Projeto de Pesquisa para Iniciação Científica

Algoritmos de Aproximação para Problemas de Clustering

Orientadora: Cristina Gomes Fernandes

Aluno: João Guilherme Alves Santos

Resumo

O objetivo desse trabalho é estudar e pesquisar algoritmos de aproximação para problemas de clustering. Será realizado um estudo abrangente de algoritmos de aproximação, focado em problemas de clustering, explorando técnicas e análises associadas a esses problemas. Em seguida, serão estudados algoritmos recentes que integram essas técnicas para alcançar melhores resultados. Três problemas de clustering serão estudados: k-centros, localização de instalações e k-mediana.

1 Introdução

Problemas de otimização têm o objetivo de encontrar um ponto ótimo de uma função definida sobre um certo domínio. Especificamente, os problemas de otimização combinatória têm domínio finito. Muitos desses problemas são NP-difíceis. Para problemas NP-difíceis, não existem algoritmos eficientes que encontrem uma solução ótima para toda instância de tais problemas a menos que $P = NP$.

Nesse contexto, algoritmos de aproximação surgiram. A ideia é abrir mão de encontrar soluções ótimas para encontrar, eficientemente, uma solução cujo valor garante uma relação pré-estabelecida com o valor ótimo.

Clustering refere-se a uma classe de problemas de otimização cujo objetivo é agrupar objetos de maneira que objetos no mesmo cluster apresentem mais semelhanças quando comparados a objetos em clusters diferentes. Tais semelhanças serão definidas pelo problema em questão. Neste projeto, iremos estudar, sob o ponto de vista de algoritmos

de aproximação, três problemas de clustering NP-difíceis: localização de instalações, k-mediana e k-centros.

Localização de instalações é um problema que visa determinar a melhor localização para instalações, como fábricas ou depósitos, com base no custo de abertura das instalações e custos de transporte. Além disso, pode ser modelado para outras aplicações como problemas de posicionamento de caches em um computador ou problemas de projeto de redes.

Existem várias versões do problema de localização de instalações, a mais simples delas é a versão sem capacidades em que as instalações não têm limitações para suprir os clientes.

No problema de localização de instalações sem capacidades, temos um grafo (F, D)-bipartido completo em que D é o conjunto de clientes a serem atendidos e F o conjunto de instalações que podem ser abertas. Para cada cliente $j \in D$ e cada instalação $i \in F$, há um custo c_{ij} para a aresta ij em associar o cliente j à instalação i. Além disso, existe um custo de abertura f_i para cada instalação $i \in F$. O objetivo do problema é escolher um conjunto $F' \subseteq F$ e uma função que associe cada cliente a uma instalação aberta tal que o custo total de abertura das instalações em F' somado ao custo de associação de cada cliente $j \in D$ à instalação em que ele está associado seja minimizado. Em outras palavras, queremos encontrar $F' \subseteq F$ e uma função σ que minimize $\sum_{i \in F'} f_i + \sum_{j \in D} c_{\sigma(j)j}$.

O problema k-mediana é muito parecido com o problema de localização de instalações. A diferença aqui é que não temos custo para a abertura de instalações e podemos abrir no máximo k delas.

Assim como no localização de instalações, no problema k-mediana temos um grafo (F, D)-bipartido completo em que D é o conjunto de clientes a serem atendidos e F o conjunto de instalações que podem ser abertas. Para cada cliente $j \in D$ e cada instalação $i \in F$, há um custo c_{ij} para a aresta ij em associar o cliente j à instalação i. Temos também um inteiro k que representa a quantidade de instalações que podem ser abertas. Então, queremos encontrar um conjunto $F' \subseteq F$ de tamanho k e uma função

que associe cada cliente a uma instalação aberta tal que o custo total de associação seja minimizado.

No problema dos k-centros não existe essa diferença entre instalações que podem ser abertas e clientes, teremos cidades e escolheremos k delas para construir instalações.

Então, temos um grafo $G(V, E)$ completo em que V são cidades e temos um custo c_e para associar as cidades que são extremos de e. Temos também um inteiro k que representa a quantidade de cidades em que uma instalação será aberta. Cada cidade será associada a uma cidade com uma instalação aberta com menor custo de associação entre elas. O objetivo do nosso problema é minimizar o maior custo de associação entre uma cidade qualquer e a cidade a qual ela está associada.

Diversos métodos podem ser utilizados para aproximar o problema de localização de instalações. Charikar e Guha desenvolveram um algoritmo com razão de aproximação 2.414 utilizando o método de busca local [3]. Esse problema também pode ser modelado como um problema de programação inteira e, por isso, técnicas envolvendo programação linear podem ser aplicadas a ele. Por exemplo, há algoritmos que fazem o arredondamento de soluções da relaxação linear do programa inteiro para obter uma solução aproximada do problema. Alguns destes algoritmos atingem boas razões de aproximação, por exemplo, chegando a 1.677 [1]. Entretanto, a melhor aproximação encontrada utiliza de vários métodos, incluindo o conhecido método primal-dual, e garante uma razão de aproximação 1.488 [10]. Essa não é muito distante do melhor que se poderia encontrar, uma vez que Guha e Khuller mostraram que não existe algoritmo para esse problema com razão de aproximação melhor que 1.463 [6], a menos que $P = NP$.

Dentre os três problemas apresentados, o k-mediana é o que tem a maior folga entre o melhor resultado de inaproximabilidade e a razão do melhor algoritmo de aproximação conhecido. Jain, Mahdian e Saberi provaram que não existe algoritmo polinomial com razão de aproximação $1 + \frac{2}{e}$ para o k-mediana [9], assumindo que $P \neq NP$, enquanto a melhor aproximação encontrada tem razão $2.675 + \epsilon$ [2].

Hsu e Nemhauser [8] mostraram que não existe algoritmo polinomial com razão de

aproximação menor que 2 para o problema dos k-centros, assumindo que $P \neq NP$. Neste caso, temos algoritmos que apresentam o melhor desempenho possível: utilizando o método do gargalo, Gonzalez [5] e independentemente Hochbaum e Shmoys [7] desenvolveram um algoritmo polinomial com razão de aproximação igual a 2.

2 Justificativa

O estudo de algoritmos de aproximação para problemas de clustering desempenha um papel crucial na pesquisa em ciência da computação. Essa área possui uma ampla gama de aplicações, desde a segmentação de clientes em marketing até a organização de dados biológicos em genômica comparativa. Os problemas de clustering são intrinsecamente desafiadores, uma vez que encontrar uma solução ótima geralmente é NP-difícil, o que torna essencial o desenvolvimento de algoritmos que possam fornecer soluções aproximadas eficientes. Além disso, à medida que a quantidade de dados continua a crescer, a capacidade de realizar tarefas de clustering de forma eficaz se torna cada vez mais crucial.

Além das aplicações, a pesquisa em algoritmos de aproximação para problemas de clustering desperta um considerável interesse teórico. Esses problemas são de suma importância na área da otimização combinatória, e os algoritmos desenvolvidos para abordá-los empregam uma ampla gama de métodos. Essa dimensão teórica é essencial para o desenvolvimento do estudante nessa área, pois proporciona uma base sólida que lhe permitirá aprimorar as soluções existentes ou criar abordagens inovadoras para problemas que possam surgir. Isso, por sua vez, contribui significativamente para o avanço contínuo do campo e capacita o estudante a enfrentar desafios complexos relacionados à otimização.

3 Plano de trabalho e cronograma

O João Guilherme começou a trabalhar nesse projeto em setembro de 2023. Antes disto, ele estudou os capítulos iniciais do livro de Carvalho et al. [4] sobre algoritmos de aproximação. Ao mesmo tempo, durante esse semestre, o João Guilherme estava cursando a disciplina de Algoritmos de Aproximação, que está cobrindo todo o material deste livro e se aprofundando em várias das técnicas de algoritmos de aproximação. Com isso, o João já obteve uma boa base nesta área.

A partir de setembro de 2023, começamos a focar os estudos em problemas de clustering. Inicialmente, o João Guilherme estudou o Capítulo 2 do livro de Williamson e Shmoys [13] (WS2011) e o Capítulo 5 do livro de Vazirani [12] (V2001), que abordam o algoritmo guloso e os resultados de inaproximabilidade para o problema dos k-centros respectivamente. O próximo passo será estudar o método do gargalo, que se aplica em particular ao problema dos k-centros. Esse método é abordado por Hochbaum e Shmoys [7] e foi apresentado por Hsu e Nemhauser [8]. Essa etapa completa o estudo da versão mais simples do problema dos k-centros.

Durante os meses seguintes, o João estudará os resultados sobre o problema de localização de instalações. Os primeiros resultados que estudaremos são os que envolvem programação linear, e em especial a técnica primal-dual. Estes resultados se encontram descritos no Capítulo 24 do livro V2001 [12] e no Capítulo 7 do livro WS2011 [13]. O João já tem a base necessária para estudar esse material uma vez que o método primal-dual foi abordado durante a disciplina de Algoritmos de Aproximação que ele está cursando. Também estudaremos o material da Seção 4.5 do livro WS2011 [13], que contém um algoritmo de arredondamento de uma solução de um problema linear que modela o problema da localização de instalações.

A seguir, vamos estudar métodos gulosos e de busca local aplicados tanto ao problema k-mediana como ao problema de localização de instalações. Inicialmente vamos usar o material do Capítulo 9 do livro WS2011. Neste mesmo livro, no Capítulo 12, são apresentadas técnicas probabilísticas que dão bons resultados para o problema da

localização de instalações, e que também serão estudados nos primeiros seis meses deste projeto.

O João já começou a escrever os resultados que ele já estudou sobre o k-centros. Esse texto está sendo revisado, e a nossa ideia é ir escrevendo tudo o que for estudado em paralelo com os estudos.

No segundo semestre, continuaremos os estudos sobre o k-mediana, usando o material do Capítulo 25 do livro V2001 [12], que aborda o método primal-dual usando relaxação Lagrangeana e o método de arredondamento probabilístico, assim como técnicas de desaleatorização aplicadas a este algoritmo.

É bom destacar que há muito material sobre estes três problemas na literatura. Aqui apresentamos apenas um levantamento inicial, mas pretendemos também complementar esse levantamento durante os seis primeiros meses de trabalho.

Após esse estudo inicial dos três problemas, escolheremos algum dos algoritmos mais recentes para o problema da localização de instalações e/ou para o k-mediana, retirados de artigos tais como [2, 10, 11], que apresentam as melhores aproximações até o momento para estes problemas, ou algum outro artigo anterior, mais acessível, porém também com material que não seja abordado nos livros da área. Os algoritmos mais recentes para estes problemas são sofisticados, assim sendo podemos levar um tempo maior do que o previsto para cobrir devidamente este material.

O cronograma estipulado para 12 meses é o seguinte.

Ativ/Mês	1	2	3	4	5	6	7	8	9	10	11	12
1	√											
2		√	√									
3				√	√							
4	√	√	√	√	√	√						
5				√								
6						√						
7							√	√				
8									√	√	√	
9							√	√	√	√	√	√
10												√

Legenda:

1. Estudo do problema k-centros: resultados de inaproximabilidade, algoritmo guloso e método do gargalo. Em curso.
2. Estudo do problema de localização de instalações: método primal-dual, de arredondamento e inaproximabilidade.
3. Estudo do problema de localização de instalações: busca local, método guloso e probabilístico.
4. Escritas parciais do texto, referente a 1, 2 e 3.
5. Estudo do problema k-mediana: busca local.
6. Preparação do relatório intermediário.
7. Estudo do problema k-mediana: método primal-dual usando relaxação Lagrangeana, método de arredondamento probabilístico e inaproximabilidade.
8. Estudo de resultados mais recentes para o problema de localização de instalações e/ou k-mediana.
9. Continuação da escrita do texto.
10. Preparação do relatório final.

4 Material e métodos

Além dos artigos citados ao longo de todo o projeto, utilizaremos livros muito conceituados na área como o "Approximation Algorithms" do Vazirani [12] e o "The Design of Approximation Algorithms" do Williamson e do Shmoys [13].

O candidato irá interagir ativamente com a supervisora ao longo do processo de pesquisa. Um dos principais aspectos dessa interação envolverá a redação de partes do estudo e material de pesquisa. Essa prática serve para diversos propósitos, incluindo avaliar o nível de compreensão do candidato e permitir que a supervisora forneça feedback e orientação pontuais conforme necessário. Esse processo iterativo de feedback garantirá que a abordagem de pesquisa do candidato permaneça no rumo certo e esteja alinhada com os objetivos do projeto.

Em conclusão, esta metodologia de pesquisa combina a aquisição de ferramentas essenciais com uma leitura extensiva de artigos de pesquisa relevantes. A interação ativa do candidato com a supervisora, incluindo a redação do material de estudo e pesquisa, assegura uma abordagem de pesquisa focada e alinhada. Ao seguir essa metodologia, o projeto tem como objetivo desenvolver uma base sólida em algoritmos de aproximação e inaproximabilidade, permitindo a exploração de tópicos avançados e o alcance dos objetivos do projeto.

5 Forma e análise dos resultados

Durante todo o período de estudo, o aluno estará preparando um texto, que, ao final do trabalho, conterá tudo que foi estudado na iniciação científica. Este é o principal objeto que pode ser usado na análise do trabalho que estará sendo desenvolvido.

Fora isso, evidentemente esperamos que o aluno mantenha o bom desempenho (ou até melhore) no Bacharelado em Ciência da Computação. Durante o próximo ano, o aluno planeja cursar, como aluno especial da graduação, algumas disciplinas da nossa pós-graduação. Seu desempenho em tais disciplinas também poderá ser usado na sua

avaliação.

Referências

[1] Jaroslaw Byrka and Karen Aardal. An optimal bifactor approximation algorithm for the metric uncapacitated facility location problem. *SIAM Journal on Computing*, 39(6):2212–2231, 2010.

[2] Jarosław Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for k-median and positive correlation in budgeted optimization. *ACM Trans. Algorithms*, 13(2), 2017.

[3] Moses Charikar and Sudipto Guha. Improved combinatorial algorithms for facility location problems. *SIAM Journal on Computing*, 34(4):803–824, 2005.

[4] Marcelo H. de Carvalho, Márcia R. Cerioli, Ricardo Dahab, Paulo Feofiloff, Cristina G. Fernandes, Carlos E. Ferreira, Katia S. Guimarães, Flavio K. Miyazawa, José C. de Pina, José A. R. Soares, and Yoshiko Wakabayashi. *Uma Introdução Sucinta a Algoritmos de Aproximação*. 2001.

[5] Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

[6] Sudipto Guha and Samir Khuller. Greedy strikes back: Improved facility location algorithms. *Journal of Algorithms*, 31(1):228–248, 1999.

[7] Dorit Hochbaum and David Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research (MOR)*, 10:180–184, 05 1985.

[8] Wen-Lian Hsu and George L. Nemhauser. Easy and hard bottleneck location problems. *Discrete Applied Mathematics*, 1(3):209–215, 1979.

[9] Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In *Proceedings of the Thirty-Fourth Annual ACM*

Symposium on Theory of Computing, STOC'02, page 731–740, New York, NY, USA, 2002. Association for Computing Machinery.

[10] Shi Li. A 1.488 approximation algorithm for the uncapacitated facility location problem. *Information and Computation*, 222:45–58, 2013. 38th International Colloquium on Automata, Languages and Programming (ICALP 2011).

[11] Shi Li and Ola Svensson. Approximating k-median via pseudo-approximation. *CoRR*, abs/1211.0243, 2012.

[12] Vijay V. Vazirani. *Approximation algorithms.* Springer, 2001.

[13] David P. Williamson and David B. Shmoys. *The Design of Approximation Algorithms.* Cambridge University Press, 2011.