad S \leftarrow \emptyset$
2: **Enquanto** $C \neq \emptyset$ **faça**
3: $\quad N(j) \leftarrow \{i \in F : v_j \geq c_{ij}\}$ para todo $j \in C$
4: $\quad$ Aumente v_j e w_{ij} uniformemente para todo $j \in C$ e $i \in N(j)$
$\quad\quad$ até que (1) ou (2) fique justo para algum i e para algum j
5: $\quad$ **Se** existe $i \notin S$ tal que $\sum_{j \in D} w_{ij} = f_i$ **então**
6: $\quad\quad S \leftarrow S + \{i\}$
7: $\quad\quad C \leftarrow C - \{j \in C : i \in N(j)\}$
8: $\quad$ **Se** existe $j \in C$ tal que $v_j = c_{ij}$ para algum $i \in S$ **então**
9: $\quad\quad C \leftarrow C - \{j\}$
10: $S' \leftarrow \emptyset$
11: **Enquanto** $S \neq \emptyset$ **faça**
12: $\quad$ Escolha $i \in S$; $S' \leftarrow S' + \{i\}$
13: $\quad S \leftarrow S - \{h \in S :$ existe j tal que $w_{ij} > 0$ e $w_{hj} > 0\}$
14: **Devolva** S'

Considere uma instância simples com duas instalações e seis clientes. Vamos simular a primeira fase do algoritmo. No desenho, as arestas de um cliente j crescem ao longo que o valor de v_j cresce e encosta em uma instalação i quando v_j chega no valor c_{ij}.



Conclusões

Vários algoritmos foram estudados para cada um dos três problemas. De todos esses, escolhemos um de cada problema para ser apresentado aqui. Além disso, também estudamos os melhores resultados de inaproximabilidade para cada um desses problemas.
Além dos algoritmos apresentados, destaco também um algoritmo sofisticado para o problema das k-medianas métrico, desenvolvido por Li e Svensson [4]. Esse algoritmo trouxe avanço para um problema que estava há mais de uma década sem novos resultados. Eles mostraram que, para construir uma $(\alpha + \varepsilon)$-aproximação para o problema das k-medianas métrico, é suficiente encontrar um algoritmo que devolve uma pseudo solução, ou seja, um conjunto com mais que k instalações, com custo α vezes o ótimo. Esse resultado ampliou as possibilidades de pesquisa, evidenciando que soluções que não se limitam exatamente a k instalações também podem trazer contribuições relevantes para o campo. Embora esse não seja o estado da arte para o problema das k-medianas métrico, os melhores resultados existentes hoje são derivados direto desse trabalho.

Referências

[1]T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, 38:293–306, 1985.
[2]V. Arya, N. Garg, R. Khandekar, K. Munagala, and V. Pandit. Local search heuristics for k-median and facility location problems. SIAM Journal on Computing, 33, 2003.
[3]K. Jain and V. Vazirani. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and lagrangian relaxation. J. ACM, 48(2):274–296, 2001.
[4]S. Li and O. Svensson. Approximating k-median via pseudo-approximation. SIAM Journal on Computing, 45(2):530–547, 2016.

Agradecimentos

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil. Processo nº 2023/16197-0.
Análises desses e de outros algoritmos para esses problemas podem ser encontradas em https://linux.ime.usp.br/ alvesjg/mac0499