Trabalho Prático 2 - Algoritmos 2

Etelvina C. S. Sá Oliveira[1], **Vitor L. de Faria**[1]

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

{etelvina.oliveira2003, vtifaria} @gmail.com

Abstract. *This work aims to analyze two 2-approximation algorithms for the k-centers problem and the KMeans algorithm, considering computational demand and solution quality.*

Resumo. *O presente trabalho tem como objetivo realizar a análise de dois algoritmos 2-aproximativos para o problema dos k-centros e o KMeans, levando em consideração aspectos de demanda computacional e qualidade da solução.*

1. Introdução

Este trabalho prático aborda a implementação e análise de algoritmos para o problema dos k-centros, um desafio central em tarefas de agrupamento em aprendizado de máquina. Foram implementados dois algoritmos 2-aproximativos e o algoritmo clássico K-Means, com o objetivo de comparar demanda computacional e qualidade das soluções.

O problema dos k-centros consiste em escolher k centros de forma a minimizar a maior distância entre um ponto e seu centro mais próximo (raio). O centro de um ponto é o ponto do conjunto de centros que possui a menor distância do ponto considerado. Os algoritmos aproximativos utilizados oferecem soluções com um raio de no máximo duas vezes o ótimo. Além disso, o K-Means foi usado como referência devido à sua eficácia conhecida em tarefas de agrupamento.

Os experimentos foram realizados utilizando datasets reais do UCI Machine Learning Repository e datasets sintéticos, analisando o desempenho dos algoritmos em termos de tempo de execução e qualidade do agrupamento. A análise visa entender a aplicabilidade prática desses algoritmos em diferentes contextos.

2. Métodos e métricas

No trabalho em questão foram implementados três algoritmos que realizam a tarefa de agrupamento, sendo dois deles 2-aproximativos e um algoritmo clássico para o problema de agrupamento, o K-Means.

O primeiro algoritmo aproximativo implementado (k-centros1) se baseia no refinamento de intervalos, em que usamos um algoritmo auxiliar que verifica se é possível gerar um agrupamento com até k centros com o raio determinado. Nesse algoritmo, a cada iteração selecionamos um ponto arbitrário do conjunto de pontos adicionando ele ao conjunto de centros e removemos todos os pontos que estão a uma distância de até 2*r do último ponto considerado. Esse processo é realizado até que todos os pontos tenham sido avaliados e o conjunto de pontos esteja vazio. Por fim, verificamos se o número de

centros encontrado é menor ou igual ao número de centros desejado. Com a aplicação desse algoritmo auxiliar, partimos de um intervalo que vai de 0 até a maior distância entre dois pontos do dataset, que é o intervalo que garantidamente contém o raio ótimo, o raio que minimiza a maior distância de um ponto ao seu centro, e o refinamos através de uma busca binária, em que a cada iteração verificamos se é possível encontrar um agrupamento de tamanho até k com o raio dado pelo ponto médio do limite superior e inferior do intervalo. Nessa implementação recebemos um parâmetro que define a porcentagem do intervalo real que o intervalo refinado deve atingir (referência). Por fim, realizamos uma busca binária no intervalo refinado para a seleção de um valor de raio que gera um conjunto de centros de tamanho menor ou igual a k.

O segundo algoritmo aproximativo implementado (k-centros2) inicialmente verifica se o número de centros é maior que o número de pontos, retornando caso seja verdadeiro o conjunto de pontos em si, senão a cada iteração escolhemos um ponto arbitrário do conjunto de pontos e, enquanto o numero de pontos no conjunto de centros é menor que o número de centros desejado, escolhemos o ponto que maximiza a distância entre os pontos que atualmente estão no conjunto de centros.

Em cada um dos algoritmos aproximativos foi utilizada como medida de distância a distância de Minkowski, que é definida da seguinte forma:

$$D(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

Quando atribuímos p = 1 ela é equivalente a distância de Manhattan e quando atribuímos p = 2 ela é equivalente a distância Euclidiana.

O terceiro algoritmo utilizado foi o K-Means, conhecido por sua simplicidade e eficácia em agrupamento. O K-Means é um algoritmo que treina um modelo para agrupar objetos semelhantes. Para isso, ele mapeia cada observação no conjunto de dados de entrada para um ponto no espaço de n dimensões (em que n é o número de atributos da observação).

Foram utilizadas três métricas para a avaliação dos algoritmos: o tempo de execução, o raio da solução, a silhueta e o índice de Rand ajustado. No caso das métricas de silhueta e índice de Rand ajustado foi utilizada a implementação disponibilizada pela biblioteca scikit-learn (https://scikit-learn.org/stable/index.html).

O raio da solução é dado como a maior distância de um ponto ao seu centro, sendo uma medida da dispersão dos pontos dentro do cluster.

O Coeficiente de Silhueta é calculado usando a distância média intracluster e a distância média do cluster mais próximo para cada amostra. O melhor valor é 1 e o pior valor é -1. Valores próximos a 0 indicam clusters sobrepostos. Valores negativos geralmente indicam que uma amostra foi atribuída ao cluster errado, pois um cluster diferente é mais similar.

O Índice Rand calcula uma medida de similaridade entre dois agrupamentos considerando todos os pares de amostras e contando os pares que são atribuídos no mesmo agrupamento ou em agrupamentos diferentes nos agrupamentos previstos e verdadeiros.O

índice Rand ajustado tem um valor próximo de 0,0 para rotulagem aleatória independentemente do número de clusters e amostras e exatamente 1,0 quando os clusters são idênticos. O índice Rand ajustado é limitado abaixo por -0,5 para clusters especialmente discordantes.

3. Bases de dados

Para a realização do trabalho utilizamos os dados de 3 fontes principais a UCI Machine Learning Repository, a geração sintética 1 por meio do scikit, e a geração sintética 2 em duas dimensões utilizando a distribuição normal multivariada. Todas as bases de dados consideradas possuem um número de instâncias maior que 700.

3.1. UCI Machine Learning Repository

Foram utilizados 10 conjuntos de dados da base de dados do UCI Machine Learning Repository, esses foram escolhidos baseados no critério de possuírem a maioria dos atributos em formato numérico, conforme solicitado no enunciado do trabalho.

3.1.1. Raisin

Imagens das variedades de passas Kecimen e Besni cultivadas na Turquia foram obtidas com CVS. Um total de 900 grãos de passas foram usados, incluindo 450 pedaços de ambas as variedades. Essas imagens foram submetidas a vários estágios de pré-processamento e 7 características morfológicas foram extraídas. Essas características foram classificadas usando três técnicas diferentes de inteligência artificial.

3.1.2. Image Segmentation

As instâncias foram extraídas aleatoriamente de um banco de dados de 7 imagens externas. As imagens foram segmentadas manualmente para criar uma classificação para cada pixel. Cada instância é uma região 3x3.

3.1.3. Maternal Health Risk

Dados foram coletados de diferentes hospitais, clínicas comunitárias e unidades de saúde materna de áreas rurais de Bangladesh por meio do sistema de monitoramento de risco baseado em IoT.

3.1.4. Mammographic Mass

Discriminação de massas mamográficas benignas e malignas com base nos atributos BI-RADS e na idade do paciente.

3.1.5. Optical Recognition of Handwritten Digits

Usamos programas de pré-processamento disponibilizados pelo NIST para extrair bitmaps normalizados de dígitos manuscritos de um formulário pré-impresso. De um total de 43 pessoas, 30 contribuíram para o conjunto de treinamento e 13 diferentes para o conjunto de teste. Bitmaps 32x32 são divididos em blocos não sobrepostos de 4x4 e o número de pixels é contado em cada bloco. Isso gera uma matriz de entrada de 8x8 onde cada elemento é um inteiro no intervalo de 0 a 16. Isso reduz a dimensionalidade e dá invariância a pequenas distorções.

3.1.6. Statlog (Vehicle Silhouettes)

Objetos 3D dentro de uma imagem 2D pela aplicação de um conjunto de extratores de características de forma às silhuetas 2D dos objetos.

3.1.7. Yeast

Atributo previsto: Sítio de localização da proteína. (não numérico). As referências abaixo descrevem um predecessor deste conjunto de dados e seu desenvolvimento. Elas também fornecem resultados (não validados de forma cruzada) para classificação por um sistema especialista baseado em regras com essa versão do conjunto de dados.

3.1.8. Blood Transfusion Service Center

Para demonstrar o modelo de marketing RFMTC (uma versão modificada do RFM), este estudo adotou o banco de dados de doadores do Blood Transfusion Service Center na cidade de Hsin-Chu em Taiwan. O centro passa seu ônibus de serviço de transfusão de sangue para uma universidade na cidade de Hsin-Chu para coletar sangue doado a cada três meses. Para construir um modelo FRMTC, selecionamos 748 doadores aleatoriamente do banco de dados de doadores. Esses 748 dados de doadores, cada um incluía R (Recência - meses desde a última doação), F (Frequência - número total de doações), M (Monetário - total de sangue doado em cc), T (Tempo - meses desde a primeira doação) e uma variável binária representando se ele/ela doou sangue em março de 2007 (1 significa doação de sangue; 0 significa não doação de sangue).

3.1.9. Abalone

Prevendo a idade do abalone a partir de medições físicas. A idade do abalone é determinada cortando a concha através do cone, colorindo-a e contando o número de anéis através de um microscópio – uma tarefa chata e demorada. Outras medições, que são mais fáceis de obter, são usadas para prever a idade. Mais informações, como padrões climáticos e localização (portanto, disponibilidade de alimentos) podem ser necessárias para resolver o problema. Dos dados originais, exemplos com valores ausentes foram removidos (a maioria com o valor previsto ausente), e os intervalos dos valores contínuos foram dimensionados para uso com uma ANN (dividindo por 200).

3.1.10. Diabetic Retinopathy Debrecen

Este conjunto de dados contém recursos extraídos do conjunto de imagens Messidor para prever se uma imagem contém sinais de retinopatia diabética ou não. Todos os recursos representam uma lesão detectada, um recurso descritivo de uma parte anatômica ou um descritor de nível de imagem.

3.2. Sintética 1

A primeira abordagem para a geração dos 10 conjuntos de dados sintéticos baseou-se no material "Comparing different clustering algorithms on toy datasets"usando scikit-learn. Foram gerados datasets como noisy circles e noisy moons, que incluem formas não lineares e ruído, além de blobs com diferentes variâncias e um dataset anisotrópico resultante de uma transformação linear em blobs. Também foram criados conjuntos baseados em quantis gaussianos. Cada dataset foi salvo em formato CSV, facilitando sua utilização nos experimentos de clustering.

3.3. Sintética 2

Na segunda abordagem, foram gerados 10 conjuntos de dados sintéticos em duas dimensões utilizando a distribuição normal multivariada. Para cada conjunto, foram criados centros de distribuição com médias aleatórias, variando o desvio padrão para controlar a sobreposição entre os grupos, desde inexistente até altamente sobrepostos. A função generate multivariate data foi utilizada para amostrar pontos em torno desses centros, resultando em uma distribuição variada de pontos para cada centro. Cada conjunto foi salvo em formato CSV, com os rótulos dos centros incluídos para identificar a origem dos pontos.

4. Descrição dos experimentos

Os experimentos foram conduzidos da seguinte forma: no caso dos algoritmos aproximativos, a escolha dos centros é arbitrária e pode influenciar na qualidade da solução, desse modo foram realizados 30 experimentos para cada conjunto de dados. Além disso, para cada uma das execuções calculamos cada uma das métricas mencionadas para possibilitar a comparação.

Para auxiliar a execução dos algoritmos aproximados foi calculada uma matriz de distâncias de todos os pontos, com o objetivo de otimizar a execução dos algoritmos, visto que é necessário realizar seu cálculo apenas uma vez para cada base de dados. Também foram realizadas 30 iterações do algoritmo K-Means para cada um dos datasets para possibilitar a comparação com os demais algoritmos.

Para cada um dos datasets foi realizada uma etapa de pré processamento em que realizamos a remoção de registros com valores nulos e realizamos um mapeamento dos possíveis valores da variável alvo para valores numéricos, nos casos em que ela é categórica.

A função carregar datasets é responsável por carregar os datasets sintéticos previamente gerados. Dependendo do tipo de dataset (sintético 1 ou 2), ela carrega os arquivos CSV correspondentes.

Após o processamento de todos os datasets, os resultados são organizados em dataframes e salvos em arquivos CSV. Isso permite uma análise posterior dos resultados

para comparar o desempenho dos algoritmos em diferentes tipos de dados. No arquivo CSV é armazedo o nome do dataset, o algoritmo utilizado, além da média e do desvio padrão de cada uma das métricas utilizadas.

4.1. K-centros 1

Nos experimentos utilizando o algoritmo K-centros 1 utilizamos como parâmetro de referência os valores 5%, 10%, 15%, 20% e 25%. Desse modo, foram realizadas 30 iterações para cada um desses parâmetros e com cada uma das medidas de distância.

Assim, inicialmente utilizamos uma função que calcula a distância máxima entre dois pontos do dataset e refinamos o intervalo que ele tenha x% do tamanho do intervalo original.

Desse modo, foi criada a função experimentos-kcentros1 que realiza 30 iterações para cada base de dados considerando cada um dos valores possíveis do parâmetro de referência e as medidas de distância. Além das informações já descritas que são salvas no CSV, também armazenamos qual o valor do parâmetro de referência utilizado.

4.2. K-centros 2

Os experimentos utilizando o K-centros 2 foram realizadas de forma semelhante ao K-centros 1. Para isso, foi criada a função experimentos-kcentros2, que realiza 30 iterações para cada base de dados considerando cada uma das medidas de distância.

4.3. K-means

Os experimentos com o algoritmo K-Means foram organizados de forma estruturada no código, começa com a função mainKmeans, que coordena todo o processo de execução.

Para cada dataset, a função processardatasetskmeans é chamada. Esta função aplica o K-Means nos dados,depois determina o número de clusters com base no número de classes únicas (y.nunique()).

A execução do K-Means é realizada pela função kmeansexec, que roda o algoritmo 30 vezes para cada dataset, variando a inicialização dos centroides.

Além dos datasets sintéticos, a função datasetUCI processa os datasets reais obtidos do UCI Machine Learning Repository. Ela carrega os dados, ajusta-os conforme necessário (remove colunas ou lida com valores ausentes), e executa o K-Means,no final, salva os resultados em arquivos CSV específicos.

5. Apresentação dos Resultados

Apresentamos abaixo alguns dos resultados obtidos após a execução dos 30 conjuntos de dados, os resultados completos estão disponíveis no repositório do github.

Em cada um das métricas os resultados são apresentados no formato: média+-desvio padrão.

Tabela 1. Resultados dos Experimentos de Clusterização com K-Centros 2 Distância de Manhattan

Dataset	Algoritmo	Tempo Execução (s)	Silhueta	Índice de Rand Ajustado	Raio Máximo
Dataset Sintético Scikit Learn 0	K-Centros 2	0.0005 ± 0.0007	0.3218 ± 0.0000	0.0081 ± 0.0000	1.6413 ± 0.0000
Dataset Sintético Scikit Learn 1	K-Centros 2	0.0006 ± 0.0007	0.4067 ± 0.0000	-0.0014 ± 0.0000	1.7892 ± 0.0000
Dataset Sintético Scikit Learn 2	K-Centros 2	0.0003 ± 0.0001	0.2903 ± 0.0000	0.0427 ± 0.0000	2.2675 ± 0.0000
Dataset Sintético Scikit Learn 3	K-Centros 2	0.0003 ± 0.0000	0.2862 ± 0.0000	0.0380 ± 0.0000	2.3329 ± 0.0000
Dataset Sintético Scikit Learn 4	K-Centros 2	0.0010 ± 0.0001	0.5798 ± 0.0000	0.8045 ± 0.0000	11.4368 ± 0.0000
Dataset Sintético Scikit Learn 5	K-Centros 2	0.0012 ± 0.0006	0.5278 ± 0.0000	0.6722 ± 0.0000	10.3719 ± 0.0000
Dataset Sintético Scikit Learn 6	K-Centros 2	0.0010 ± 0.0000	0.3028 ± 0.0000	0.2174 ± 0.0000	4.8318 ± 0.0000
Dataset Sintético Scikit Learn 7	K-Centros 2	0.0010 ± 0.0000	0.3506 ± 0.0000	0.5066 ± 0.0000	9.5999 ± 0.0000
Dataset Sintético Scikit Learn 8	K-Centros 2	0.0010 ± 0.0000	0.4429 ± 0.0000	0.5867 ± 0.0000	9.0621 ± 0.0000
Dataset Sintético Scikit Learn 9	K-Centros 2	0.0001 ± 0.0000	0.2196 ± 0.0000	0.1639 ± 0.0000	3.6349 ± 0.0000

Tabela 2. Resultados dos Experimentos de Clusterização com K-Centros 2 Distância Euclidiana

Dataset	Algoritmo	Tempo Execução (s)	Silhueta	Índice de Rand Ajustado	Raio Máximo
Dataset Sintétido Distribuição Multivariada 0	K-Centros 2	0.0017 ± 0.0010	0.4874 ± 0.0000	0.6875 ± 0.0000	14.0229 ± 0.0000
Dataset Sintétido Distribuição Multivariada 1	K-Centros 2	0.0055 ± 0.0031	0.6053 ± 0.0000	0.5194 ± 0.0000	6.1819 ± 0.0000
Dataset Sintétido Distribuição Multivariada 2	K-Centros 2	0.0041 ± 0.0004	-0.0070 ± 0.0000	0.1937 ± 0.0000	11.1692 ± 0.0000
Dataset Sintétido Distribuição Multivariada 3	K-Centros 2	0.0024 ± 0.0002	0.4351 ± 0.0000	0.3266 ± 0.0000	7.3422 ± 0.0000
Dataset Sintétido Distribuição Multivariada 4	K-Centros 2	0.0077 ± 0.0007	0.6392 ± 0.0000	0.4481 ± 0.0000	11.3726 ± 0.0000
Dataset Sintétido Distribuição Multivariada 5	K-Centros 2	0.0018 ± 0.0002	0.1411 ± 0.0000	0.3267 ± 0.0000	18.9975 ± 0.0000
Dataset Sintétido Distribuição Multivariada 6	K-Centros 2	0.0074 ± 0.0007	0.2143 ± 0.0000	0.1335 ± 0.0000	18.3425 ± 0.0000
Dataset Sintétido Distribuição Multivariada 7	K-Centros 2	0.0053 ± 0.0002	0.2985 ± 0.0000	0.3614 ± 0.0000	14.5563 ± 0.0000
Dataset Sintétido Distribuição Multivariada 8	K-Centros 2	0.0066 ± 0.0009	0.2588 ± 0.0000	0.1981 ± 0.0000	11.6828 ± 0.0000
Dataset Sintétido Distribuição Multivariada 9	K-Centros 2	0.0049 ± 0.0005	0.4447 ± 0.0000	0.5085 ± 0.0000	13.3017 ± 0.0000

Tabela 3. Resultados dos Experimentos de Clusterização com K-Centros 1 Distância Euclidiana

Dataset	Algoritmo	Referência	Tempo Execução (s)	Silhueta	Índice de Rand Ajustado	Raio Máximo
Dataset Sintético Scikit Learn 1	K-Centros 1	0.05	0.1699 ± 0.0051	-1.0000 ± 0.0000	0.0000 ± 0.0000	1.9000 ± 0.0000
Dataset Sintético Scikit Learn 1	K-Centros 1	0.1	0.2245 ± 0.0701	-1.0000 ± 0.0000	0.0000 ± 0.0000	1.9000 ± 0.0000
Dataset Sintético Scikit Learn 1	K-Centros 1	0.15	0.1692 ± 0.0057	-1.0000 ± 0.0000	0.0000 ± 0.0000	1.9000 ± 0.0000
Dataset Sintético Scikit Learn 1	K-Centros 1	0.2	0.1775 ± 0.0323	-1.0000 ± 0.0000	0.0000 ± 0.0000	1.9000 ± 0.0000
Dataset Sintético Scikit Learn 1	K-Centros 1	0.25	0.2090 ± 0.0658	-1.0000 ± 0.0000	0.0000 ± 0.0000	1.9000 ± 0.0000

Tabela 4. Resultados dos Experimentos de Clusterização com K-Centros 1 Distância de Manhattan

Dataset	Algoritmo	Referência	Tempo Execução (s)	Silhueta	Índice de Rand Ajustado	Raio Máximo
Dataset Sintético Scikit Learn 1	K-Centros 1	0.05	0.1853 ± 0.0523	-1.0000 ± 0.0000	0.0000 ± 0.0000	2.6117 ± 0.0000
Dataset Sintético Scikit Learn 1	K-Centros 1	0.1	0.1690 ± 0.0069	-1.0000 ± 0.0000	0.0000 ± 0.0000	2.6117 ± 0.0000
Dataset Sintético Scikit Learn 1	K-Centros 1	0.15	0.2179 ± 0.0680	-1.0000 ± 0.0000	0.0000 ± 0.0000	2.6117 ± 0.0000
Dataset Sintético Scikit Learn 1	K-Centros 1	0.2	0.1721 ± 0.0057	-1.0000 ± 0.0000	0.0000 ± 0.0000	2.6117 ± 0.0000
Dataset Sintético Scikit Learn 1	K-Centros 1	0.25	0.2275 ± 0.0670	-1.0000 ± 0.0000	0.0000 ± 0.0000	2.6117 ± 0.0000

Tabela 5. Resultados dos Experimentos de Clusterização com K-Centros 1 Distância Euclidiana

Dataset	Algoritmo	Referência	Tempo Execução (s)	Silhueta	Índice de Rand Ajustado	Raio Máximo
Dataset Sintético Distribuição Multivariada 9	K-Centros 1	0.05	0.2351 ± 0.0041	0.2859 ± 0.0000	0.3825 ± 0.0000	11.0314 ± 0.0000
Dataset Sintético Distribuição Multivariada 9	K-Centros 1	0.1	0.2818 ± 0.0651	0.4338 ± 0.0000	0.4725 ± 0.0000	11.0314 ± 0.0000
Dataset Sintético Distribuição Multivariada 9	K-Centros 1	0.15	0.2629 ± 0.0536	0.4338 ± 0.0000	0.4725 ± 0.0000	11.0314 ± 0.0000
Dataset Sintético Distribuição Multivariada 9	K-Centros 1	0.2	0.2537 ± 0.0445	0.4338 ± 0.0000	0.4725 ± 0.0000	11.0314 ± 0.0000
Dataset Sintético Distribuição Multivariada 9	K-Centros 1	0.25	0.2769 ± 0.0630	0.4338 ± 0.0000	0.4725 ± 0.0000	11.0314 ± 0.0000

Tabela 6. Resultados dos Experimentos de Clusterização com K-Centros 1 Distância de Manhattan

Dataset	Algoritmo	Referência	Tempo Execução (s)	Silhueta	Índice de Rand Ajustado	Raio Máximo
Dataset Sintético Distribuição Multivariada 9	K-Centros 1	0.05	0.2692 ± 0.0593	0.3236 ± 0.0000	0.4980 ± 0.0000	12.9845 ± 0.0000
Dataset Sintético Distribuição Multivariada 9	K-Centros 1	0.1	0.2715 ± 0.0616	0.3982 ± 0.0000	0.4831 ± 0.0000	12.9845 ± 0.0000
Dataset Sintético Distribuição Multivariada 9	K-Centros 1	0.15	0.2374 ± 0.0052	0.4698 ± 0.0000	0.5016 ± 0.0000	15.5639 ± 0.0000
Dataset Sintético Distribuição Multivariada 9	K-Centros 1	0.2	0.2773 ± 0.0596	0.4698 ± 0.0000	0.5016 ± 0.0000	15.5639 ± 0.0000
Dataset Sintético Distribuição Multivariada 9	K-Centros 1	0.25	0.2751 ± 0.0619	0.4698 ± 0.0000	0.5016 ± 0.0000	15.5639 ± 0.0000

Tabela 7. Resultados dos Experimentos de Clusterização com K-Centros 2 Distância de Manhattan Dados Reais

Dataset	Algoritmo	Tempo Execução (s)	Silhueta	Índice de Rand Ajustado	Raio Máximo
Blood Transfusion Service Center Data	K-Centros 2	0.0004 ± 0.0000	0.8504 ± 0.0000	0.0368 ± 0.0000	6048.0000 ± 0.0000
Yeast	K-Centros 2	0.0293 ± 0.0029	0.4233 ± 0.0000	0.0092 ± 0.0000	1.5500 ± 0.0000
Statlog Vehicle Silhouettes	K-Centros 2	0.0042 ± 0.0022	0.1330 ± 0.0000	0.0296 ± 0.0000	596.0000 ± 0.0000
Optional Recognition of Handwritten Digits	K-Centros 2	0.0757 ± 0.0216	0.0350 ± 0.0000	0.0773 ± 0.0000	347.0000 ± 0.0000
Abalone	K-Centros 2	0.4197 ± 0.1282	0.2360 ± 0.0000	0.0314 ± 0.0000	2.7530 ± 0.0000
Maternal Health Risk	K-Centros 2	0.0014 ± 0.0001	0.3886 ± 0.0000	0.0022 ± 0.0000	90.0000 ± 0.0000
Diabetic Retinopathy Debrecen	K-Centros 2	0.0006 ± 0.0000	0.5459 ± 0.0000	-0.0019 ± 0.0000	562.8169 ± 0.0000
Raising	K-Centros 2	0.0008 ± 0.0012	0.6393 ± 0.0000	0.0158 ± 0.0000	156470.4476 ± 0.0000
Image Segmentation	K-Centros 2	0.0052 ± 0.0019	0.1081 ± 0.0000	0.0818 ± 0.0000	957.7717 ± 0.0000
Mammographic Mass Data	K-Centros 2	0.0004 ± 0.0000	0.6819 ± 0.0000	0.0001 ± 0.0000	56.0000 ± 0.0000

Tabela 8. Resultados dos Experimentos de Clusterização com K-Centros 2 Distância Euclidiana

Dataset	Algoritmo	Tempo Execução (s)	Silhueta	Índice de Rand Ajustado	Raio Máximo
Optional Recognition of Handwritten Digits	K-Centros 2	0.0874 ± 0.0273	0.0657 ± 0.0000	0.1229 ± 0.0000	56.2494 ± 0.0000
Blood Transfusion Service Center Data	K-Centros 2	0.0004 ± 0.0000	0.8504 ± 0.0000	0.0368 ± 0.0000	6000.0887 ± 0.0000
Image Segmentation	K-Centros 2	0.0039 ± 0.0020	0.1024 ± 0.0000	0.0338 ± 0.0000	371.1860 ± 0.0000
Mammographic Mass Data	K-Centros 2	0.0007 ± 0.0012	0.6819 ± 0.0000	0.0001 ± 0.0000	49.2138 ± 0.0000
Maternal Health Risk	K-Centros 2	0.0016 ± 0.0001	0.2909 ± 0.0000	-0.0069 ± 0.0000	60.9672 ± 0.0000
Diabetic Retinopathy Debrecen	K-Centros 2	0.0006 ± 0.0000	0.5672 ± 0.0000	-0.0013 ± 0.0000	227.8358 ± 0.0000
Abalone	K-Centros 2	0.4543 ± 0.1470	0.2395 ± 0.0000	0.0309 ± 0.0000	1.4062 ± 0.0000
Raising	K-Centros 2	0.0007 ± 0.0005	0.6366 ± 0.0000	0.0131 ± 0.0000	114332.7300 ± 0.0000
Statlog Vehicle Silhouettes	K-Centros 2	0.0043 ± 0.0027	0.1242 ± 0.0000	0.0355 ± 0.0000	335.2357 ± 0.0000
Yeast	K-Centros 2	0.0305 ± 0.0027	0.3350 ± 0.0000	-0.0090 ± 0.0000	0.9103 ± 0.0000

Tabela 9. Resultados dos Experimentos de Clusterização com K-Means Dados Reais

Dataset	Tempo Execução (s)	Silhueta	Índice de Rand Ajustado	Raio Máximo
Raising	0.0572 ± 0.0209	0.6542 ± 0.0001	0.1665 ± 0.0005	411985.0973 ± 8.0829
Abalone	0.4528 ± 0.1275	0.2722 ± 0.0045	0.0452 ± 0.0014	2.6730 ± 0.0820
Blood Transfusion Service Center Data	0.0316 ± 0.0080	0.7025 ± 0.0000	0.0795 ± 0.0000	15223.1532 ± 0.0000
Diabetic Retinopathy Debrecen	0.0700 ± 0.0262	0.4407 ± 0.0007	-0.0016 ± 0.0000	2255.7181 ± 0.1906
Image Segmentation	0.0933 ± 0.0445	0.3326 ± 0.0055	0.4055 ± 0.0186	2768.7261 ± 19.1158
Mammographic Mass Data	0.0330 ± 0.0135	0.5427 ± 0.0000	0.1367 ± 0.0000	111.0518 ± 0.0000
Maternal Health Risk	0.0416 ± 0.0151	0.3753 ± 0.0000	0.1264 ± 0.0000	178.3239 ± 0.0000
Optional Recognition of Handwritten Digits	0.9101 ± 0.4208	0.1871 ± 0.0023	0.6719 ± 0.0025	1884.7318 ± 1.0507
Statlog Vehicle Silhouettes	0.0362 ± 0.0061	0.4459 ± 0.0133	0.1204 ± 0.0009	1943.4479 ± 5.4125
Yeast	0.1492 ± 0.0534	0.1762 ± 0.0069	0.1415 ± 0.0101	3.5747 ± 0.0435

6. Análise dos resultados

Faremos a análise dos resultados considerando cada uma das métricas obtidas a partir de nossos experimentos, no geral, desvio padrão baixo indica que os resultados obtidos são consistentes e não apresentam grandes flutuações em torno do valor obtido.

6.1. Tempo de execução

Após a execução dos experimentos foi possível observar que em média o K-Centros 2 apresenta o menor tempo de execução, o que pode ser justificado pela simplicidade do algoritmo na realização do agrupamento. O K-Centros 1 apresenta um tempo de execução maior, visto que, a cada iteração em que é realizado um refinamento, é necessário rodar uma nova clusterização e verificar se ela é valida. Já o K-Means possui um tempo de execução intermediário entre os dois algoritmos apresentados.

É válido mencionar que não foi considerado no cálculo do tempo de execução da matriz de distâncias, o que poderia elevar o tempo de execução dos algoritmos aproximativos, dando uma vantagem maior para o K-Means nesse sentido.

Também vemos que o tempo de execução tende a ser menor ao se utilizar a distância euclidiana, em comparação com a distância de Manhattan. Pode-se observar que o tempo de execução é menor com o aumento da porcentagem do refinamento no algoritmo K-Clusters 1 para alguns casos. Em geral o tempo de execução é menor nos dados sintéticos pelo menor número de dimensões em comparação com os dados reais.

6.2. Raio da solução

Em relação a métrica de raio da solução os algoritmos aproximativos apresentaram resultados melhores que o K-Means. O K-Centros 1 e 2 apresentaram desempenho similar, mas no geral pode-se observar que o K-Centros 1 tende a ser melhor, o que é esperado visto que buscamos obter um raio de x% do raio do intervalo inicial. Em alguns casos valores menores da porcentagem do parâmetro de referência apresentam resultados melhores. Além disso, também foi possível observar que ao considerar a distância euclidiana obtemos valores de raio menor que ao considerar a distância de Manhattan.

6.3. Silhueta

Em relação a métrica de silhueta pode-se observar que em alguns casos não são observadas diferenças significativas entre os algoritmos, de modo que o o K-Means e o K-Clusters 1 apresentaram o melhor desempenho, o que indica que foram gerados clusters mais separados, com menos sobreposição utilizando essa abordagem. Além disso, temos resultados melhores com a distância euclidiana, em comparação com a distância de Manhattan. Também foi possível observar que a redução da porcentagem do parâmetro de refinamento produz resultado melhores em relação à métrica em questão.

6.4. Índice de Rand ajustado

Em relação a métrica de índice de Rand ajustado pode-se observar que os algoritmos apresentam um melhor desempenho nos dados sintético em comparação aos dados reais. O que mostra que em alguns casos os algoritmos obtiveram como resultados clusters diferentes dos reais. O K-Means e o K-Clusters 1 apresentaram desempenho similar na métrica em questão.

Vale ressaltar que o K-Centros 1 apresenta um melhor desempenho nessa métrica considerando valores menores da porcentagem do parâmetro de refinamento. Ambos os algoritmos possuem resultados melhores considerando a distância Euclidiana.

7. Conclusão

O trabalho prático permitiu a aplicação dos conceitos aprendidos nas aulas relacionados a algoritmos aproximativos, especificamente o problema de agrupamento. Após a implementação dos algoritmos e a execução dos experimentos foi possível observar que apesar de mandar um tempo de execução maior devido aos refinamento realizados o K-Centros 1 tem desempenho similar ao K-Means, que é o algoritmo clássico utilizado para a tarefa de agrupamento, nas métricas consideradas nesse trabalho. Além disso o K-Centros 2 se destaca por apresentar um tempo de execução menor e ser um algoritmo mais simples em comparação com os anteriores. Também foi possível comparar o desempenho dos algoritmos em dados reais e em dados sintéticos, além da comparação da aplicação da utilização de diferentes métricas de distância e seu impacto no tempo de execução e qualidade dos algoritmos.

Referências

[1] sklearn.metrics.silhouette_score, *Scikit-learn Documentation*, 2024. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html. Acessado em: 15 de agosto de 2024.

[2] Notas técnicas do algoritmo k-means, *Amazon Web Services Documentation*, 2024. Disponível em: https://docs.aws.amazon.com/pt_br/sagemaker/latest/dg/algo-kmeans-tech-notes.html#:~:text=O%20k%2Dmeans%20%C3%A9%20um,n%C3%BAmero%20de%20atributos%20da%20observa%C3%A7%C3%A3o). Acessado em: 15 de agosto de 2024.

[3] sklearn.metrics.adjusted_rand_score, *Scikit-learn Documentation*, 2024. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html. Acessado em: 15 de agosto de 2024.