Trabalho Prático 2
Algoritmo 2-aproximado para solução do k-centros

Arthur Pontes Nader[1]

[1]Universidade Federal de Minas Gerais

arthurnader@dcc.ufmg.br

Abstract. *This report presents the results obtained in the implementation of a 2-approximate algorithm to solve the k-centers problem. These results were compared with those obtained using the KMeans method available in the Scikit-Learn library.*

Resumo. *Este relatório apresenta os resultados obtidos na implementação de um algoritmo 2-aproximado para solução do problema de k-centros. Esses resultados foram comparados com os obtidos por meio do método KMeans disponibilizado na biblioteca Scikit-Learn.*

1. Introdução

O problema de k-centros consiste em selecionar k elementos de um conjunto de pontos de tal forma que se minimize a distância máxima dos pontos ao centro que estiver mais próximo. Problemas de agrupamento, tal como o k-centros, são muito comuns em tarefas de Machine Learning e Data Mining.

Por se tratar de um problema NP-difícil, não se conhece algoritmo em tempo polinomial para resolvê-lo de forma exata. Por isso, utiliza-se um algoritmo aproximativo que fornece um resultados que é no máximo 2 vezes pior do que o ótimo.

2. Métodos e métricas

Para implementação desse algoritmo 2-aproximado, utilizou-se a linguagem Python e as seguintes bibliotecas: numpy, pandas, sklearn e time. A metodologia utilizada consistiu nas seguintes etapas: normalização dos dados, cálculo das distâncias entre cada instância do dataset e utilização do algoritmo de k-centros implementado e do KMeans para realização do agrupamento.

Utilizou-se as seguintes métricas para avaliação dos resultados obtidos:

- Tempo: refere-se ao tempo gasto para realizar a tarefa de agrupamento
- Raio máximo: distância máxima dentre todos os pontos em relação ao centro mais próximo.
- Índice de Rand ajustado: métrica usada para comparação dos valores verdadeiros de cada instância e das classes preditas pelo algoritmo.
- Silhueta: mede o quanto um objeto é similar ao seu cluster comparado com os demais clusters.

3. Implementação

Para construção do algoritmo e das métricas de avaliação, foi necessário a implementação das seguintes funções:

- Função normalizacao_dataset:

Essa função subtrai de cada atributo do dataset seu valor médio e divide a coluna pelo seu respectivo desvio padrão, retornando assim o dataset normalizado.

- Função distancia_de_minkowski:

Já essa função calcula a distância de Minkowski para dois vetores de acordo com o valor passado como parâmetro.

- Função distancias_entre_pontos:

Dado um conjunto de instâncias com atributos numéricos, essa função utiliza o cálculo da distância de Minkowski para gerar as distâncias entre cada par de instâncias do conjunto de dados.

- Função ponto_mais_distante:

Essa função recebe um conjunto de centros e a matriz de distâncias. Assim, ela realiza os cálculos para encontrar o ponto que está mais distante de um centro. Esse ponto será o próximo centro a ser adicionado à solução.

- Função k_centros:

A função k_centros escolhe um ponto aleatório para ser o primeiro centro e, após isso, faz a escolha dos k-1 centros restantes por meio do resultado da chamada da função ponto_mais_distante.

- Função raio_maximo_kmeans:

Após a execução do KMeans, tem-se os resultados dos centros escolhidos pelo algoritmo. Com esses centros, a função raio_maximo_kmeans é capaz de calcular o raio máximo obtido no agrupamento realizado.

- Função gerar_resultados_dataset:

Essa função é responsável por gerar os resultados de cada configuração de agrupamento. Esses resultados são guardados em um arquivo csv.

- Função main:

Na main é realizada a chamada da função que irá gerar os resultados para cada um dos datasets.

4. Experimentos

Os experimentos realizados consistiram em avaliar o tempo de execução, o raio máximo, o índice de Rand ajustado e a silhueta em 30 iterações de uma dada configuração para 10 datasets que deveriam ser apropriadamente escolhidos no seguinte site: https://archive.ics.uci.edu/ml/index.php. Em cada iteração, os valores obtidos são

guardados em uma lista e, após a trigésima, calcula-se a média e o desvio-padrão para cada uma dessas métricas.

As diferentes configurações se referem ao algoritmo de k-centros implementado para distâncias de Minkowski com valor de p = 1, 2, 3 e 4. Além disso, a quinta configuração utiliza o método KMeans da biblioteca Scikit-learn. O valor de k é determinado pela quantidade de classes presentes no dataset em questão.

Os conjuntos de dados utilizados nos experimentos precisaram ser adaptados e filtrados para execução do algoritmo. Assim, o seguinte link contém os datasets após o processamento inicial: https://github.com/arthurnader/datasets-tp2.

A seguir, há uma breve descrição de cada um dos dataset, bem como um link para a página em que ele foi encontrado.

- accelerometer: este conjunto de dados busca relacionar como a vibração de um motor em cada um dos 3 eixos está associado com o modo que ele está configurado. Existem 3 configurações possíveis.

link: https://archive.ics.uci.edu/ml/datasets/Accelerometer

- avila: esse é um conjunto de dados que se relaciona com métricas extraídas de um livro espanhol. Essas métricas se relacionam com 12 classes diferentes.

link: https://archive.ics.uci.edu/ml/datasets/Avila

- credit: os diversos atributos desse conjunto de dados são usados para classificar pessoas de acordo com seu risco de crédito. Há 2 classes possíveis.

link: https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

- frogs: no dataset frogs estão presentes informações sobre as frequências de emissões sonoras de anfíbios. Há 60 classes de anfíbios diferentes no conjunto de dados.

link: https://archive.ics.uci.edu/ml/datasets/Anuran+Calls+%28MFCCs%29

- images: este conjunto de dados apresenta diversos atributos numéricos de imagens, como saturação e intensidade de pixels. Esses atributos se relacionam a 7 diferentes tipos de cenas.

link: https://archive.ics.uci.edu/ml/datasets/Image+Segmentation

- plants: o dataset plants possui 64 atributos numéricos relacionados a 100 classes de plantas diferentes.

link:https://archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set

- pulsars: neste conjunto de dados, há atributos astronômicos que se relacionam a duas classes de estrelas: pulsar e não pulsar.

link: https://archive.ics.uci.edu/ml/datasets/HTRU2

- satellite: os atributos deste dataset são relacionados a imagens de um satélite. O conjunto possui 6 classes, sendo cada uma delas associadas a um tipo de ambiente.

link: https://archive.ics.uci.edu/ml/datasets/Statlog+%28Landsat+Satellite%29

- vehicles: o dataset possui como atributos as medidas de contorno para instâncias de 4 tipos de veículos diferentes.

link: https://archive.ics.uci.edu/ml/datasets/Statlog+%28Vehicle+Silhouettes%29

- waveform: por fim, neste dataset estão presentes 21 atributos, dentre os quais um é o ruído, que são usados para associação com 3 tipos diferentes de ondas.

link:https://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+%28Version+1%29

5. Resultados

As tabelas a seguir mostram os resultados obtidos para cada um dos datasets escolhidos. Após essas tabelas, há uma análise que compara os valores obtidos para cada um dos métodos utilizados.

Table 1. Resultados do dataset accelerometer.csv

k = 3

	Raio (média)	Raio (desvio-padrão	Tempo (média)	Tempo (desvio-padrão)	Índice de Rand ajustado (média)	Índice de Rand ajustado (desvio-padrão)	Silhueta (média)	Silhueta (desvio-padrão)
K-centros (p=1)	14.957310	1.277211	0.000800	0.002401	0.000537	2.399873e-04	0.704483	0.026756
K-centros (p=2)	9.897578	0.354607	0.001563	0.004688	0.000551	1.028202e-03	0.714761	0.059955
K-centros (p=3)	8.896203	0.302318	0.000521	0.002805	0.000242	1.590613e-04	0.729675	0.031237
K-centros (p=4)	8.662773	0.394428	0.000521	0.002805	0.000242	1.127906e-04	0.731649	0.026615
KMeans	9.774762	0.087108	0.148480	0.011473	0.000225	2.710505e-20	0.363676	0.000840

Table 2. Resultados do dataset avila.csv

k = 12

	Raio (média)	Raio (desvio-padrão	Tempo (média)	Tempo (desvio-padrão)	Índice de Rand ajustado (média)	Índice de Rand ajustado (desvio-padrão)	Silhueta (média)	Silhueta (desvio-padrão)
K-centros (p=1)	19.523085	0.610565	0.001563	0.004689	-0.005792	1.064455e-02	0.399623	0.121357
K-centros (p=2)	8.425290	0.180647	0.001035	0.003873	-0.004109	6.257054e-03	0.439158	0.111174
K-centros (p=3)	6.480250	0.083462	0.002097	0.004930	-0.006663	5.179093e-03	0.424454	0.075665
K-centros (p=4)	5.959604	0.145352	0.002883	0.005525	-0.000200	2.747030e-02	0.375513	0.111300
KMeans	9.636748	0.792639	0.816148	0.021945	-0.031023	6.938894e-18	0.162534	0.007166

Table 3. Resultados do dataset credit.csv

k = 2

	Raio (média)	Raio (desvio-padrão	Tempo (média)	Tempo (desvio-padrão)	Índice de Rand ajustado (média)	Índice de Rand ajustado (desvio-padrão)	Silhueta (média)	Silhueta (desvio-padrão)
K-centros (p=1)	37.387014	2.400108	0.001041	0.003895	0.008540	2.672491e-02	0.195400	0.083724
K-centros (p=2)	11.235318	0.351625	0.000000	0.000000	0.001223	1.150444e-02	0.300199	0.059265
K-centros (p=3)	8.068500	0.162483	0.000000	0.000000	0.002499	4.956768e-03	0.336823	0.010918
K-centros (p=4)	6.988683	0.162336	0.000000	0.000000	0.000841	2.751890e-03	0.339658	0.010784
KMeans	10.434586	0.060708	0.154470	0.026194	0.001352	2.168404e-19	0.114238	0.001430

Table 4. Resultados do dataset frogs.csv

k = 60

	Raio (média)	Raio (desvio-padrão	Tempo (média)	Tempo (desvio-padrão)	Índice de Rand ajustado (média)	Índice de Rand ajustado (desvio-padrão)	Silhueta (média)	Silhueta (desvio-padrão)
K-centros (p=1)	17.189053	0.239919	0.012301	0.006949	0.251647	3.761752e-02	0.199652	0.031210
K-centros (p=2)	4.758780	0.073443	0.012450	0.005891	0.246749	2.349750e-02	0.186091	0.027898
K-centros (p=3)	3.274144	0.058483	0.009149	0.007595	0.243177	3.073243e-02	0.184574	0.029519
K-centros (p=4)	2.825700	0.044700	0.011130	0.007025	0.237858	3.192430e-02	0.168404	0.025157
KMeans	5.526777	0.607996	1.914655	0.051058	0.222333	2.775558e-17	0.218918	0.011054

Table 5. Resultados do dataset images.csv

k = 7

	Raio (média)	Raio (desvio-padrão	Tempo (média)	Tempo (desvio-padrão)	Índice de Rand ajustado (média)	Índice de Rand ajustado (desvio-padrão)	Silhueta (média)	Silhueta (desvio-padrão)
K-centros (p=1)	25.987138	1.709354	0.000000	0.000000	0.131276	7.573560e-02	0.321564	0.077471
K-centros (p=2)	10.906458	0.394587	0.002084	0.005314	0.002184	8.637947e-03	0.388857	0.093290
K-centros (p=3)	9.446246	0.299830	0.001563	0.004689	0.000424	2.895527e-04	0.384384	0.071064
K-centros (p=4)	8.942345	0.240383	0.000521	0.002806	0.000604	2.965827e-04	0.342141	0.042540
KMeans	16.600793	0.000000	0.320311	0.022677	0.000938	1.084202e-19	0.336903	0.026456

Table 6. Resultados do dataset plants.csv

k = 100

	Raio (média)	Raio (desvio-padrão	Tempo (média)	Tempo (desvio-padrão)	Índice de Rand ajustado (média)	Índice de Rand ajustado (desvio-padrão)	Silhueta (média)	Silhueta (desvio-padrão)
K-centros (p=1)	34.614836	0.330698	0.047956	0.005609	0.267904	1.522267e-02	0.101494	0.006915
K-centros (p=2)	8.127014	0.058256	0.048253	0.003979	0.189257	2.804663e-02	0.106451	0.006050
K-centros (p=3)	5.733196	0.030384	0.045389	0.004712	0.137244	3.428728e-02	0.087034	0.010287
K-centros (p=4)	5.057838	0.031412	0.046459	0.007655	0.108931	3.338756e-02	0.071652	0.011898
KMeans	10.068569	0.838312	3.642825	0.558377	0.098044	2.775558e-17	0.167840	0.002613

Table 7. Resultados do dataset pulsars.csv

k = 2

	Raio (média)	Raio (desvio-padrão	Tempo (média)	Tempo (desvio-padrão)	Índice de Rand ajustado (média)	Índice de Rand ajustado (desvio-padrão)	Silhueta (média)	Silhueta (desvio-padrão)
K-centros (p=1)	22.867622	3.227118	0.001042	0.003899	0.355097	1.805691e-01	0.611974	0.164575
K-centros (p=2)	12.027406	1.234970	0.000000	0.000000	0.289743	1.972267e-01	0.629721	0.108294
K-centros (p=3)	10.607022	0.642328	0.000033	0.000179	0.113217	1.596814e-01	0.657212	0.041370
K-centros (p=4)	9.794533	0.568895	0.000000	0.000000	0.044650	1.093426e-01	0.661467	0.026431
KMeans	11.808353	0.001761	0.125461	0.013931	-0.001591	2.168404e-19	0.575186	0.000998

Table 8. Resultados do dataset satellite.csv

k = 8

	Raio (média)	Raio (desvio-padrão	Tempo (média)	Tempo (desvio-padrão)	Índice de Rand ajustado (média)	Índice de Rand ajustado (desvio-padrão)	Silhueta (média)	Silhueta (desvio-padrão)
K-centros (p=1)	44.008362	2.704188	0.000200	0.000400	0.242145	7.278493e-02	0.217841	0.060451
K-centros (p=2)	8.959655	0.391608	0.002084	0.005313	0.241960	7.056955e-02	0.249049	0.053641
K-centros (p=3)	5.697437	0.228803	0.000521	0.002806	0.230729	7.505775e-02	0.258031	0.035339
K-centros (p=4)	4.528650	0.256370	0.000521	0.002806	0.234458	8.675658e-02	0.255528	0.056513
KMeans	8.069628	0.000237	0.320977	0.018122	0.279848	5.551115e-17	0.352531	0.000224

Table 9. Resultados do dataset vehicles.csv

k = 4

	Raio (média)	Raio (desvio-padrão	Tempo (média)	Tempo (desvio-padrão)	Índice de Rand ajustado (média)	Índice de Rand ajustado (desvio-padrão)	Silhueta (média)	Silhueta (desvio-padrão)
K-centros (p=1)	24.753926	0.894134	0.000036	0.000180	0.052763	3.255060e-02	0.233326	0.074134
K-centros (p=2)	8.090925	0.572016	0.000033	0.000179	0.030094	2.022909e-02	0.309481	0.050716
K-centros (p=3)	5.503207	0.402234	0.000033	0.000180	0.024619	1.785438e-02	0.314498	0.046035
K-centros (p=4)	4.922035	0.156765	0.000035	0.000191	0.025916	1.893759e-02	0.317325	0.087161
KMeans	6.084864	1.866516	0.105516	0.008068	0.016522	3.469447e-18	0.302989	0.010835

Table 10. Resultados do dataset waveform.csv

k = 3

	Raio (média)	Raio (desvio-padrão	Tempo (média)	Tempo (desvio-padrão)	Índice de Rand ajustado (média)	Índice de Rand ajustado (desvio-padrão)	Silhueta (média)	Silhueta (desvio-padrão)
K-centros (p=1)	29.830386	1.707307	0.000331	0.001467	0.208554	7.183633e-02	0.155117	0.038407
K-centros (p=2)	7.978020	0.350682	0.001042	0.003900	0.180832	7.371402e-02	0.151095	0.035506
K-centros (p=3)	5.603142	0.275415	0.000891	0.003128	0.182450	6.694018e-02	0.138609	0.036063
K-centros (p=4)	4.916770	0.173826	0.001042	0.003900	0.140803	7.026695e-02	0.107483	0.044677
KMeans	5.780141	0.001908	0.454993	0.018561	0.173923	5.551115e-17	0.216338	0.000024

Primeiramente, é possível notar que, nas execuções do algoritmo de k-centros, o raio máximo e seu desvio-padrão tendem a cair com o aumento do parâmetro p usado no cálculo da distância de Minkowski. Também é perceptível que, na maioria dos datasets, o raio máximo produzido pelo KMeans, que por padrão utiliza a distância euclidiana para realização dos cálculos, ficou bem próximo do raio obtido para o agrupamento de k-centros com p = 2.

Outro fato relevante a ser destacado é que o tempo de agrupamento depende principalmente de k, a quantidade de centros a serem escolhidos. Dessa forma, para casos em que há poucos centros, algumas medidas resultam em valor de tempo igual a 0. Isso ocorre principalmente devido ao modo como computadores conseguem representar valores numéricos de ponto flutuante muito pequenos. Percebe-se também que em todos os datasets o KMeans necessita de um tempo relativamente maior para executar o agrupamento comparado com as demais configurações de k-centros.

Em relação a precisão dos métodos, notou-se que o método KMeans é mais preciso na maioria das métricas, pois seus desvios-padrão foram bem menores que os obtidos com as diversas configurações de k-centros, o que de certa forma parece compensar o tempo gasto para execução desse método.

A métrica índice de Rand ajustado expressa como os resultados de agrupamento obtidos se comparam com os valores reais de cada classe das instâncias do dataset. Quanto mais

perto de 1, melhor é o agrupamento feito. Assim, percebe-se que os resultados não foram muito promissores ao realizar agrupamento das instâncias dos conjuntos de dados. Esse comportamento pode ter sido causado por outliers que necessitam de centros muitos específicos que abrangem poucas instâncias ou porque provavelmente as instâncias de classes diferentes se localizam misturadas, de tal forma que não seja possível separá-las de forma adequada por meio desses métodos.

Uma outra observação relevante a ser feita é sobre os valores de silhuetas obtidos. Em nenhum dos resultados houve um valor negativo dessa métrica, o que é um indicativo de que, de modo geral, os elementos não foram atribuídos a clusters que diferem muito de suas características. Entretanto, ocorreram diversos casos de valores bem próximos de zero, o que significa que provavelmente houve sobreposição de clusters.

6. Conclusão

Os resultados obtidos permitem concluir que nos datasets escolhidos parece haver muitas semelhanças entre os atributos de elementos de classes diferentes, o que faz com que pareça ser difícil a utilização de um algoritmo de agrupamento para separação adequada das instâncias em cada um dos conjuntos de dados.

O desenvolvimento deste trabalho prático foi uma boa oportunidade de colocar em prática os conceitos e métodos sobre algoritmos aproximados vistos durante a disciplina. Além disso, a sua realização possibilitou uma maior aprendizagem acerca da experimentação de algoritmos, bem como de duas importantes métricas utilizadas para medir a qualidade de um agrupamento.

7. Referências

Kleinberg, J. and Tardos, E. (2006) “Algorithm Design”, Usa, Addison Wesley, First Edition, p. 606-612