Assignment 4 - Clustering - Roteiro da resolução

Caroline Pereira de Sena[1]

[1]Universidade Tecnológica Federal do Paraná

carolinesena@alunos.utfpr.edu.br

1. Respostas

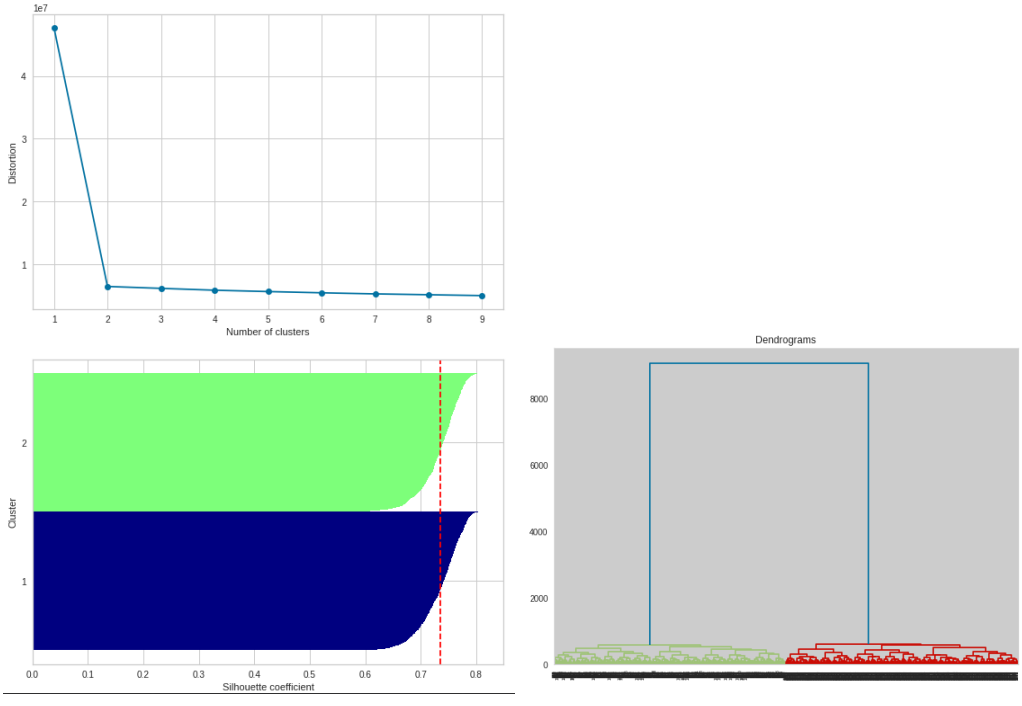
Os números de *clusters* para cada um dos *datasets* foi obtido através da comparação entre os resultados das análises de cada método: *elbow*, coeficiente de silhueta, e dendograma. Após isso, foram comparados os resultados das abordagens de *clustering*: k-means, hierárquico e dbscan. Foram obtidos os gráficos para cada caso, seguindo o código disponibilizado pelo professor e algumas bibliotecas do *python*.

Para o método "elbow", observa-se na curva o ponto em que há uma dobra. Tal ponto indica o número de clusters. Para o método de silhueta, observa-se se o gráfico tem a forma esperada e se cada curva tinge valores próximos de 1, para cada cluster. No dendograma, observa-se a área em que as ramificações tem maior altura, e então se conta o número de linhas dessa área.

1.1. *Dataset* 1

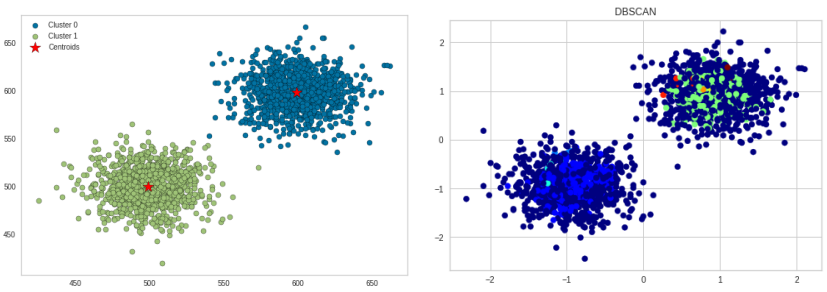
Esse conjunto de dados é composto por 8 colunas e 2048 entradas. Analisando seus gráficos, todos indicaram a divisão em 2 clusters.

Observa-se para os três gráficos da figura 1 que o número de clusters indicado é 2.

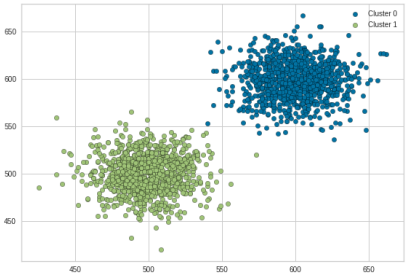


(a) (b)

Figura 1. Resultados para o Dataset 1



(a) (b)

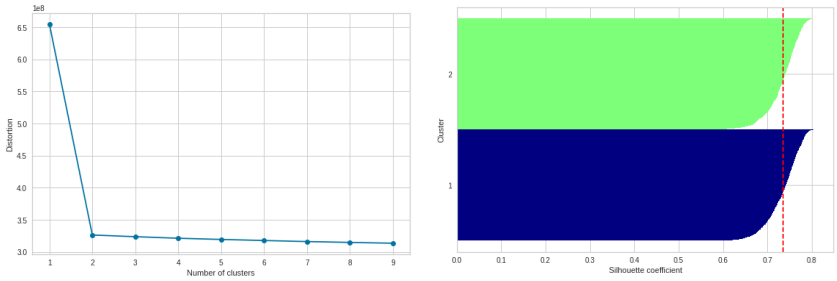


(c)

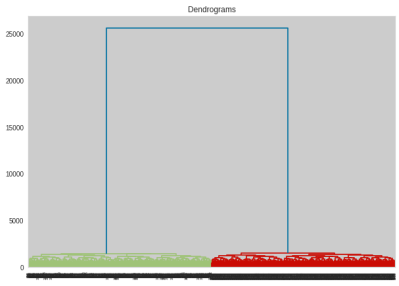
Figura 2. Resultados de (a) Treino (b) Teste

1.2. *Dataset* 2

O segundo conjunto de dados é composto por 64 colunas e 2048 entradas. Analisando seus gráficos, todos indicaram a divisão em 2 clusters.



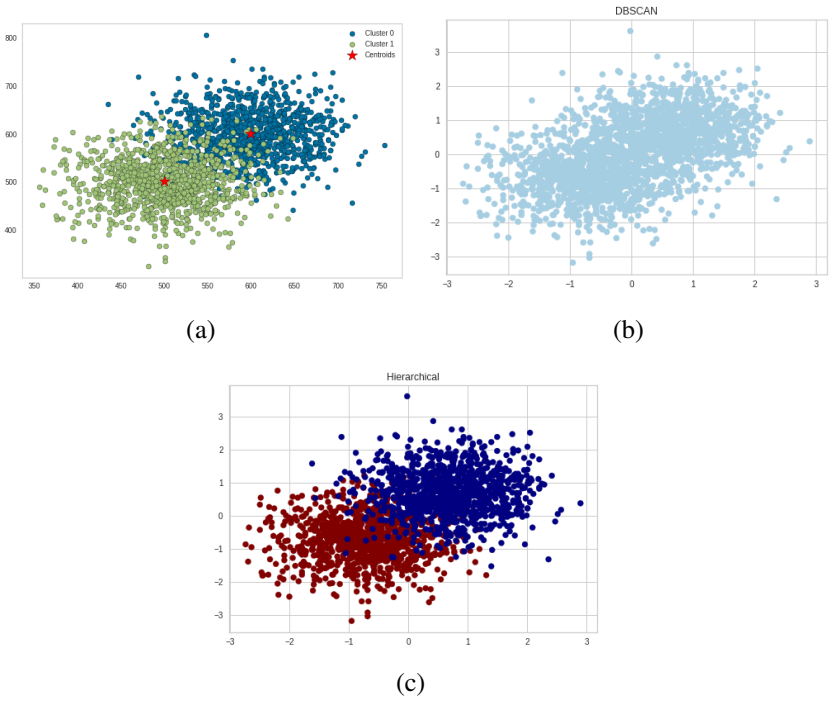
(a) (b)



(c)

Figura 3. Resultados de (a) Treino (b) Teste

Olhando os gráficos gerados na figura 4, é possível ver que o K-means e a clusterização hierarquica mostraram dois clusters, porém o DBScan parece ter encontrado apenas 1 (ou pode ser um erro de configuração da biblioteca utilizada). Porém, considerando que todos os outros métodos indicaram 2 clusters, essa é minha resposta final.



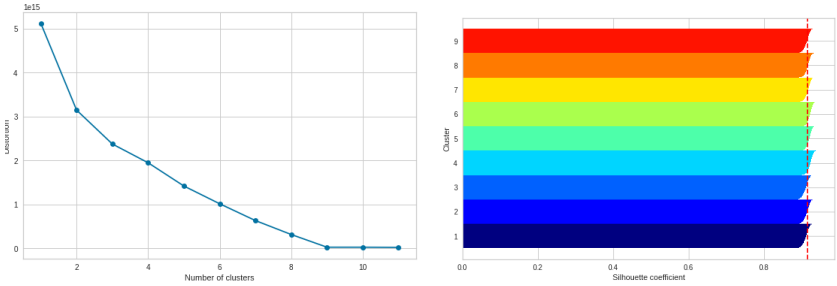
(a) (b)

(c)

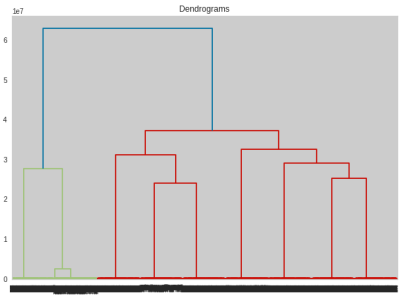
Figura 4. Resultados de (a) Treino (b) Teste

1.3. *Dataset* 3

O terceiro conjunto de dados é composto por 15 colunas e 10126 entradas. Os gráficos obtidos apontaram para uma divisão em 9 clusters.

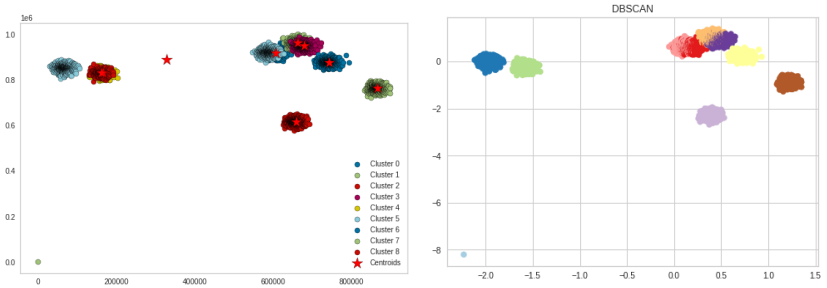


(a) (b)

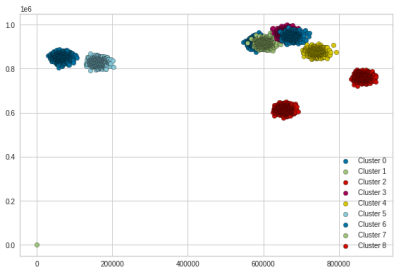


(c)

Figura 5. Resultados de (a) Treino (b) Teste



(a) (b)

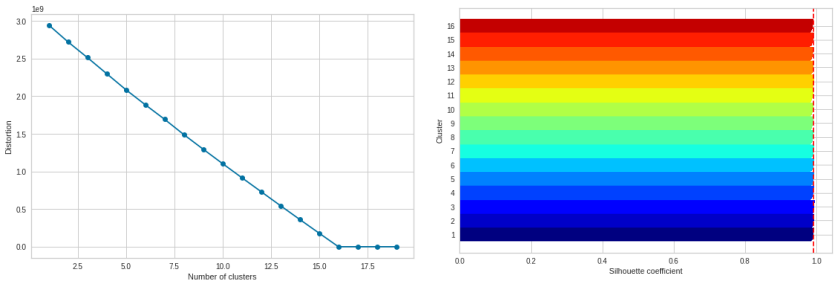


(c)

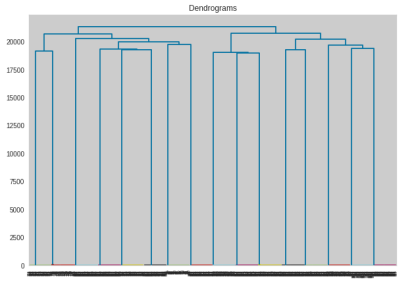
Figura 6. Resultados de (a) Treino (b) Teste

1.4. *Dataset* 4

O terceiro conjunto de dados é composto por 1024 colunas e 1024 entradas. Os gráficos obtidos apontaram para uma divisão em 16 clusters.

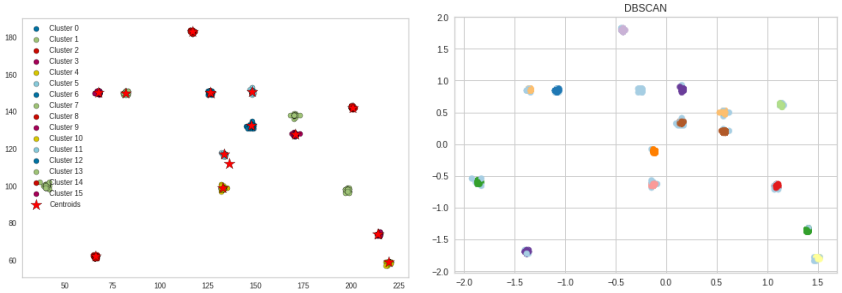


(a) (b)

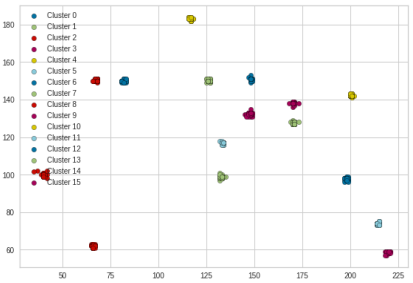


(c)

Figura 7. Resultados de (a) Treino (b) Teste



(a) (b)



(c)

Figura 8. Resultados de (a) Treino (b) Teste