Universidade de Brasília - UnB
Programa de Pós-Graduação em Computação Aplicada - PPCA

A tutorial on various clustering evaluation metrics

Autor: João Vitor Rodrigues Baptista
Professor: Dr. Marcelo Ladeira

Agenda

- Introdução
- Analise

Artigo

Published on March 23, 2022 In Mystery Vault

A tutorial on various clustering evaluation metrics

In this article, we will be learning about different performance metrics for clustering and implementation of them

By Sourabh Mehta

[CITAÇÃO] A tutorial on various clustering evaluation metrics
S Mehta - Analytics India Magazine, 2022
Salvar Citar Citado por 2 Artigos relacionados

Figure: *Artigo na integra pelo Analytics India Magazine. Autoria própria*

Publicação

Analytics India Magazine

- Empresa de mídia tecnológica que publica notícias. Foi fundada em 2012 e está sediada em Bengaluru, na Índia.
- Publica artigos, tutoriais e entrevistas com especialistas na área de ciência de dados e aprendizado de máquina. Também organiza conferências e workshops para ajudar os profissionais a se manterem atualizados sobre as últimas tendências (i.e Data Science Central, Eureka, Towards Data Science e etc)

Data de publicação

- 23 de Março, 2022

Autor

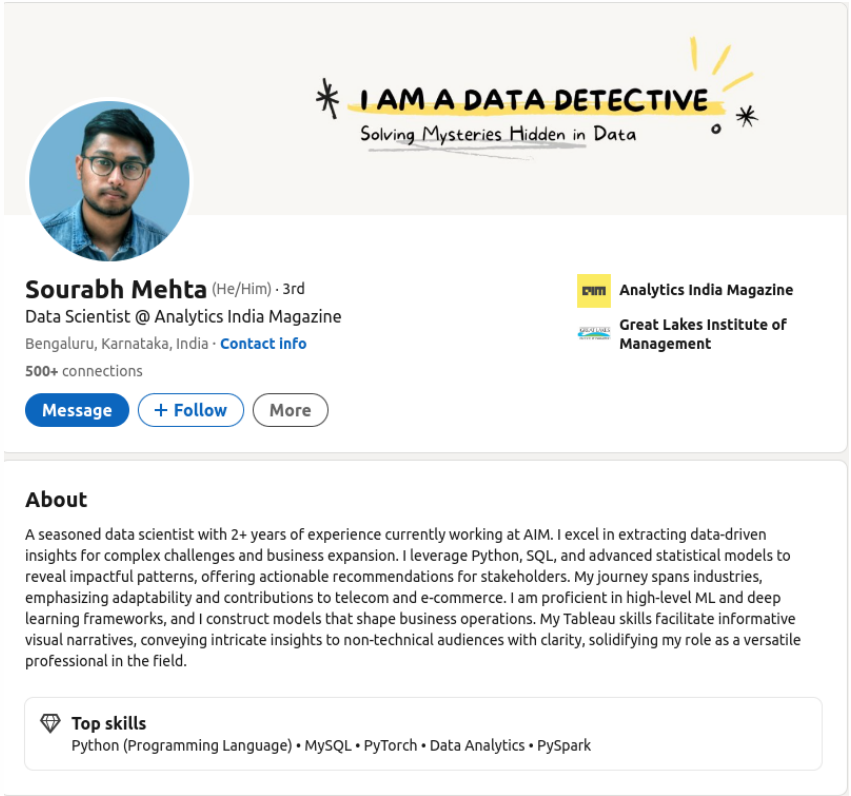


Figure: *Perfil do autor no Linkedin.*
Autoria própria

Problematica

O problema de avaliação do desempenho de algoritmos de clustering. Clustering é uma tarefa de aprendizagem não supervisionada que envolve agrupar pontos de dados em clusters com base em sua similaridade. No entanto, pode ser difícil determinar o desempenho de um algoritmo de agrupamento sem conhecer os rótulos verdadeiros dos pontos de dados.

Abordagem utilizada

Within-Cluster Sum of Square

WCSS é a soma da distância quadrada entre cada ponto e o centroide em um cluster. Quando traçamos o WCSS com o valor K, o gráfico se parece com um cotovelo. À medida que o número de clusters aumenta, o valor WCSS começará a diminuir.

Abordagem utilizada

Silhouette Score

Mede quão bem cada ponto de dados é atribuído ao seu cluster. Uma Silhouette score alta indica que um ponto de dados corresponde bem ao seu cluster, enquanto uma pontuação de silhueta baixa indica que o ponto de dados está atribuído incorretamente.

$$s(i) = \frac{N_c(i) - I_c(i)}{\max(I_c(i), N_c(i))}$$

I_c - mean of the intra-cluster distance
N_c - mean of the nearest-cluster distance

Abordagem utilizada

Calinski-Harabasz index

Mede a compactação dos clusters e a separação entre os clusters. Um alto índice de Calinski-Harabasz indica que os clusters são bem separados e compactos.

$$CH(k) = [B(k).W(k)][(n-k)(k-1)]$$

n - data points
k - clusters
$W(k)$ - within cluster variation
$B(k)$ - between cluster variation

Abordagem utilizada

Davies Bouldin index

Mede a similaridade média entre clusters. Um índice de Davies-Bouldin baixo indica que os clusters estão bem separados.

$$P = \frac{1}{n} \sum_{i=1}^{n} \max_{j} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \tag{1}$$

P - peformance index
c - centroid (geometric center)
σ - avarage disntace of cluster elements to the centroid
n - number of clusters
i - index over elements
j - index over elements

Experimentos

Dataset utilizado

Conjunto de dados contém informações sobre os clientes como **sexo, estado civil, idade, escolaridade, renda, ocupação e etc**, para segmentação de clientes. Retirado de um repositório do kaggle.

Implementação

1. Foi criado do algorítimo usando K-Means Clustering utilizando a lib Pandas.
2. Foi aplicado as três métricas mencionadas utilizando a lib Sklearn.
3. Foi utilizando de valores de 2 até 6 para os números de clusters durante os experimentos.

Resultados obtidos

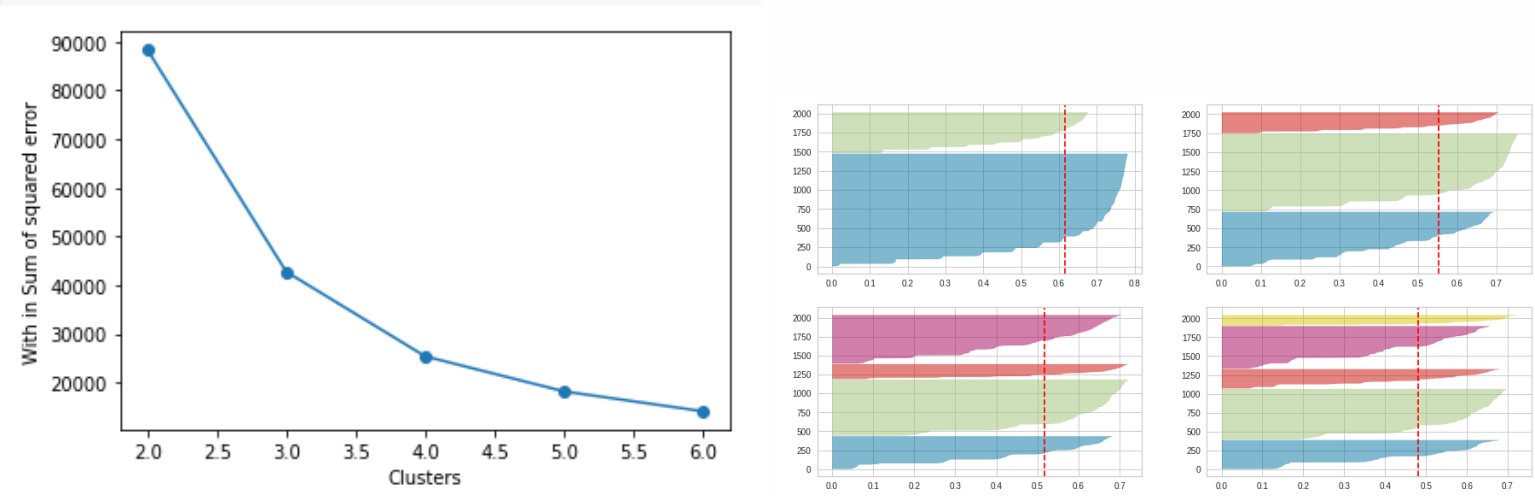
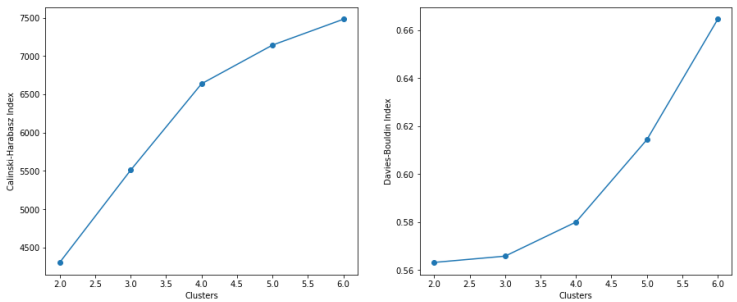


Figure: *Plote dos experimentos utilizando a metrica WCSS e a Silhouette Score*

Resultados obtidos




```
With in sum of squared error (WCSS) at K = 2 : 88313.58766373781
Silhouette Coefficient: 0.616
Calinski-Harabasz Index: 4304.782
Davies-Bouldin Index: 0.563
---------------------------------------------------------------
With in sum of squared error (WCSS) at K = 3 : 42744.08758922211
Silhouette Coefficient: 0.552
Calinski-Harabasz Index: 5509.332
Davies-Bouldin Index: 0.566
---------------------------------------------------------------
With in sum of squared error (WCSS) at K = 4 : 25381.401102352607
Silhouette Coefficient: 0.517
Calinski-Harabasz Index: 6637.444
Davies-Bouldin Index: 0.580
---------------------------------------------------------------
With in sum of squared error (WCSS) at K = 5 : 18185.11273492832
Silhouette Coefficient: 0.481
Calinski-Harabasz Index: 7141.917
Davies-Bouldin Index: 0.614
---------------------------------------------------------------
With in sum of squared error (WCSS) at K = 6 : 14102.514771759765
Silhouette Coefficient: 0.443
Calinski-Harabasz Index: 7479.326
Davies-Bouldin Index: 0.665
---------------------------------------------------------------
```

Figure: *Plote dos experimentos utilizando a métrica Calinksi e Davies, na direita é apresentado o resultado resumido de cada experimento.*

Conclusões apresentadas

Resultado dos experimentos

O autor menciona que pela métrica WCSS e Silhouette Coefficient o valor ótimo de numero de clusters é 3 e pelas métricas Calinski-Harabasz Index e Davies-Bouldin Index os melhores resultados estariam com o 4 clusters. Porem o autor segue com a implementação de 3 clusters.

Conclusão do autor

Os autores concluem que as quatro métricas propostas são uma ferramenta valiosa para avaliação de clustering. As métricas são fáceis de entender e implementar e demonstraram ser eficazes em diversos cenários de cluster.

Pontos fortes

1. O autor consegue apresentar de forma clara a problemática.
2. Aplica técnicas conhecidas para tentar solucionar o problema.
3. Promove uma aplicação pratica através dos experimentos.

Pontos fracos

1. *'various'* da entender que serão muitas técnicas de avaliação que seriam abordadas.
2. Estrutura do texto poderia ter a seção de discussão.
3. Não explica a escolha do numero de clusters.
4. Não compara as diferenças entre as técnicas.
5. Não exemplifica quais técnicas utilizar em determinas situações.

Obrigado