Análise de qualidade de *Clusters*

Algumas análises feitas para saber se uma busca binária seria viável para achar um cluster de boa qualidade, porém, com base nas análises, a busca não seria possível.

Existe uma medida proposta pelo estatístico Peter J. Rousseeuw chamada silhueta ou largura de silhueta que mede o quão bom um sample está dentro de seu cluster sendo necessário o uso de uma métrica, seja ela distância Eucliana ou City-Block. Para cada sample **i** do dataset, a silhueta é definida por $s_i = \frac{b_i - a_i}{max\{a_i, b_i\}}$ (I), onde a_i e b_i são , respectivamente, a dissimilariade média(distância média) de **i** em relação aos samples do seu próprio cluster e a dissimilaridade média de **i** em relação aos samples que não estão no mesmo cluster de **i**. Analisando os casos limites da equação (I):

Se b_i >> a_i , teremos $s_i \rightarrow 1$, o que significa que a dissimilaridade de **i** no seu próprio cluster é desprezível, ou seja, os dados foram apropriadamente clusterizados.

Se a_i >> b_i , teremos $s_i \rightarrow -1$, o que significa que a dissimilaridade de **i** no seu cluster é maior com relação ao cluster vizinho mais próximo, ou seja, seria melhor que o sample **i** estivesse no cluster vizinho.

Se $a_i \rightarrow b_i$ ou vice-versa, teremos $s_i \rightarrow 0$, significa que o sample **i** está entre dois clusters naturais.

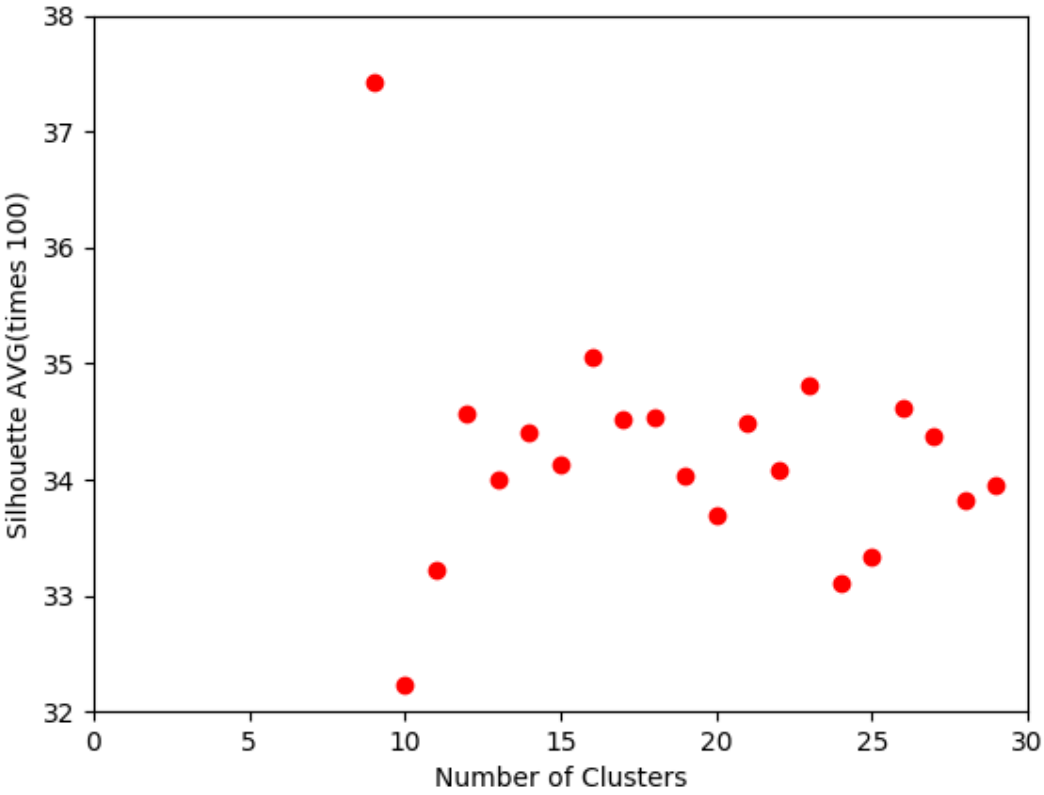
Visto isso, foi calculado o coeficiente de silhueta(SWC) que é a media aritmética de todas as silhuetas calculadas no dataset.

O coefieciente de silhueta (SWC) varia também no intervalo [-1, 1], já que é a média aritmética das silhuetas. O quão mais próximo SWC for de 1 for, melhor foi a seleção de clusters . Valores negativos de SWC implicam em uma seleção ruim, onde vários samples foram mal escolhidos pelo algoritmo ultilizado.

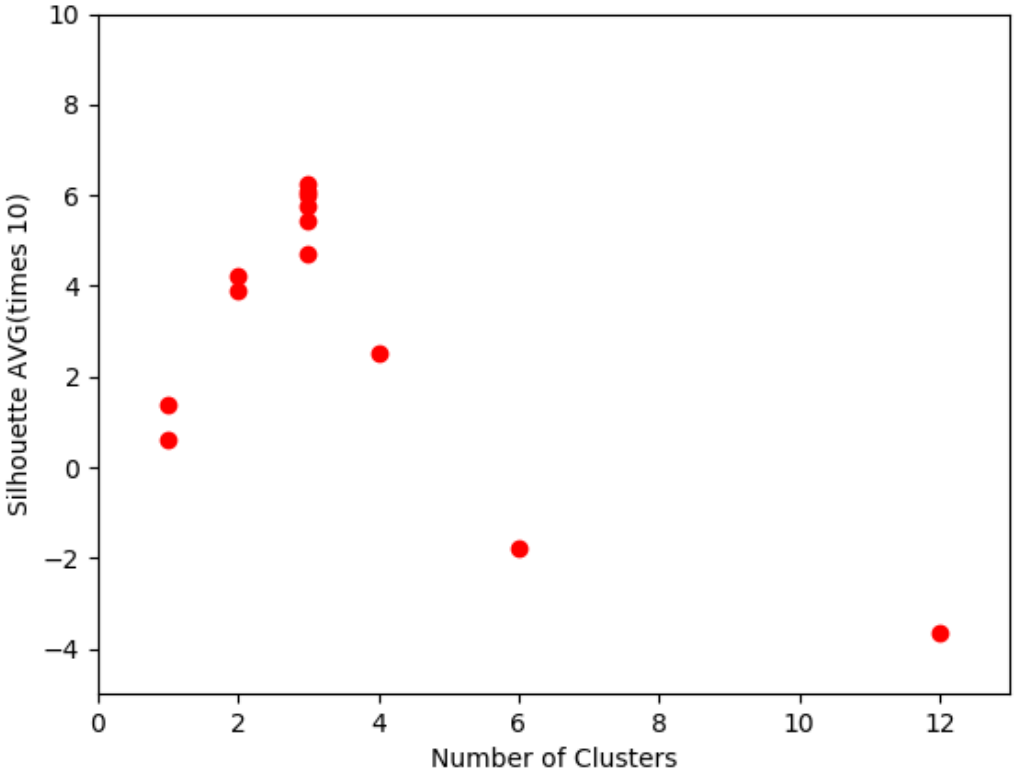
Teste empíricos

Para saber se uma busca binária seria possível, foram calculados os coeficientes de silheta para diferentes quantidades de cluster(K-Means) e para diferentes tipos de épsilon(DBSCAN) e, com base nisso, saber se as função é monótona. Os seguintes gráficos foram obtidos:

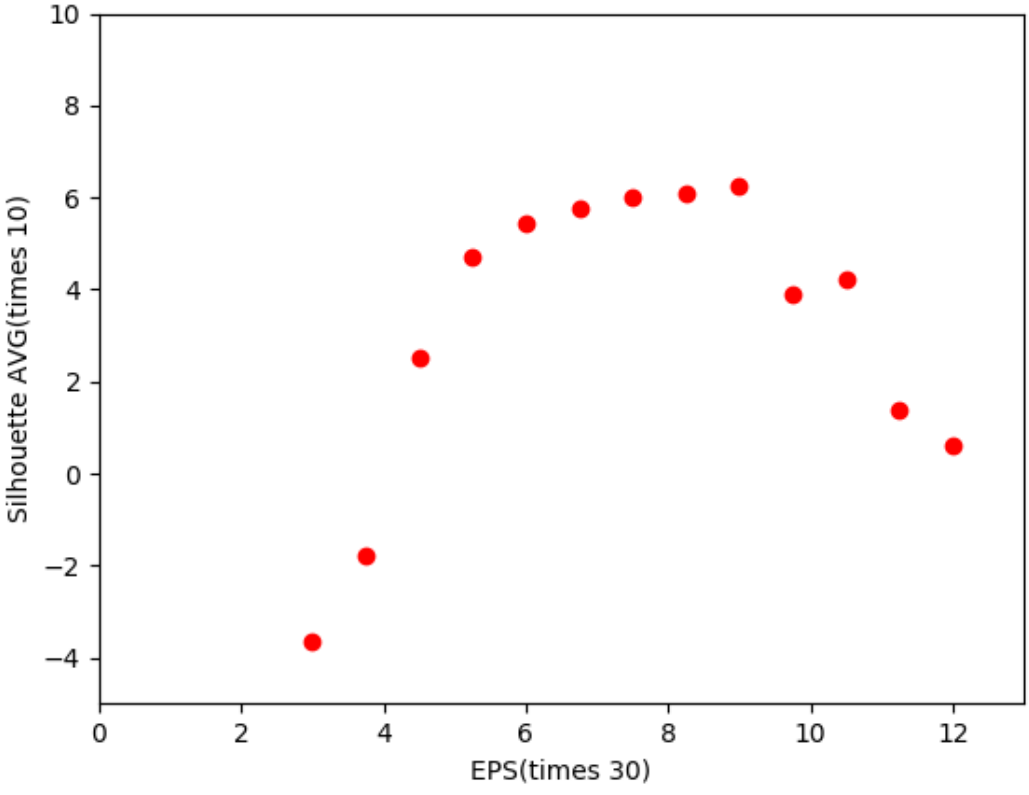
K-Means++



DBSCAN(Clusters)



DBSCAN(EPS)



Note que em nenhum dos casos a função swc(K) ou swc(EPS) é monótona. Uma busca binária precisaria que swc(X) fosse estritamente crescente ou decrescente, o que não é verdade.

Todos os cálculos feitos seguem anexados ao email

Humberto Pires Marques

fonte
http://svn.donarmstrong.com/don/trunk/projects/research/papers_to_read/statistics/silhouettes_a_graphical_aid_to_the_interpretation_and_validation_of_cluster_analysis_rousseeuw_j_comp_app_math_20_53_1987.pdf