Métodos de Agrupamento

Bruno Stafuzza Maion[1], **Rodrigo da Rosa**[1]

[1] Departamento de Ciência da Computação
Universidade Estadual do Oeste do Paraná. (Unioeste)
Cascavel – PR – Brasil

1. Introdução

Neste estudo, será abordado uma análise de três métodos de agrupamento aplicados a uma base de dados gerada artificialmente. O objetivo principal é avaliar o desempenho desses métodos com base em diversas métricas para medir a eficácia do agrupamento.

Os métodos de agrupamento são o K-Means, o DBScan e o AGNES (Aglomerative). Cada um configurado com seus próprios parâmetros específicos.

Para avaliar os agrupamentos, será utilziado as métricas que incluem a soma dos quadrados, coesão, índice de silhueta, índice randômico, homogeneidade e completude.

2. Base de dados

A base de dados utilizada foi "Base 5", que consiste em 2.000 instâncias, geradas artificalmente contendo eixo X, eixo Y e sua classe. A figura 1 representa graficamente a distribuição dos pontos, onde os pontos verdes representam a classe 1 e os pontos azuis a classe 2. Os centróides de cada classe está representado pelo ponto vermelho.

- **Mínimo (x):** -13.405
- **Máximo (x):** 3.254
- **Mínimo (y):** -0.995
- **Máximo (y):** 10.162
- **Desvio Padrão Classe 1** em (x): 1.585 (y): 1.574
- **Desvio Padrão Classe 2** em (x): 1.583 (y): 1.590
- **Ponto Médio Classe 1** em (x): -8.471 (y): 5.476
- **Ponto Médio Classe 2** em (x): -1.252 (y): 4.495

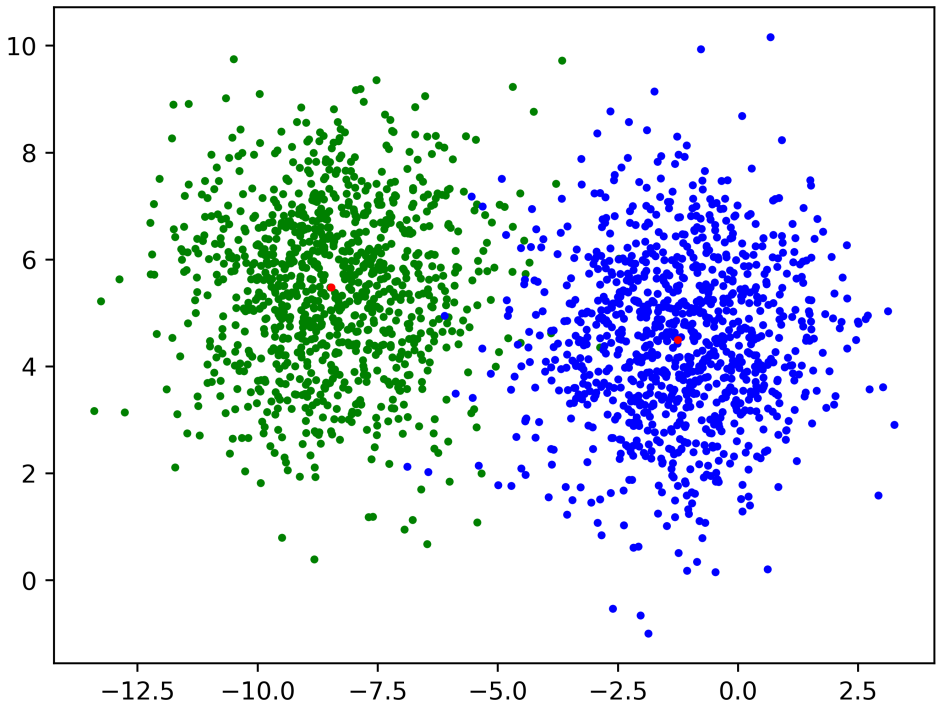

Figura 1. Plotagem dos clusters originais em X e Y.

3. Métodos de Agrupamento e Parâmetros

Este trabalho utilizou os métodos de agrupamento K-Means, DBScan e AGNES (Aglomerative). Suas características e seus parâmetros estão detalhados abaixo.

3.1. K-Means

No K-Means, o algoritmo divide um conjunto de dados em clusters, onde cada ponto de dados pertence ao cluster com o centróide mais próximo. Os parâmetros utilizados foram:

- **Número de clusters:** agrupamentos máximos possíveis. [2, 3, 4]
- **Iterações máximas:** número máximo de iterações. [1, 10, 50, 100]

3.2. DBScan

O DBScan é um algoritmo de clusterização baseado em densidade, sendo capaz de identificar clusters de diferentes formas e densidades, portanto, lida facilmente com outliers e ruído. Os parâmetros utilizados foram:

- **Épsilon:** distância máxima para que seja considerado vizinho. [0.5, 0.7, 1.0].
- **Amostras mínimas:** número de amostras mínimas para constituir um cluster. [10, 25, 50].

3.3. AGNES (Aglomerative)

O Agglomerative Clustering é um algoritmo de agrupamento hierárquico que começa com cada ponto de dados como seu próprio cluster, em seguida, une os clusters mais próximos em clusters maiores. O resultado pode ser visualizada como uma árvore. Os parâmetros utilizados foram:

- **Número de clusters:** agrupamentos máximos possíveis. [2, 3, 4]
- **Ligação:** ward - menor variância; average - distância média; complete - distância máxima; single - distância mínima.

4. Resultados

A tabela 1 sintetiza os resultados das métricas sobre o algoritmo K-Means, observa-se que os melhores resultados foram obtidos nas primeiras 4 iterações por se tratar de apenas 2 clusters alvos. Destaca-se o índice rândomico com 0.98, a homogeniedade dos cluster com 0.90 e maior completude obtida de 0.90. Nota-se também a rápida convergência com apenas 1 iteração já alcançou o máximo global, podendo ser visualizado na figura 2.

Tabela 1. K-Means

Iter	Clusters	MaxIter	SomaQua	Coesão	Silhueta	Rand	Homoge.	Completude
1	2	1	9777.38	49.44	0.62	0.98	0.90	0.90
2	2	10	9777.38	49.44	0.62	0.98	0.90	0.90
3	2	50	9777.38	49.44	0.62	0.98	0.90	0.90
4	2	100	9777.38	49.44	0.62	0.98	0.90	0.90
5	3	1	8062.64	29.93	0.42	0.85	0.87	0.58
6	3	10	8062.64	29.93	0.42	0.85	0.87	0.58
7	3	50	8062.64	29.93	0.42	0.85	0.87	0.58
8	3	100	8062.64	29.93	0.42	0.85	0.87	0.58
9	4	1	6499.63	20.16	0.31	0.74	0.89	0.45
10	4	10	6499.63	20.16	0.31	0.74	0.89	0.45
11	4	50	6499.63	20.16	0.31	0.74	0.89	0.45
12	4	100	6499.63	20.16	0.31	0.74	0.89	0.45

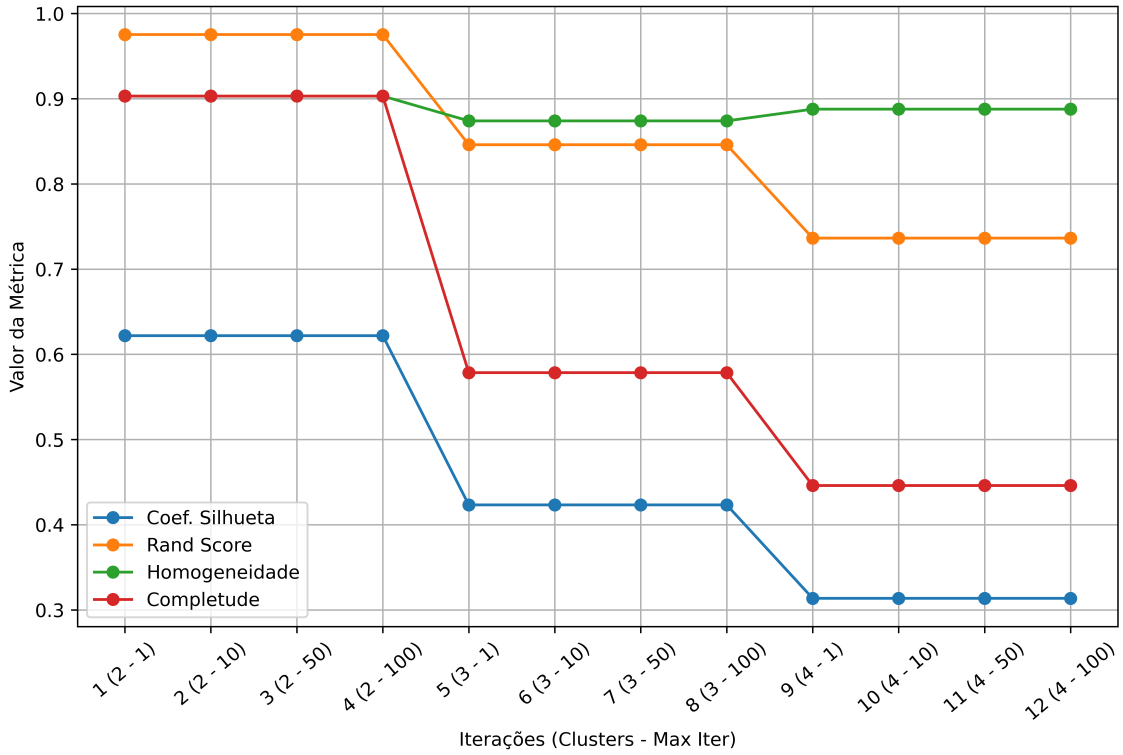

Figura 2. Gráfico das métricas por épocas.

A figura 3, abaixo, representa a plotagem dos pontos obtidos na primeira iteração do algoritmo.

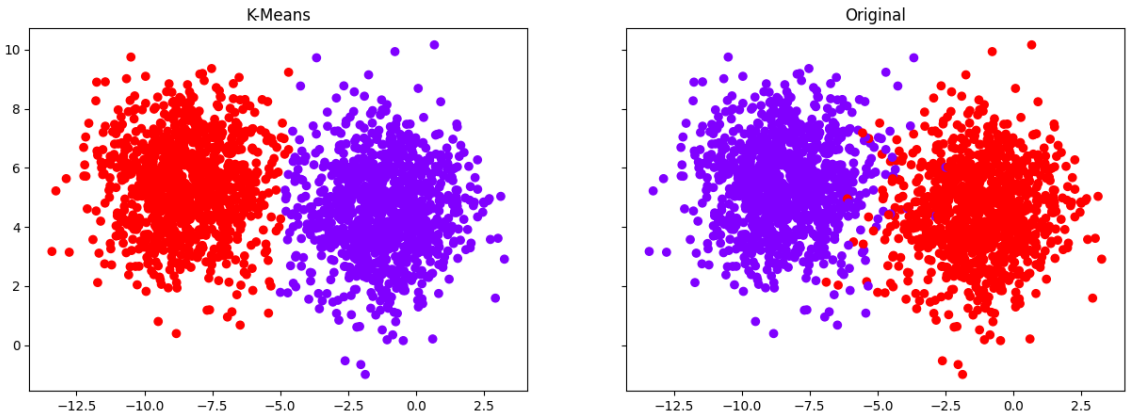

Figura 3. Gráfico da primeira iteração.

Já no DBScan, os melhores resultados foram obtidos na última iteração, onde o épsilon é igual a 1 e o número de amostras mínimas é de 50. Nota-se na tabela 2 e na figura 2.

Tabela 2. DBScan

Iter	Eps	Min Samples	Coef. Silhueta	Rand Score	Homogeneidade	Completude
1	0.50	10	0.23	0.89	0.85	0.58
2	0.50	25	0.34	0.76	0.70	0.45
3	0.50	50	-0.39	0.52	0.22	0.19
4	0.70	10	0.21	0.50	0.00	0.00
5	0.70	25	0.49	0.90	0.86	0.63
6	0.70	50	0.36	0.78	0.73	0.47
7	1.00	10	0.29	0.50	0.00	0.00
8	1.00	25	0.23	0.50	0.00	0.00
9	1.00	50	0.53	0.93	0.88	0.70

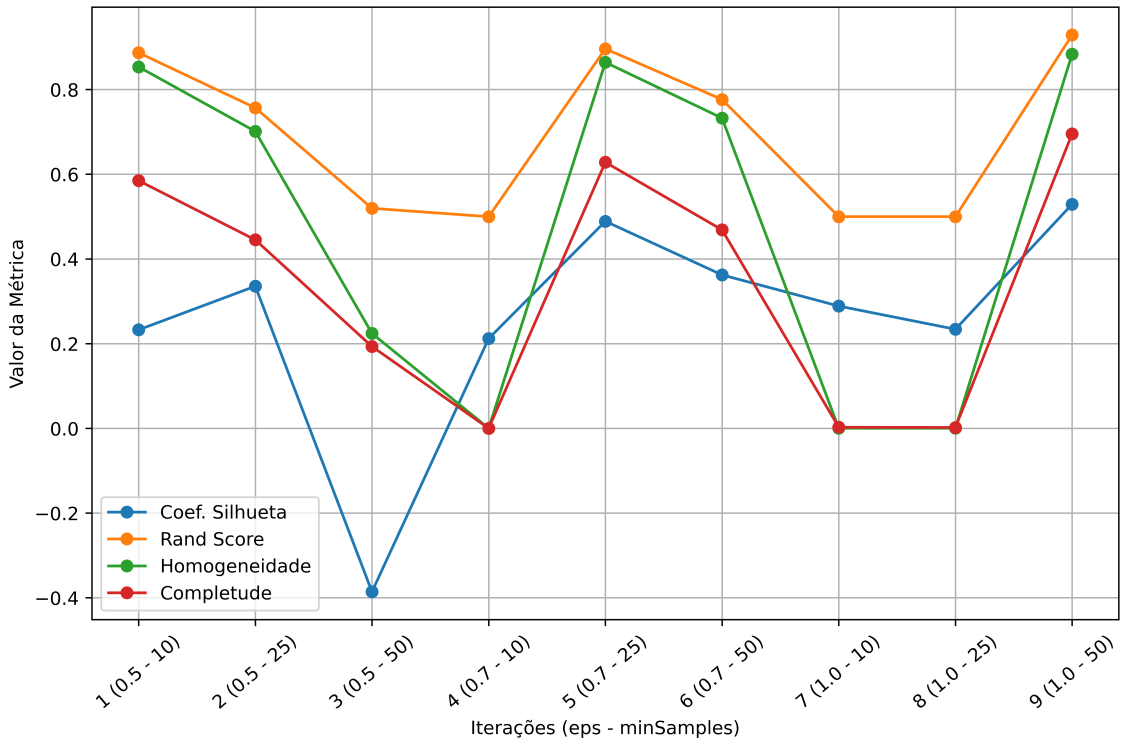

Figura 4. Gráfico das métricas por épocas.

A figura 5 representa a comparação da plotagem do algoritmo DBScan com o alvo, observa-se uma boa aproximação onde as cores verde vermelho seria os clusters identificados e a cor roxa é tratada como ruído.

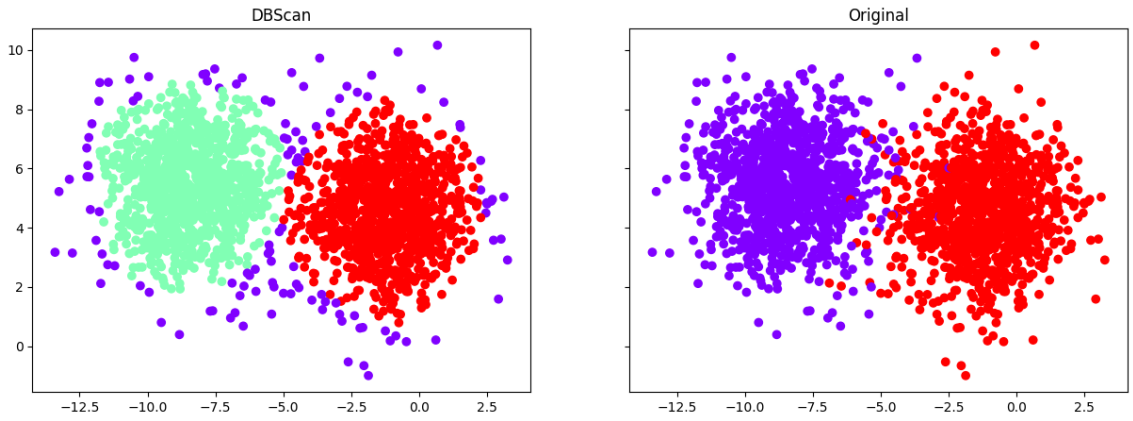

Figura 5. Gráfico da última iteração.

Observa-se na tabela 2 que algoritmo AGNES obteve os melhores resultados na segunda iteração, com a utilização dos seguintes parâmetros: número de cluster em 2 e a ligação completa, que consiste na mesclagem dos cluster a partir da distância máxima.

Tabela 3. AGNES

Iter	Clusters	Linkage	Coef. Silhueta	Rand Score	Homogeneidade	Completude
1	2	ward	0.61	0.95	0.84	0.84
2	2	complete	0.61	0.96	0.86	0.86
3	2	average	0.61	0.94	0.83	0.83
4	2	single	0.31	0.50	0.00	0.08
5	3	ward	0.44	0.85	0.84	0.56
6	3	complete	0.41	0.87	0.88	0.60
7	3	average	0.53	0.94	0.83	0.82
8	3	single	0.20	0.50	0.00	0.08
9	4	ward	0.29	0.74	0.84	0.43
10	4	complete	0.27	0.78	0.89	0.47
11	4	average	0.44	0.94	0.83	0.81
12	4	single	-0.03	0.50	0.00	0.09

A figura 6 representa as iterações, observa-se a grande similaridade nas 3 primeiras iterações e uma queda bricas das métricas na quarta.

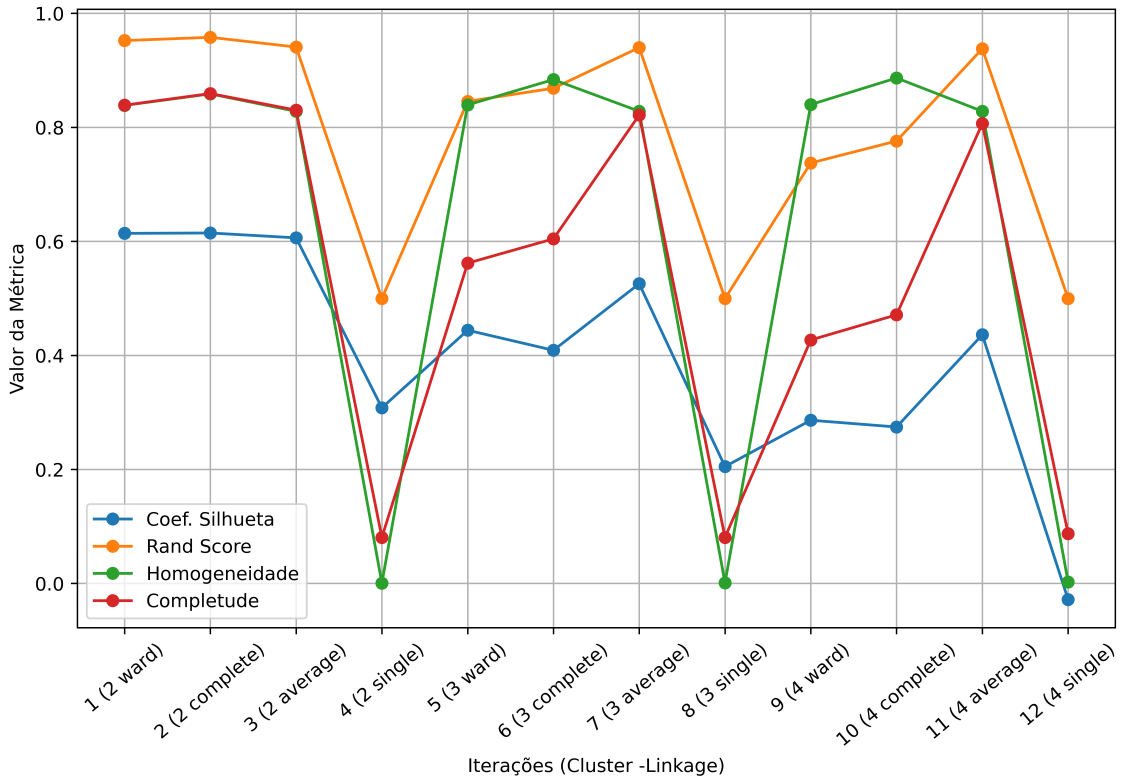

Figura 6. Gráfico das métricas por épocas.

A figura 7 representa a comparação do algortimo na segunda iteração com o alvo. Nota-se que na região central há muitos erros, principalmente por conta dos ruídos.

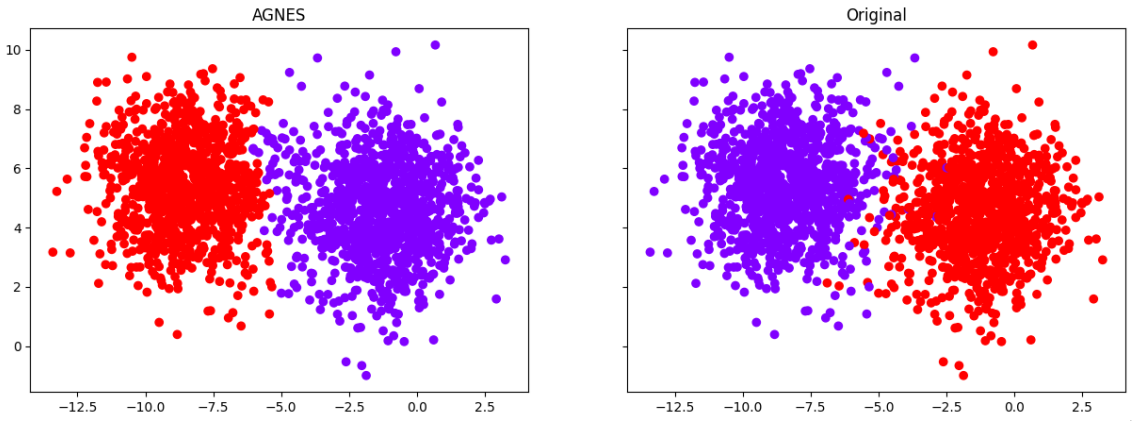

Figura 7. Gráfico da segunda iteração.

5. Conclusões

Neste estudo, examinamos a base de dados artificial “Base 5”e aplicamos três métodos de agrupamento: K-Means, DBScan e AGNES. Cada algoritmo foi ajustado com diferentes parâmetros, e seus resultados foram analisados usando diferentes métricas. Após

as análises, observou-se que o algortimo com o melhor desempenho, considerando principalmente a homegenidade e o índice rândomico, foi o K-Means, atingindo 90% e 98% respectivamente.