ISSN 2177-9139

REVISITANDO CONJUNTOS E DISTÂNCIAS PARA ENCONTRAR PONTOS VIZINHOS

Eduardo Braun - eduardotbraun@gmail.com
Universidade Federal de Santa Maria, Campus Camobi, 97105-900 - Santa Maria, RS, Brasil
Alice de Jesus Kozakevicius - alicek@ufsm.br
Universidade Federal de Santa Maria, Campus Camobi, 97105-900 - Santa Maria, RS, Brasil

Resumo. *Neste trabalho os conceitos de conjunto e distância são revisitados para explicitar o algoritmo K-means, que obtém de forma iterativa, K subconjuntos a partir de um conjunto inicial de pontos e das distâncias relativas entre seus elementos. No final do processo, dentro de cada um dos K subconjuntos, seus elementos têm distância mínima em relação a um ponto central do subconjunto, dito centróide, que também é atualizado por este processo iterativo. Em aplicações de mineração de dados (data-mining) em conjuntos com milhões de elementos, estes subconjuntos são denominados de clusters e todos os vizinhos compartilham propriedades em comum. Neste trabalho, são apresentadas diferentes métricas e suas influências na formação de clusters em conjuntos de pontos do plano.*
Palavras Chave: *Conjuntos, Subconjuntos, Clusters, Distâncias, Algoritmo K-means*

1 INTRODUÇÃO

Em matemática, um dos conceitos mais intuitivos é o de conjunto como sendo uma coleção de elementos. Foi através de Georg Cantor(1845-1918) que a teoria dos conjuntos começou a ser construída, e a noção de que um conjunto pode ser também definido a partir das propriedades que seus elementos compartilham começou a ser melhor entendida e explorada [JOHNSON,1972]. De fato, ao longo de toda nossa formação, a definição de conjunto ou aplicações envolvendo conjuntos nos acompanham de forma intuitiva, como por exemplo quando uma criança associa elementos a números para poder contar. Depois, esse conceito é estendido quando somos convidados a reconhecer padrões e classificar elementos que preservem estes padrões. Por exemplo, quem não se lembra das aulas de biologia quando eram classificados animais e plantas segundo suas características morfológicas?

Outro conceito intuitivo é o de distância, com o qual temos contato em diferentes aplicações do dia a dia. Quando pensamos em distância, imediatamente estamos pensando em distância Euclidiana. Na verdade, esta não é a única função possível a ser calculada para designar a distância entre dois pontos, ou dois objetos, ou entre dados quaisquer [LIMA,1977]. Neste trabalho, o foco principal é explorar diferentes métricas, funções distância, na determinação de vizinhanças. Vizinhanças são conjuntos nos quais seus elementos estão todos próximos em relação ao um ponto central, denominado centróide. Estes conjuntos podem ser diferentes dependendo da métrica escolhida e são exatamente essas diferenças que queremos estudar.

Esta ideia de vizinhança está por traz do conceito de cluster, muito empregado em problemas de mineração de dados, ou reconhecimento de padrões. Clusters são conjuntos considerados para seleção de dados conforme algum critério de similaridade [ESTIVILL-CASTRO, 2002]. Assim, os elementos em um mesmo cluster são todos mais similares uns aos outros, do que em relação a elementos de outros clusters. Um dos algoritmos bem estabelecidos para a obtenção de clusters é o algoritmo K-means [MACQUEEN,1967], cuja ideia é formar K subconjuntos

considerando como critério de similaridade a menor distância possível em relação ao centróide do cluster. O valor K é escolhido pelo usuário, uma vez que nas aplicações [BRAUN,2014], estes K subconjuntos estão relacionados a K padrões que estão sendo procurados nos dados.

Neste trabalho iremos apresentar várias funções distâncias e o efeito de suas escolhas como critério de similaridade na geração de clusters de pontos no plano, gerados através do algoritmo K-means. As seções a seguir apresentam com mais detalhes o algoritmo K-means, as definições das diferentes distâncias e os resultados da dinâmica na formação dos subconjuntos.

2 K-means

O algoritmo K-means, proposto por MacQueen em 1967 [MACQUEEN,1967], é muito utilizado na construção de K clusters, que são subconjuntos, obtidos através de algum critério de escolha que evidencie a similaridade entre seus elementos. Um desses critérios de seleção é o de vizinhança, ou seja, um ponto p pertencerá a um subconjunto s_j, j=1,2...,K, se $d(p,s_j)$, a distância de p a s_j, for menor do que a distância de p aos demais subconjuntos $s_i, i \neq j$.

Aqui, inicialmente a $d(p,s_j)$ é obtida através da distância do ponto p a um ponto de referência fixado em s_j, chamado de centróide de s_j. Desta forma, todos os elementos de um subconjunto terão distância mínima em relação ao seu centróide, e serão todos denominados vizinhos entre si.

O algoritmo K-means é iterativo e depende de um chute inicial para os K primeiros centróides. A cada iteração são (i) calculadas as distâncias dos pontos aos centróides; (ii) são reagrupados os pontos de acordo com as distâncias obtidas; e (iii) para cada cluster formado, um novo centróide é obtido a partir da média aritmética dos valores contidos no clusters. Diz-se que o algorítmo convergiu quando os centróides não variarem entre duas iterações.

3 Distância

Intuitivamente distância é uma medida da separação entre dois pontos ou eventos e uma métrica é um conceito que generaliza a ideia geométrica de distância. Assim, uma métrica em um conjunto S é uma função $d : S \times S \to \mathbb{R}$ que satisfaz as seguintes propriedades:

- é positivamente definida: $d(x,y) \geq 0, \forall x,y \in S$;
- é simétrica: $d(x,y) = d(y,x), \forall x,y \in S$;
- obedece à desigualdade triangular: $d(x,z) \geq d(x,y) + d(y,z), \forall x,y,z \in S$;
- é nula apenas para pontos coincidentes: $d(x,y) = 0 \Leftrightarrow x = y$;

Apesar da distância Euclidiana ser a mais conhecida dentre as métricas, existe um grande número de métricas que são utilizadas em diferentes aplicações. A seguir, são apresentadas as formulações das métricas mais populares e que podem ser utilizadas com o algoritmo K-means.

Para todas as formulações apresentadas na próxima seção, são considerados pontos $p = (p_1, p_2, ...p_{n-1}, p_n) \in \mathbb{R}^n$, cujas coordenadas são dadas em relação a um sistema de coordenadas canônico, com $O = (0,0,...,0,0)$ sendo a origem e $B = e_1, ...e_n e_i, k = 0$ se $i \neq k$, e 1 se $i = k$, a base canônica do $\mathbb{R}^n$.

3.1 Euclidiana

A distância Euclidiana entre os pontos p e q é o comprimento do segmento de linha que os conecta. Se p e $q \in \mathbb{R}^n$, temos:

$$d(p,q) = d(q,p) = \sqrt{(p_1 - q_1)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}. \quad (1)$$

3.2 Cityblock

A distância Cityblock, também chamada de norma 1, entre dois pontos p e q em um espaço com coordenadas cartesianas fixas, é a soma dos comprimentos das projeções do segmento entre os dois pontos nos eixos coordenados. Para p e $q \in \mathbb{R}^n$.

$$d(p,q) = ||p - q||_1 = \sum_{i=1}^{n} |p_i - q_i|. \quad (2)$$

3.3 Chebyshev

A distância de Chebyshev é também conhecida como norma infinito. Dados p e $q \in \mathbb{R}^n$ temos:

$$d(p,q) = ||p - q||_\infty = \max_i |p_i - q_i|. \quad (3)$$

3.4 Canberra

A distância de Canberra entre dois pontos p e q é dada por:

$$d(p,q) = \sum_i \frac{|p_i - q_i|}{|p_i| + |q_i|}. \quad (4)$$

3.5 Bray-Curtis

A distância de Bray-Curtis entre dois pontos p e q é dada por:

$$d(p,q) = \sum |p_i - q_i| / \sum |p_i + q_i|. \quad (5)$$

A distância de Bray-Curtis está no intervalo $[0,1]$ se todas as coordenadas são positivas, e é indefinida se $\sum |p_i + q_i| \neq 0$, caso $p = -q$ a distância não está definida.

3.6 Mahalanobis

A distância de Mahalanobis entre dois vetores p e q considera a correlação entre os vetores p e q e é dada por:

$$\sqrt{(p-q)V^{-1}(p-q)^T}, \quad (6)$$

sendo $V = \begin{pmatrix} <p,p> & <q,p> \\ <p,q> & <q,q> \end{pmatrix}$ a matriz de covariância e $< .,. >$ o produto interno usual entre vetores do $\mathbb{R}^n$.

3.7 Cosseno

A distância do Cosseno entre p e $q \in \mathbb{R}^n$ é dada por:

$$1 - \frac{< p, q >}{||p||_2 ||q||_2}. \tag{7}$$

Onde $||\cdot||_2$ representa a norma euclidiana de um vetor do $\mathbb{R}^n$, dada por $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$.

4 Resultados

Nesta seção vamos mostrar os resultados obtidos na formação de clusters quando partimos de um conjunto inicial formado por 3000 pontos aleatoriamente gerados no $[-4,4] \times [-4,4] \in \mathbb{R}^2$. O valor de K escolhido para as simulações foi 3 ou 4, e os centróides correspondentes foram também escolhidos de forma aleatória.

As figuras a seguir apresentam 4 iterações do algoritmo K-means, considerando o cálculo via alguma das métricas definidas anteriormente. Em cada uma das iterações é indicado o novo valor do centróide, que é recalculado através da média aritmética dos elementos associados a cada um dos clusters, a cada iteração.

Podemos observar que diferentes agrupamentos são obtidos em cada uma das iterações, cada vez que diferentes métricas são consideradas. No entanto, apesar das diferentes dinâmicas nas evoluções dos centróides e de seus clusters, mesmo assim, há uma coerência entre os clusters finais obtidos.

Além disso, o algoritmo K-means também é sensível à escolha feita para o número de clusters. Assim, o número de clusters K também influencia na dinâmica do algoritmo, permitindo diferentes agrupamentos ao final do processo iterativo. No caso da distância Euclidiana, uma simulação com K=4 também é apresentada, evidenciando a diferença entre as vizinhanças obtidas para K=3 e K=4.

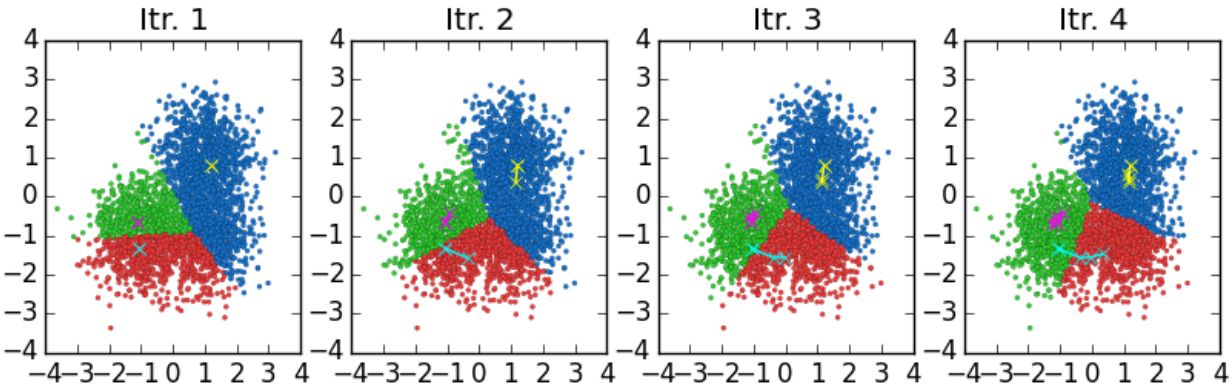

Figura 1: Quatro primeiras iterações do algoritmo K-means utilizando a métrica Euclidiana em um cenário com K=3 centróides iniciais.

Observamos, então, nas Figuras 1 e 2 os diferentes agrupamentos influêciados pela escolha do parâmetro K.

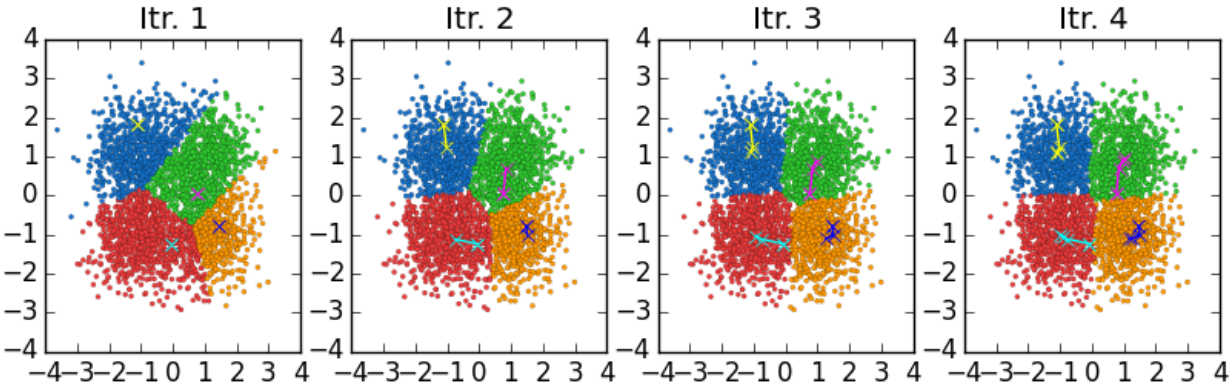

Figura 2: Quatro primeiras iterações do algoritmo K-means utilizando a métrica Euclidiana em um cenário com K=4 centróides iniciais..

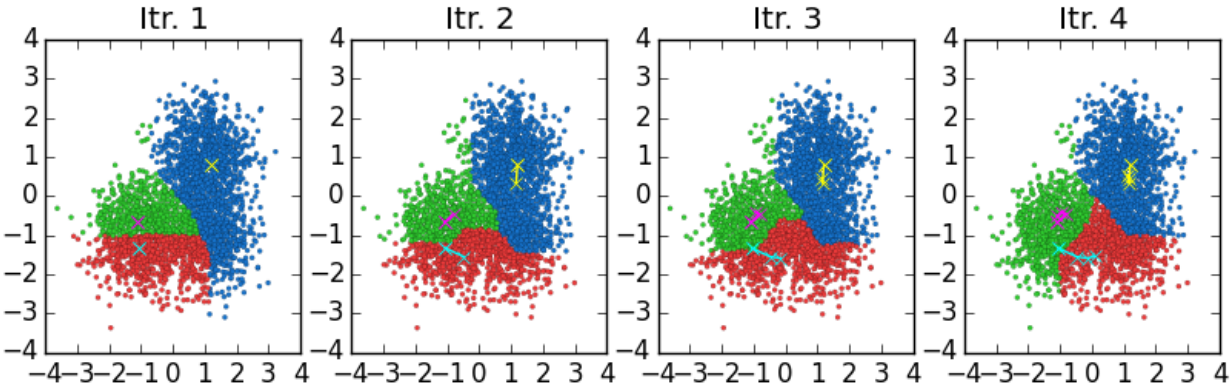

Figura 3: Quatro primeiras iterações do algoritmo K-means utilizando a métrica Cityblock em um cenário com K=3 centróides.

Observamos que as Figuras 1 e 3 apresentam dinâmicas diferentes influênciadas pela escolha das funções de distância, uma vez que o conjunto inicial de pontos é o mesmo.

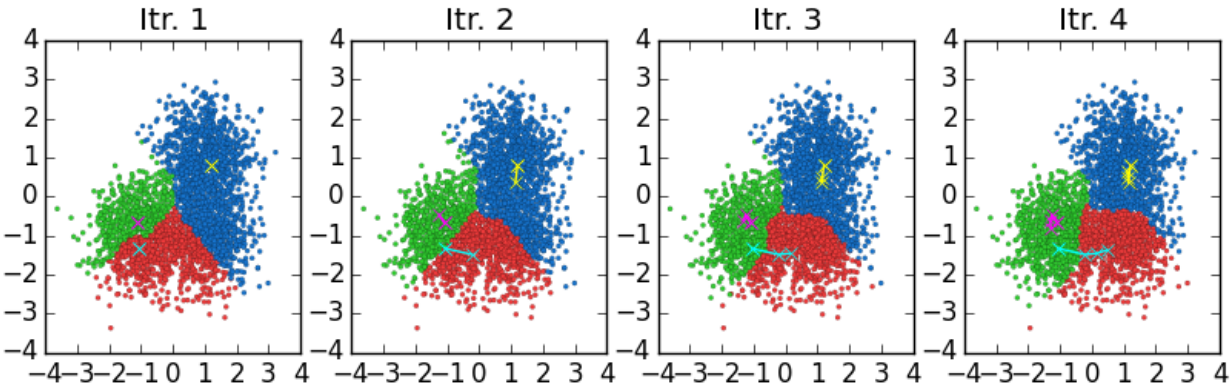

Figura 4: Quatro primeiras iterações do algoritmo K-means utilizando a métrica Chebyshev em um cenário com K=3 centróides.

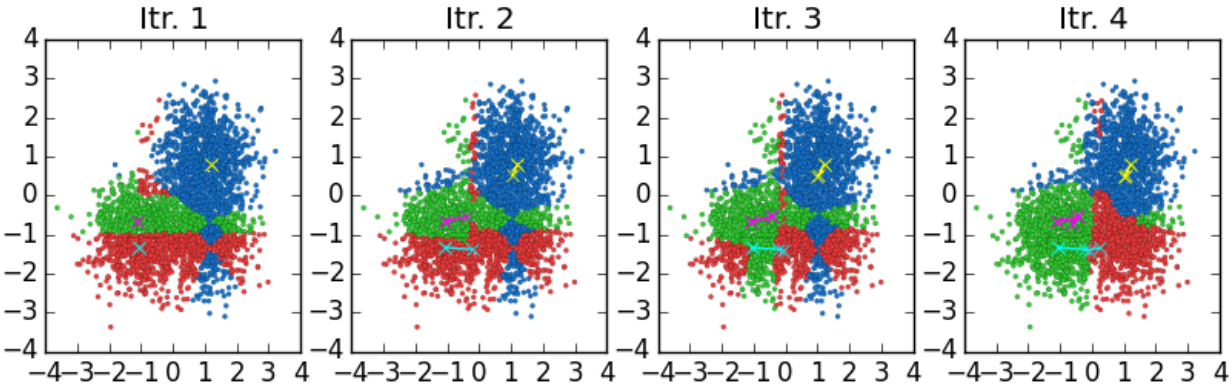

Figura 5: Quatro primeiras iterações do algoritmo K-means utilizando a métrica Canberra em um cenário com K=3.

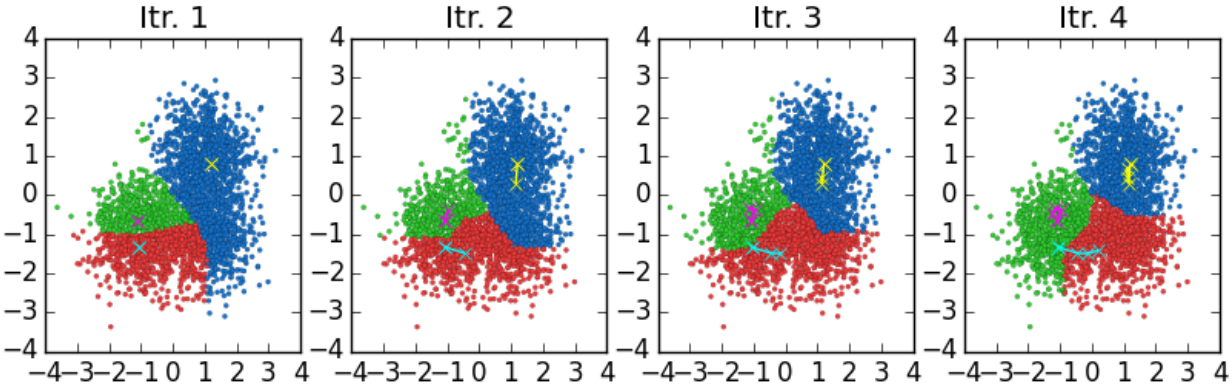

Figura 6: Quatro primeiras iterações do algoritmo K-means utilizando a métrica Bray-Curtis em um cenário com K=3.

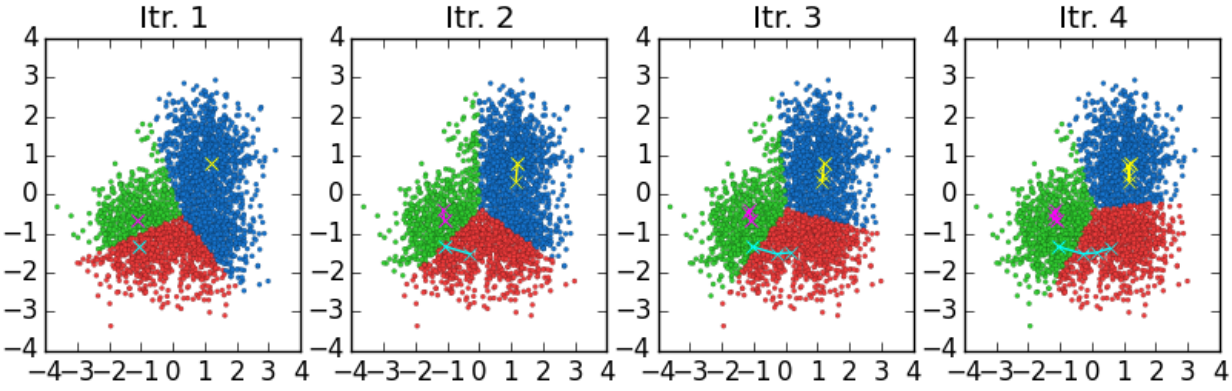

Figura 7: Quatro primeiras iterações do algoritmo K-means utilizando a métrica Mahalanobis em um cenário com K=3.

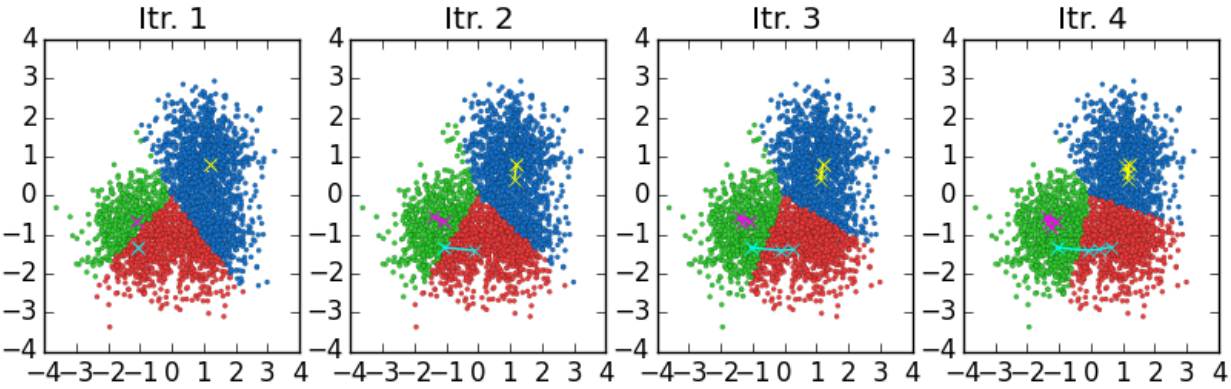

Figura 8: Quatro primeiras iterações do algoritmo K-means utilizando a métrica Cosseno em um cenário com K=3.

5 CONCLUSÃO

Neste trabalho, exploramos diferentes métricas (funções distância) na formação de agrupamentos de pontos do plano. As diferentes escolhas influenciaram na formação das K=3 vizinhanças a partir dos K=3 centróides escolhidos inicialmente para que o algoritmo K-means pudesse ser iniciado. Os gráficos das dinâmicas ao longo de 4 iterações do algoritmo K-means explicitaram a diferença nos valores obtidos para os centróides e para as configurações finais de cada uma das 3 vizinhanças obtidas. Apesar das diferenças, todas as métricas produziram resultados compatíveis e comparáveis com os demais. No caso da distância Euclidiana também foram comparados os agrupamentos obtidos a partir de K=4 centróides. Essa experiência evidencia que a classificação de elementos como pertinentes a um conjunto ou não é dependente do critério de comparação, que neste trabalho foi a métrica considerada. Pertencer a um conjunto no caso das aplicações significa possuir ou não as propriedades analisadas.

REFERÊNCIAS

BRAUN, E, RODRIGUES, C. R., BARATTO, G., KOZAKEVICIUS, A. Algoritmo K-means associado a transformadas na classificação de sinais EEG. **XXXV CNMAC CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL**, Natal, 2014. Anais do XXXV CNMAC, São Carlos, SBMAC, 2014.

ESTIVILL-CASTRO, V.; Why so many clsutering algorithms-Aposition paper. **ACM SIGKDD Explorations Newsletter**, Vol. 4, Issue 1. page 65-75, 2002. doi>10.1145/568574.568575

JOHNSON, P., **A History of Set Theory**, Prindle, Weber &Schmidt, 1972, ISBN 0871501546.

LIMA, E. L., **Espaços Métricos.** Rio de janeiro, Instituto de Matemática Pura e Aplicada-IMPA, 1977. Coleção Projeto Euclides, ISBN: 9788524401589, 337 páginas.

MACQUEEN, J. B., Some Methods for classification and Analysis of Multivariate Observations. **Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1**. University of California Press. pp. 281-297, 1967.