Análise e Comparação Empírica de Algoritmos 2-Aproximados para o Problema de K-Centros

Diogo T. Chaves[1], **Beatriz R. G. Barbosa**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

diogochaves@dcc.ufmg.br, beatriz.barbosa@dcc.ufmg.br

Resumo. *Este relatório apresenta os resultados de uma análise comparativa entre dois algoritmos 2-aproximados para o problema dos K-centros, além de comparar seus resultados com o algoritmo KMeans, como produto da disciplina Algoritmos II.*

1. Introdução

Zhi-Hua Zhou e Bo Liu em [Zhou and Liu 2002] descrevem um cluster como um conjunto de pontos que estão mais próximos entre si do que com os pontos de outros conjuntos. Com base nessa definição, o problema dos K-centros pode ser formulado da seguinte forma: dado um conjunto de pontos no espaço, o objetivo é dividir esses pontos em k clusters de maneira a minimizar a maior distância entre qualquer ponto e o centro do cluster ao qual ele pertence, conhecida como raio. Em outras palavras, queremos selecionar k centros de clusters de modo que a distância máxima entre os pontos e seus centros seja a menor possível.Um exemplo dessa divisão pode ser vista na Figura 1 no qual os clusters e seus raios podem ser diferidos pela cor, tendo seus centros são definidos e marcados por um 'X'.

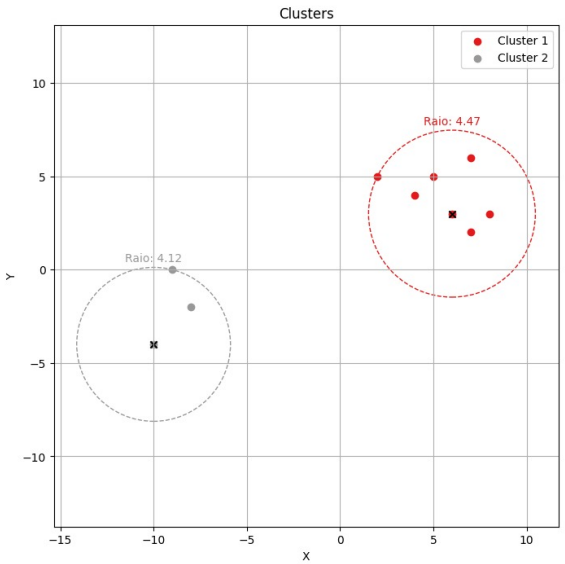

Figura 1. Exemplo de clusterização

Esse problema é bastante conhecido na computação e possui aplicações em redes de comunicações [Gupta et al. 2006], na área de visão computacional [Liu et al. 2020] e

até mesmo em logística [Kleinberg and Tardos 1999]. Todavia, um desafio significativo do problema é sua classificação como NP-completo por [Garey and Johnson 1979], não existindo solução polinomial para o ele, salvo se P=NP. Tendo isso em vista foram propostos os chamados algoritmos aproximativos para resolver esse problema que, apesar de não garantir a otimalidade, encontrem boas soluções em tempo polinomial. O problema de K-centros é limitado ao fator de aproximação 2, isto é, todos os algoritmos aproximativos para ele possuem solução ótima no máximo duas vezes pior que a solução ótima real.

Tendo em vista a importância desse problema e a dificuldade de suas soluções, o intuito desse trabalho é comparar o desempenho entre dois algoritmos-2 aproximativos entre si e com o algoritmo KMeans, uma ferrmenta eficaz na tarefa de clusterização.

2. Descrição dos Algoritmos

Ambos os algoritmos implementados têm fator de aproximação 2, o melhor conhecido para esse problema.

2.1. Refinamento do raio

O primeiro dos algoritmos implementados baseia-se no refinamento do raio ótimo até uma largura definida. Adotamos a seguinte estratégia:

- O intervalo inicial [0,r_{max}], em que rmax representa a maior distância entre 2 pontos, contém o valor ótimo do raio.

$$r_{max} = \max dist(s_i, s_j) \tag{1}$$

- Em seguida, determinamos se é possível agrupar os pontos com um raio r = r_{max}/2, utilizando uma função auxiliar. Se for possível, repetimos o processo como em uma busca binária, com o intervalo [0,r_{max}/2].
- Enquanto for possível, refinamos o intervalo até que ele seja menor que uma porcentagem determinada do raio inicial r_{max}.

2.2. Maximização de distâncias

O segundo algoritmo consiste na busca por centros que maximizem a distância dos pontos do conjunto aos centros já escolhidos. Adotamos o procedimento:

- Selecionamos um ponto qualquer do conjunto e fazemos dele um centro.
- Enquanto não obtivermos a quantidade desejada de clusters, escolhemos como próximo centro aquele que estiver mais distante daqueles já escolhidos.

2.3. KMeans

O terceiro algoritmo usado para comparação é o Kmeans, que funciona da seguinte maneira:

- São selecionados aleatoriamente k centróides, que são pontos iniciais que representam os centros dos clusters.
- Cada ponto de dado é então atribuído ao cluster cujo centróide está mais próximo.
- Após a atribuição de todos os pontos, o centróide de cada cluster é recalculado como a média dos pontos de dados que foram atribuídos a esse cluster.

3. Configuração e Metodologia dos Experimentos

Para a avaliação dos métodos, visando a realização de testes extensos que capturem o desempenho dos algoritmos para diferentes cenários e parâmetros, foi desenhada a seguinte metodologia:

- Foram escolhidos três tipos de dados, totalizando 30 conjuntos. Todos os datasets estão disponíveis neste repositório. A escolha consistiu em:
    1. Dez conjuntos de dados gerados a partir de uma distribuição normal multivariada, em torno do número de centros, variando controlando o desvio padrão para que a sobreposição entre os clusters varie entre inexistente até altamente sobrepostos. Os dados gerados podem ser obsverados na Figura 2

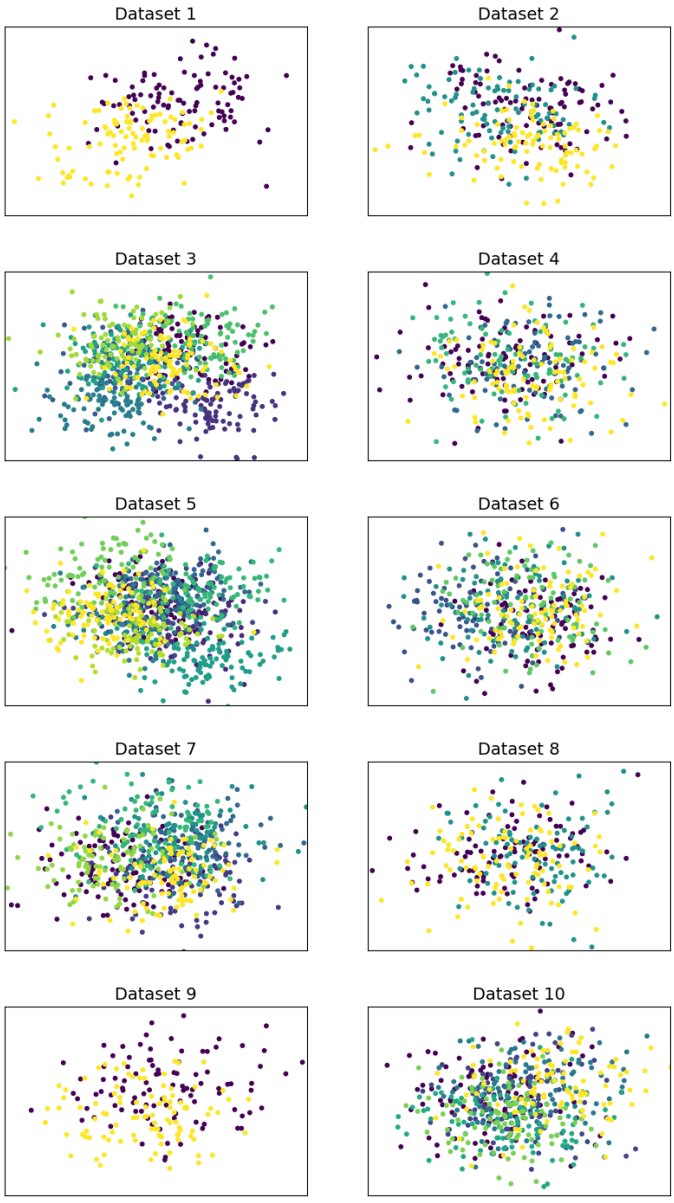

Figura 2. Dados gerados a partir da distribuição normal multivariada.

2. Dez conjuntos diversos disponíveis na biblioteca scikit-learn. Os dados selecionados estão representados na Figura 3

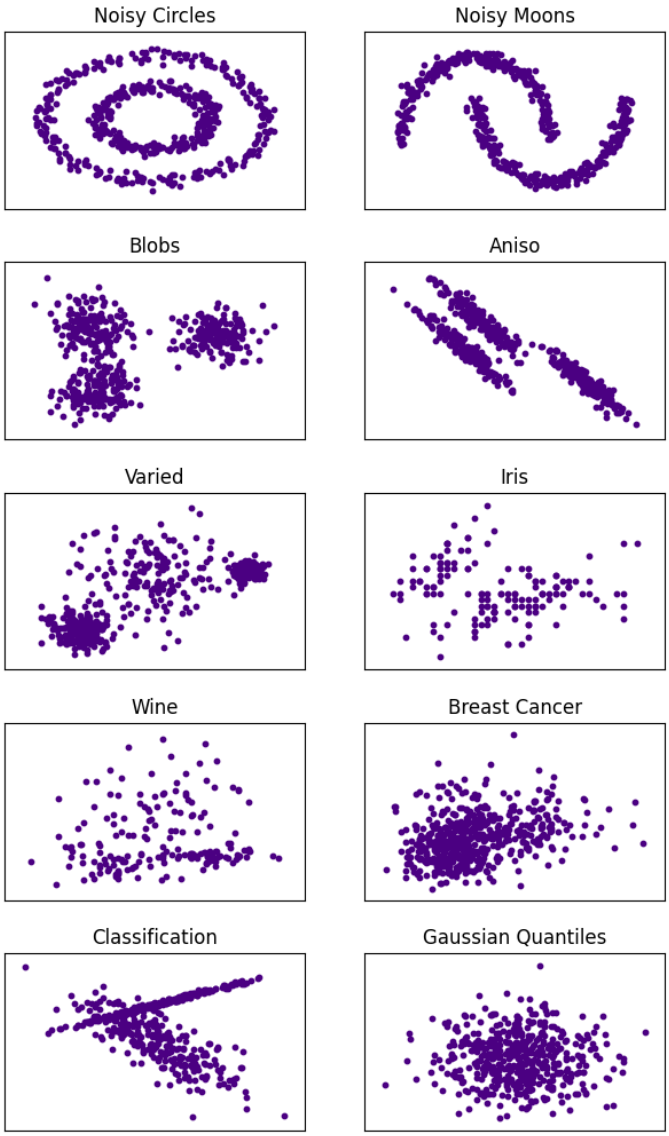

Figura 3. Dados disponíveis na biblioteca scikit-learn.

3. Dez conjuntos reais, selecionados no repositório UC Irvine Machine Learning Repository. No pré-processamento dos dados, para permitir a implementação da clusterização em duas dimensões, foram selecionados dois atributos de cada base de dados. Além disso, labels que identificam os clusters passaram por One-hot-encoding e a base foi reduzida por meio de uma seleção aleatória de instâncias. Os dados processados foram salvos em arquivos do tipo "csv", disponíveis no repositório do trabalho. A identificação de cada dataset está na Tabela 1

Dataset	ID
Hepatitis	503
Raisin	850
Mice Protein	342
Yeast	110
Maternal Risk	863
Credit	144
Adult	2
Diabetic Retinopathy	329
Dry Beans	602
Wine Quality	186

Tabela 1. Dataset IDs no repositório

- Para a execução do primeiro algoritmo, foram selecionados cinco valores distintos para determinar a porcentagem do raio inicial, ou seja, o refinamento. Os valores foram escolhidos em função do poder computacional disponível para a execução dos testes, além do objetivo de observar o desempenho do algoritmo para valores variados. As escolhas foram 0.14, 0.168, 0.195, 0.222 e 0.25.
- Para cada um tipos de dados, foram testados duas métricas de distância: Manhattan e Euclidiana.
- Para cada configuração de teste, foram realizadas 30 repetições, das quais foram extraídas a duração, o média dos maiores raios, o raio médio e as métricas silhueta e índice de Rand ajustado (ARI). Nas tabelas da seção 4, são apresentados tanto a média quanto o desvio padrão para os 30 testes.

Assim, obtemos 6 tabelas de resultados, apresentadas na seção 4, em que é reportados o desempenho de cada configuração de algoritmo para os 3 tipos de dados, com duas métricas de distância para cada.

4. Análise e Discussão dos Resultados

Os algoritmos foram executados conforme especificado na 3 para todos os 30 conjuntos de dados. Para facilitar a análise, agrupamos os resultados dos conjuntos de dados que tinham a mesma origem (gerados a partir de uma normal multivariada, gerados no scikit learn ou dados reais) e tiramos a média, representada pelo número à esquerda, e o desvio padrão,representado pelo número da direita logo após o "+-".

4.1. Conjunto de Dados Gerado a Partir de uma Normal Multivariada

Modelo	Duração (ms)	Maior Raio	Média Raios	Silhueta	ARI
Alg 1 (0.14)	36.69 ± 37.36	3.81 ± 0.99	2.83 ± 1.03	0.205 ± 0.11	0.031 ± 0.06
Alg 1 (0.168)	37.92 ± 38.22	3.81 ± 0.99	2.83 ± 1.03	0.205 ± 0.11	0.031 ± 0.06
Alg 1 (0.195)	37.72 ± 37.57	3.81 ± 0.99	2.83 ± 1.03	0.205 ± 0.11	0.031 ± 0.06
Alg 1 (0.222)	38.98 ± 39.41	3.81 ± 0.99	2.83 ± 1.03	0.205 ± 0.11	0.031 ± 0.06
Alg 1 (0.25)	38.90 ± 38.06	3.81 ± 0.99	2.83 ± 1.03	0.205 ± 0.11	0.031 ± 0.06
Alg 2	250.70 ± 370.92	3.06 ± 0.89	2.64 ± 0.78	0.265 ± 0.13	0.123 ± 0.27
Kmeans	63.11 ± 66.57	2.36 ± 0.44	1.88 ± 0.46	0.364 ± 0.10	0.162 ± 0.29

Tabela 2. Resultado Médio dos Algoritmos Utilizando a Distância de Manhattan

Modelo	Duração (ms)	Maior Raio	Média Raios	Silhueta	ARI
Alg 1 (0.14)	37.16 ± 37.41	2.83 ± 0.74	1.998 ± 0.71	0.213 ± 0.12	0.026 ± 0.04
Alg 1 (0.168)	37.51 ± 39.23	2.83 ± 0.74	1.998 ± 0.71	0.213 ± 0.12	0.026 ± 0.04
Alg 1 (0.195)	38.03 ± 40.00	2.83 ± 0.74	1.998 ± 0.71	0.213 ± 0.12	0.026 ± 0.04
Alg 1 (0.222)	38.43 ± 39.47	2.83 ± 0.74	1.998 ± 0.71	0.213 ± 0.12	0.026 ± 0.04
Alg 1 (0.25)	38.76 ± 40.78	2.83 ± 0.74	1.998 ± 0.71	0.213 ± 0.12	0.026 ± 0.04
Alg 2	245.01 ± 346.16	2.35 ± 0.58	2.105 ± 0.52	0.247 ± 0.12	0.103 ± 0.22
Kmeans	61.32 ± 70.26	2.36 ± 0.44	1.878 ± 0.46	0.364 ± 0.10	0.162 ± 0.29

Tabela 3. Resultado Médio dos Algoritmos Utilizando a Distância Euclidiana

Nesse conjunto de dados, é interessante notar que os resultados do primeiro modelo não variaram em relação ao percentual do raio inicial. O limiar 0.14 foi escolhido devido à nossa capacidade computacional, tendo em vista que a execução para valores menores era muito custosa ou interrompida. Escolhas de percentual menores ocasionariam em uma mudança significativa na duração da execução e, provavelmente, nas métricas de qualidade. Ainda sobre esse modelo, apesar de ele possuir valores de silhueta próximos aos demais, seus raios são grandes e seu índice de Rand ajustado é muito próximo de 0, o que significa que seus agrupamentos não são muito melhores do que os gerados pelo acaso.

O segundo modelo, por sua vez, apresentou um índice de Rand ajustado bem próximo ao do KMeans, todavia seus valores de raio são bem maiores, apresentando desvios padrões altos, o que significa que existe uma maior variação de tamanho entre eles.

Em relação ao tempo, todos os modelos apresentam valores de desvio padrão próximos à média, significando uma variação imensa. Definitivamente o tempo do modelo 2 é muito maior que os demais, todavia vale lembrar que o modelo 1 está relacionado ao seus percentuais e que valores menores que 0.14 ocasionaram em um tempo inviável e, com isso, muito maior que o do modelo 2.

4.2. Dataset Gerado no Scikit Learn

Modelo	Duração (ms)	Maior Raio	Média Raios	Silhueta	ARI
Alg 1 (0.14)	23.23 ± 16.05	7.78 ± 7.41	5.875 ± 5.18	0.145 ± 0.24	0.132 ± 0.21
Alg 1 (0.168)	24.54 ± 17.42	7.78 ± 7.41	5.875 ± 5.18	0.145 ± 0.24	0.132 ± 0.21
Alg 1 (0.195)	25.69 ± 18.82	7.78 ± 7.41	5.875 ± 5.18	0.145 ± 0.24	0.132 ± 0.21
Alg 1 (0.222)	23.99 ± 17.06	7.78 ± 7.41	5.875 ± 5.18	0.145 ± 0.24	0.132 ± 0.21
Alg 1 (0.25)	24.13 ± 17.06	7.78 ± 7.41	5.875 ± 5.18	0.145 ± 0.24	0.132 ± 0.21
Alg 2	54.51 ± 34.72	6.45 ± 6.70	6.029 ± 6.40	0.291 ± 0.19	0.250 ± 0.21
Kmeans	56.53 ± 59.82	337.27 ± 900.24	207.21 ± 551.48	0.530 ± 0.12	0.448 ± 0.32

Tabela 4. Resultado Médio dos Algoritmos Utilizando a Distância de Manhattan

Modelo	Duração (ms)	Maior Raio	Média Raios	Silhueta	ARI
Alg 1 (0.14)	39.45 ± 40.42	5.64 ± 4.71	4.65 ± 4.08	0.212 ± 0.25	0.136 ± 0.21
Alg 1 (0.168)	29.54 ± 19.14	5.65 ± 4.71	4.65 ± 4.08	0.211 ± 0.25	0.134 ± 0.21
Alg 1 (0.195)	26.19 ± 18.12	5.67 ± 4.69	4.70 ± 4.03	0.188 ± 0.26	0.128 ± 0.21
Alg 1 (0.222)	26.08 ± 18.93	5.67 ± 4.69	4.70 ± 4.03	0.188 ± 0.26	0.128 ± 0.21
Alg 1 (0.25)	26.11 ± 19.29	5.67 ± 4.69	4.70 ± 4.03	0.188 ± 0.26	0.128 ± 0.21
Alg 2	55.07 ± 32.03	4.90 ± 4.91	4.55 ± 4.84	0.323 ± 0.17	0.260 ± 0.22
Kmeans	42.65 ± 42.68	337.27 ± 900.24	207.21 ± 551.48	0.530 ± 0.12	0.448 ± 0.32

Tabela 5. Resultado Médio dos Algoritmos Utilizando a Distância Euclidiano

Nesse conjunto de dados, é interessante notar que os resultados das métricas são muito superiores ao do primeiro conjunto de dados para todos modelos, pois, uma vez que foram obtidos na biblioteca Scikit Learn, é provável que tenham passado por algum processamento. Como no primeiro conjunto de dados, é interessante notar que os resultados do primeiro modelo não variaram significativamente em relação ao percentual do raio inicial, provavelmente pelo mesmo motivo.

A ordem de qualidade dos modelos é a mesma do conjunto de dados anterior, sendo o KMeans o melhor e o algoritmo 1 o pior. Nesse conjunto de dados, os raios encontrados pelo algoritmo KMeans contém dois outliers (nos datasets wine quality e breast cancer), o que impactou significativamente as médias e desvios obtidos.

Observamos aqui métricas superiores às da seção 4.1, provavalmente devido à natureza dos dados selecionados.

Em relação ao tempo, todos os modelos apresentam valores de desvio padrão próximos à média, significando uma variação imensa. Todavia, comparando com o primeiro conjunto de dados, observamos uma diminuição do tempo, possivelmente porque, uma vez que os dados são reais, é mais provável que eles apresentem tendências mais claras que os dados sintéticos gerados ao acaso, o que facilita a tarefa de clusterização.

4.3. Datasets Reais

Modelo	Duração (ms)	Maior Raio	Média Raios	Silhueta	ARI
Alg 1 (0.14)	54.69 ± 7.15	38009.70 ± 95782.30	33121.11 ± 82456.99	0.452 ± 0.19	0.001 ± 0.01
Alg 1 (0.168)	56.22 ± 7.34	38009.70 ± 95782.30	33121.11 ± 82456.99	0.452 ± 0.19	0.001 ± 0.01
Alg 1 (0.195)	56.93 ± 6.54	38009.70 ± 95782.30	33121.11 ± 82456.99	0.452 ± 0.19	0.001 ± 0.01
Alg 1 (0.222)	56.19 ± 6.93	38009.70 ± 95782.30	33121.11 ± 82456.99	0.452 ± 0.19	0.001 ± 0.01
Alg 1 (0.25)	55.86 ± 8.65	38009.70 ± 95782.30	33121.11 ± 82456.99	0.452 ± 0.19	0.001 ± 0.01
Alg 2	224.50 ± 167.88	37746.44 ± 94724.08	30290.93 ± 72482.27	0.488 ± 0.10	0.047 ± 0.09
Kmeans	119.07 ± 23.58	24747.58 ± 54912.51	21564.26 ± 49588.35	0.517 ± 0.10	0.071 ± 0.12

Tabela 6. Resultado Médio dos Algoritmos Utilizando a Distância de Manhattan

Modelo	Duração (ms)	Maior Raio	Média Raios	Silhueta	ARI
Alg 1 (0.14)	55.02 ± 7.42	37843.45 ± 95291.44	32820.42 ± 81551.28	0.513 ± 0.21	0.0012 ± 0.01
Alg 1 (0.168)	55.09 ± 7.57	37843.45 ± 95291.44	32820.42 ± 81551.28	0.513 ± 0.21	0.0012 ± 0.01
Alg 1 (0.195)	56.61 ± 8.48	37843.45 ± 95291.44	32820.42 ± 81551.28	0.513 ± 0.21	0.0012 ± 0.01
Alg 1 (0.222)	56.74 ± 6.85	37843.45 ± 95291.44	32820.42 ± 81551.28	0.513 ± 0.21	0.0012 ± 0.01
Alg 1 (0.25)	57.96 ± 8.92	37843.45 ± 95291.44	32820.42 ± 81551.28	0.513 ± 0.21	0.0012 ± 0.01
Alg 2	237.39 ± 192.46	37672.53 ± 94521.23	30032.65 ± 71732.18	0.477 ± 0.13	0.051 ± 0.09
Kmeans	111.95 ± 40.35	24747.58 ± 54912.51	21564.26 ± 49588.35	0.517 ± 0.10	0.071 ± 0.12

Tabela 7. Resultado Médio dos Algoritmos Utilizando a Distância Euclidiana

Nesse conjunto de dados, o que se destaca é o aumento expressivo das médias e desvios relativos aos raios, o que ocorre naturalmente devido ao comportamento dos dados selecionados.

Em relação aos demais aspectos, os dados estão consistentes com os resultados apresentados anteriormente, em que o segundo modelo apresenta um tempo maior de execução, conforme discutido previamente. As métricas estão consistentes com a análise apresentada na seção 4.1.

Finalmente, apesar de os dados terem maior inclinação a apresentarem tendências mais claras, assim como os dados reais da seção 4.2, o maior volume de dados provavelmente impactou negativamente o tempo de execução.

5. Conclusão

Em resumo, ambos os algoritmos têm suas vantagens e desvantagens dependendo do cenário de aplicação. Os experimentos mostraram que, embora o algoritmo de maximização de distâncias forneça soluções mais consistentes em termos de qualidade de clusterização em relação ao primeiro, ele também é mais custoso em termos de tempo de execução. O algoritmo de refinamento do raio, apesar de simples, demonstrou limitações, especialmente em termos de flexibilidade e adaptação a diferentes cenários. Vale ressaltar que, nos cenários em que estiver disponível maior poder computacional, o refinamento do raio provavelmente apresentará melhores resultados, apesar do aumento no custo temporal da execução.

Referências

Garey, M. R. and Johnson, D. S. (1979). A Guide to the Theory of NP-Completeness. Wiley, New York.

Gupta, S., Agarwal, P. K., and Yao, S. L. (2006). A k-center problem for network design. IEEE Transactions on Networking, 14(3):519–530.

Kleinberg, J. M. and Tardos, E. (1999). Facility location and allocation: The k-center problem. Operations Research, 47(1):20–32.

Liu, X., Lin, W., and Chen, R. J. (2020). Image segmentation using k-center clustering. IEEE Transactions on Image Processing, 29(8):4432–4445.

Zhou, Z.-H. and Liu, B. (2002). A comparative evaluation of clustering algorithms. IEEE Transactions on Knowledge and Data Engineering, 14(5):930–944.