Projeto de Pesquisa para Iniciação Científica

O Algoritmo K-médias Geodésico para Agrupamento de Dados Baseado em Grafos

Aluno: Antônio Cícero de Amorim Azevedo
Orientador: Alexandre Luis Magalhães Levada

SETEMBRO/2023

Sumário

Resumo

O agrupamento de dados, também conhecido como *clustering*, é uma técnica fundamental no campo do aprendizado de máquina e da análise de dados. Sua importância reside no fato de ser uma abordagem não supervisionada que permite identificar padrões e estruturas ocultas nos dados, sem a necessidade de rótulos ou informações prévias sobre as classes. Um dos algoritmos mais conhecidos para essa finalidade é o K-médias. Apesar de muito utilizado, ele possui limitações, por exemplo, conseguir detectar apenas agrupamentos circulares. Neste projeto de pesquisa, propõe-se o desenvolvimento de um algoritmo K-médias geodésico, que substitui as distâncias Euclidianas pelas distâncias geodésicas em grafos. Com isso, espera-se conseguir melhores resultados no agrupamento de dados comparação com o algoritmo K-médias padrão, o que pode ser considerado um avanço científico e tecnológico em diversas aplicações de aprendizado de máquina e reconhecimento de padrões.

1. Caracterização e justificativa

A importância do algoritmo K-médias no aprendizado de máquina está na sua capacidade de encontrar grupos em grandes conjuntos de dados não rotulados, permitindo que padrões ou relacionamentos sejam descobertos sem a necessidade de conhecimento prévio dos rótulos ou categorias dos dados. Isso é particularmente útil na análise exploratória de dados. Além disso, o algoritmo K-médias é eficiente computacionalmente e relativamente fácil de entender e implementar, o que o torna uma ferramenta útil para uma ampla gama de aplicações de análise de dados. Embora o algoritmo K-médias seja amplamente utilizado e eficaz para muitas aplicações de análise de dados, ele também possui algumas limitações. Algumas das principais limitações incluem:

1. Sensibilidade aos valores iniciais: o desempenho do algoritmo K-médias pode ser fortemente afetado pelos valores iniciais dos centroides. Uma inicialização ruim pode levar a soluções subótimas ou a um número excessivo de iterações para chegar a uma solução satisfatória [1].
2. Dependência do número de agrupamentos: o número de clusters a serem criados deve ser especificado antes da execução do algoritmo. Esse número é geralmente determinado empiricamente ou por meio de métodos estatísticos, mas pode ser difícil escolher um valor ideal, especialmente quando não há conhecimento prévio sobre a estrutura dos dados [2].
3. Sensibilidade a outliers: o algoritmo K-médias pode ser sensível a valores atípicos (outliers) nos dados, uma vez que a atribuição de objetos a grupos é baseada na distância dos objetos aos centroides. Outliers podem distorcer a localização dos centroides e resultar em grupos mal definidos [3,4,5].
4. Restrição à forma dos agrupamentos: o algoritmo K-médias assume que os clusters têm formas esféricas e de tamanhos semelhantes. Isso pode ser uma limitação para conjuntos de dados com clusters de formas irregulares ou tamanhos muito diferentes [3,6,7].
5. Limitações em conjuntos de dados de alta dimensão: o algoritmo K-médias pode enfrentar dificuldades em conjuntos de dados de alta dimensão, uma vez que a distância entre objetos pode se tornar menos discriminativa em altas dimensões e a geometria dos clusters pode se tornar mais complexa [3,8].

Como forma de atenuar os efeitos negativos das limitações 4 e 5 discutidas acima, propõe-se uma versão topológica do algoritmo K-médias em que as distâncias Euclidianas são substituídas pelas distâncias geodésicas computadas a partir do grafo de adjacências induzido a partir do conjunto de dados, denominada de K-médias geodésico. Com o auxílio do grafo de adjacências, é possível capturar a forma dos dados, ou seja, diminuir a restrição de agrupamentos circularmente simétricos imposta pela distância Euclidiana. Além disso, como na definição do grafo utiliza-se apenas o cálculo de distâncias Euclidianas locais, isto é, entre pontos vizinhos, o fenômeno do espaço vazio [9] e o fenômeno da concentração [10, 11] tendem a ser atenuados, uma vez que o conjunto formado por uma amostra e seus vizinhos pode ser bem representado por um patch aproximadamente linear.

A proposta consiste em desenvolver e implementar o algoritmo K-médias geodésico em linguagem Python. Experimentos computacionais serão projetados para comparar o desempenho do método proposto com o algoritmo padrão. Espera-se conseguir melhora sensível na qualidade dos agrupamentos obtidos em diversos conjuntos de dados multivariados.

2. Objetivos

Nesta seção serão descritos os objetivos gerais e específicos a serem atingidos ao longo do desenvolvimento do projeto.

2.1. Objetivos gerais

Propor e implementar um novo algoritmo para agrupamento de dados, denominado K-médias geodésico, que substitui a distância Euclidiana por distâncias geodésicas obtidas a partir de caminhos mínimos em grafos.

2.2. Objetivos específicos

Os objetivos específicos ajudam a nortear o desenvolvimento e o cronograma de execução do projeto:

• Estudar o algoritmo K-médias padrão e suas implementações computacionais.

• Estudar ao problema dos caminhos mínimos em grafos e seus algoritmos (Djikstra).

• Estudar métricas quantitativas para medir a qualidade de agrupamentos.

• Projetar e desenvolver o algoritmo K-médias geodésico.

• Desenvolver uma implementação própria do algoritmo K-médias geodésico em Python.

• Testar e validar o método proposto em comparação com o algoritmo K-médias padrão.

3. Fundamentação teórica

Esta seção apresenta a fundamentação teórica que embasa a metodologia proposta a ser desenvolvida pelo projeto de pesquisa em questão.

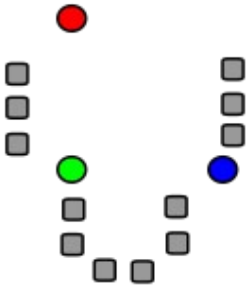
3.1. O algoritmo K-médias

Trata-se de um método de aprendizado não supervisionado, uma vez que não faz uso de rótulos disponíveis no conjunto de treinamento. A ideia básica consiste em agrupar amostras do conjunto de dados de acordo com similaridade, sendo que na versão padrão do algoritmo, emprega-se a distância Euclidiana. Variações com outras medidas de distância como a norma L1, a distância de Mahalanobis e distância do cosseno já foram desenvolvidas na literatura [12, 13]. O problema de agrupamento pelo algoritmo K-médias é formulado como segue. Dado um conjunto de dados $X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}$ com $\vec{x}_i \in R^d$, o algoritmo tem como objetivo particionar as n amostras em k < n grupos

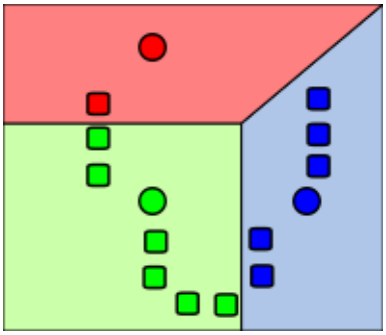
$S=\{s_1, s_2, \ldots, s_k\}$ de modo a minimizar o espalhamento intra-cluster, ou seja, o objetivo consiste em encontrar a partição ótima no sentido de minimizar o seguinte critério:

$$\underset{S}{argmin} \sum_{i=1}^{k} \sum_{\vec{x} \in s_i} \|\vec{x} - \vec{\mu}_i\|^2 \quad (1)$$

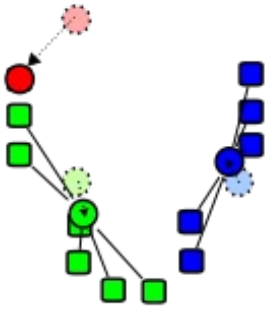
onde μ_i é o centroide (ponto médio) da partição s_i . Note que, pela formulação matemática, deseja-se que os centros dos grupos estejam o mais próximo possível dos dados. Pode-se verificar que encontrar a solução ótima para esse problema é NP-Hard para um número arbitrário de dimensões. A heurística adotada pelo K-médias para simplificar o problema consiste em fixar o valor de K, ou seja, definir o número de agrupamentos desejados a priori. Em resumo, o funcionamento do algoritmo segue a seguinte lógica:



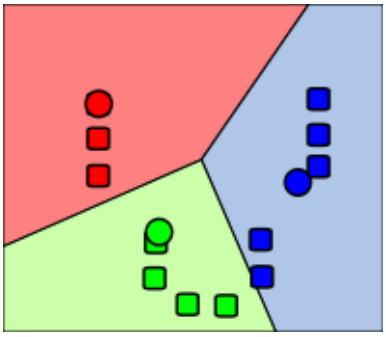
A) K sementes são geradas aleatoriamente



B) K grupos são criados associando cada amostra a semente mais próxima



C) Computa-se o novo centroide de cada um dos K grupos como sendo a média dos elementos do grupo



D) Os passos anteriores são repetidos até que a convergência seja atingida (centroides não se movem)

O algoritmo a seguir resume os passos do algoritmo K-médias [14].

1. Especificar um conjunto de K centros (escolher K amostras aleatoriamente)

    - K é o parâmetro do método (representa o número desejado de agrupamentos).

    - Escolha aleatória do conjunto de amostras.

2. Associar cada amostra x_i ao agrupamento mais próximo (centroide mais próximo), de acordo com a distância Euclidiana.

3. Atualiza os centros dos agrupamentos (média dos pontos do agrupamento)

$$\mu_j^{(t+1)} = \frac{1}{N_j} \sum_{x_i \in \omega_j} \vec{x}_i \quad \text{para} \quad j = 1, 2, \ldots, c \tag{2}$$

4. Se não houverem mudanças nos rótulos (centros não mudaram), algoritmo convergiu. Senão, retorne ao passo 2.

Esse algoritmo possui algumas limitações, tais como [15]:

- Sensibilidade à escolha aleatória do conjunto de centros iniciais. Diversas vezes é comum utilizar-se de algum tipo de supervisão na seleção dos centros iniciais, utilizando uma amostra representativa de cada classe.
- Escolha incorreta do número de agrupamentos (K).
- Dados com anisotropia (dados fortemente correlacionados) geralmente levam a resultados inadequados, devido a distância Euclidiana.
- Dados de alta dimensionalidade sofrem com o fenômeno da concentração, devido a distância Euclidiana.

3.2. Distâncias geodesicas em grafos

Encontrar caminhos mínimos em grafos é um dos mais importantes problemas da computação, em grande parte por ser utilizado em aplicações nas mais diversas áreas da ciência. Nesta seção serão apresentados conceitos sobre como obter distâncias geodésicas em grafos ponderados a partir do algoritmo de Dijkstra.

Def: Caminho ótimo

Seja G = (V, E, w) com $w: E \rightarrow R^{+}$ uma função de custo para as arestas. Um caminho P* de v_0 a v_n é ótimo se seu peso

$$w(P^{*}) = \sum_{i=0}^{n-1} w(v_i, v_{i+1}) = w(v_0, v_1) + w(v_1, v_2) + \ldots + w(v_{n-1}, v_n) \quad (3)$$

é o menor possível. Veremos a seguir que caminhos ótimos possuem a propriedade de subestrutura ótima.

Teorema: Seja G = (V, E, w) um grafo ponderado e $P=v_0 v_1 v_2 ... v_n$ o caminho mínimo de v_0 a v_n. Seja ainda, para $0 \leq i < j \leq n$, $P' = v_i v_{i+1} ... v_j$ o subcaminho de P de v_i a v_j. Então, P' é o caminho mínimo de v_i a v_j.

Prova:

1. Podemos decompor P em 3 subcaminhos:

2. Logo, o peso total do caminho P é dado por:

$$w(P) = w(P_{0i}) + w(P_{ij}) + w(P_{jn}) \quad (4)$$

3. Suponha que exista um caminho P'_{ij} de v_i a v_j tal que $w(P'_{ij}) < w(P_{ij})$.

4. Então, o caminho $\overline{P}$, definido como:

com peso $w(\overline{P})=w(P_{0i})+w(P'_{ij})+w(P_{jn})$ possui $w(\overline{P})<w(P)$, o que contradiz a hipótese de que P é o caminho mínimo. Portanto, não existe P'_{ij} diferente de P_{ij} com $w(P'_{ij})<w(P_{ij})$. Esse resultado é muito importante pois permite a aplicação da programação dinâmica, o que acelera muito o cômputo de distâncias geodésicas.

A seguir, é apresentado o algoritmo de Dijkstra para obtenção de caminhos mínimos em grafos ponderados utilizando programação dinâmica. A ideia do algoritmo de Dijkstra é utilizar uma estratégia Bottom-Up para a reversão da recursão. Mas porque o algoritmo de Dijkstra é um algoritmo de programação dinâmica? Basicamente, por dois motivos:

1. Não recalcula os custos dos caminhos a toda iteração: armazena os valores dos caminho mínimos em uma Fila de Prioridades Q

2. Constrói a solução ótima a partir das soluções dos subproblemas menores (subestrutura ótima de caminhos mínimos possui sobreposição)

A seguir, são definidas as principais variáveis utilizadas no algoritmo:

λ(v): menor custo até o momento para o caminho de s (raiz) até v.

π(v): predecessor do vértice v na árvore de caminhos mínimos (de onde vim ao entrar no vértice v).

Q: fila de prioridades dos vértices.

A fila de prioridades Q possui três primitivas básicas:

a) Insert(Q, v): insere um vértice v no fim da fila Q.

b) ExtractMin(Q): remove da fila o vértice de maior prioridade.

c) DecreaseKey(Q, v, λ(v)): modifica a prioridade do vértice v da fila Q, atualizando o seu valor de λ(v) (note que sempre irá diminuir esse valor que é o tamanho do caminho mínimo)

Uma observação importante é que quanto menor o valor de λ(v), maior a prioridade do vértice v. Além disso, no algoritmo, define-se o conjunto S como o conjunto dos vértices já finalizados (vértices para os quais a distância geodésica já foi obtida). O pseudo-código a seguir ilustra o algoritmo de Dijkstra [16].

```
Dijkstra(G, w, s) {
     for each v ∈ V {
          λ(v) = ∞
          π(v) = nil
     }
     λ(s) = 0
     S = ∅
     Q = ∅
     for each v ∈ V
          Insert(Q, v)
     while Q ≠ ∅ {
          u = ExtractMin(Q)
          S = S ∪ {u}
          for each v in N(u) {
               λ(v) = min{λ(v), λ(u) + w(u,v)}
               if λ(v) was updated {
                    π(v) = u
                    Decrease_Key(Q, v, λ(v))
               }
          }
     }
}
```

O algoritmo inicia atribuindo valor infinito à distância geodésica de todos os vértices, com exceção da raiz, que recebe valor zero. Em seguida todos os vértices são inseridos na fila de prioridades Q. O laço principal remove a cada iteração o vértice u com menor valor de λ(v) e para cada um de seus vizinhos atualiza o valor de λ(v) verificando se é uma boa ideia passar por u para atingir v. Em caso de modificação da distância, atualiza-se o predecessor de v na árvore de caminhos mínimos e modifica-se a prioridade do vértice v na fila de prioridades. O resultado a seguir, garante que ao fim do algoritmo de Dijkstra, temos as distâncias geodésicas da raiz s a todos os demais vértices do grafo.

Teorema: O algoritmo de Dijkstra termina com $\lambda(v) = d(s,v), \forall\, v \in V$, onde d(s, v) é a distância geodésica de s a v (menor distância possível).

Para uma demonstração formal detalhada, recomenda-se consultar o livro "*Introduction to Algorithms*" de Thomas Cormen e colegas [16].

4. Metodologia proposta

A utilização de distâncias geodésicas em grafos para fins de segmentação de imagens e classificação supervisionada foram exploradas com sucesso em algoritmos como a Transformada Imagem Floresta (IFT) [17] e o algoritmo Floresta de Caminhos Ótimos (OPF) [18]. Porém, no contexto de agrupamento de dados, em que não se tem acesso aos rótulos das classes, tais distãncias ainda não foram muito exploradas. As principais contribuições do emprego da distância geodésica no lugar da distância Euclidiana são:

- Aumentar a robustez em relação a presença de ruídos e outliers nos dados.
- *Manifold hypothesis*: estudos tem mostrado que conjuntos de dados de alta dimensionalidade podem ser representados por uma variedade de menor

dimensão imersa em um espaço ambiente de alta dimensão, o que faz com que a distância Euclidiana não seja uma métrica adequada.

- Aumentar o poder discriminante da métrica, uma vez que em espaços de alta dimensão, a capacidade discriminatória da distância Euclidiana é severamente prejudicada.

O algoritmo a ser proposto, denominado K-médias geodésico, busca induzir um grafo a partir do conjunto de amostras, com intuito de aproximar a variedade que representa a estrutura geométrica oculta dos dados. Dessa forma, torna-se possível utilizar as distâncias geodésicas no grafo como medida de distância entre os vértices. Em resumo, o algoritmo K-médias geodésico é dado pelos seguintes passos:

1. A partir dos dados observados, induzir um grafo KNN, isto é, um grafo G em que uma amostra x_i se liga a uma amostra x_j se ela for um dos seus K vizinhos mais próximos. Nesta etapa, utiliza-se a distância Euclidiana para identificar os vizinhos mais próximos. Essa escolha é razoável, pois localmente, uma variedade é como um espaço Euclidiano.
2. Escolher K vértices do grafo para serem sementes aleatórias, ou seja, os centroides de cada classe. Pode-se aplicar algoritmos de particionamento de grafos para que cada semente seja proveniente de uma partição.
3. Utilizando o algoritmo de Dijkstra, calcular as distâncias geodésicas entre cada amostra e os respectivos centroides, atribuindo a cada amostra o rótulo da semente mais próxima.
4. Calcule os novos centroides, com base na rotulação nova.
5. Se a movimentação dos centros é pequena o suficiente, pare. Senão, volte para o passo 3.

Nos experimentos computacionais, a expectativa é medir o desempenho do método proposto com métricas de avaliação de qualidade de agrupamentos, como o índice de Rand [19], o índice de Fowlkes-Mallows [20], o coeficiente Silhoueta [21], o índice de Calinski-Harabasz [22] e o índice de Davies-Bouldin [23]. Pretende-se selecionar um conjunto com 20 a 30 conjuntos de dados distintas, de modo que ao final, um teste de hipóteses não paramétrico, como o teste de Wilcoxon com nível de significância de 5% [24], possa ser aplicado para verificar se existem diferenças significativas entre o desempenho do algoritmo proposto e do K-médias padrão.

4.1. Bases de dados

O repositório openML, acessível a partir da URL www.openml.org, possui inúmeros conjuntos de dados reais para utilização em aplicações de aprendizado de máquina e reconhecimento de padrões. Ele é mantido principalmente para apoiar o desenvolvimento de pesquisas na área de inteligência artificial e ciência de dados. Em todos os experimentos a serem executados neste projeto, serão utilizados conjuntos de dados disponíveis nesse repositório. Dentre tais conjuntos de dados, destacam-se os conhecidos MNIST (Modified National Institute of Standards and Technology), que é composto por imagens de dígitos numéricos escritos a mão, o Fashion-MNIST, que é composto por imagens de roupas e acessórios vestíveis, e o CIFAR-10, que é composto por imagens de automóveis e animais.

5. Resultados Esperados

Os resultados desse projeto de pesquisa são caracterizados de duas maneiras:

Desenvolvimento científico: Os métodos de agrupamento de dados são considerados fundamentais em diversas aplicações de ciência de dados e aprendizado de máquina. Assim, ao propor e desenvolver um novo algoritmo K-médias geodésico que combina

caminhos mínimos em grafos com agrupamento de dados na tentativa de superar as limitações do K-médias padrão, espera-se contribuir com a evolução das técnicas de aprendizado não supervisionado. A expectativa é que seja possível melhorar a qualidade dos agrupamentos detectados. Além disso, após o término das pesquisas, a implementação em Python do algoritmo K-médias geodésico será compartilhada com toda a comunidade científica por meio de repositórios online. Portanto, em termos científicos, as principais contribuições deste projeto estão diretamente relacionadas com a produção de novos conhecimentos na área de aprendizado de máquina e reconhecimento de padrões.

Desenvolvimento pessoal: o desenvolvimento desse projeto proporciona o(a) aluno(a) a aprendizagem de métodos de pesquisa e estimula o desenvolvimento do pensamento científico através da metodologia e da pesquisa, que são fundamentais para a formação acadêmica.

6. Cronograma de Atividades

1. Realizar estudos sobre a fundamentação teórica do algoritmo K-médias.

2. Realizar estudos sobre a fundamentação teórica do algoritmo de Dijkstra.

3. Realizar estudos sobre medidas de qualidade de agrupamento.

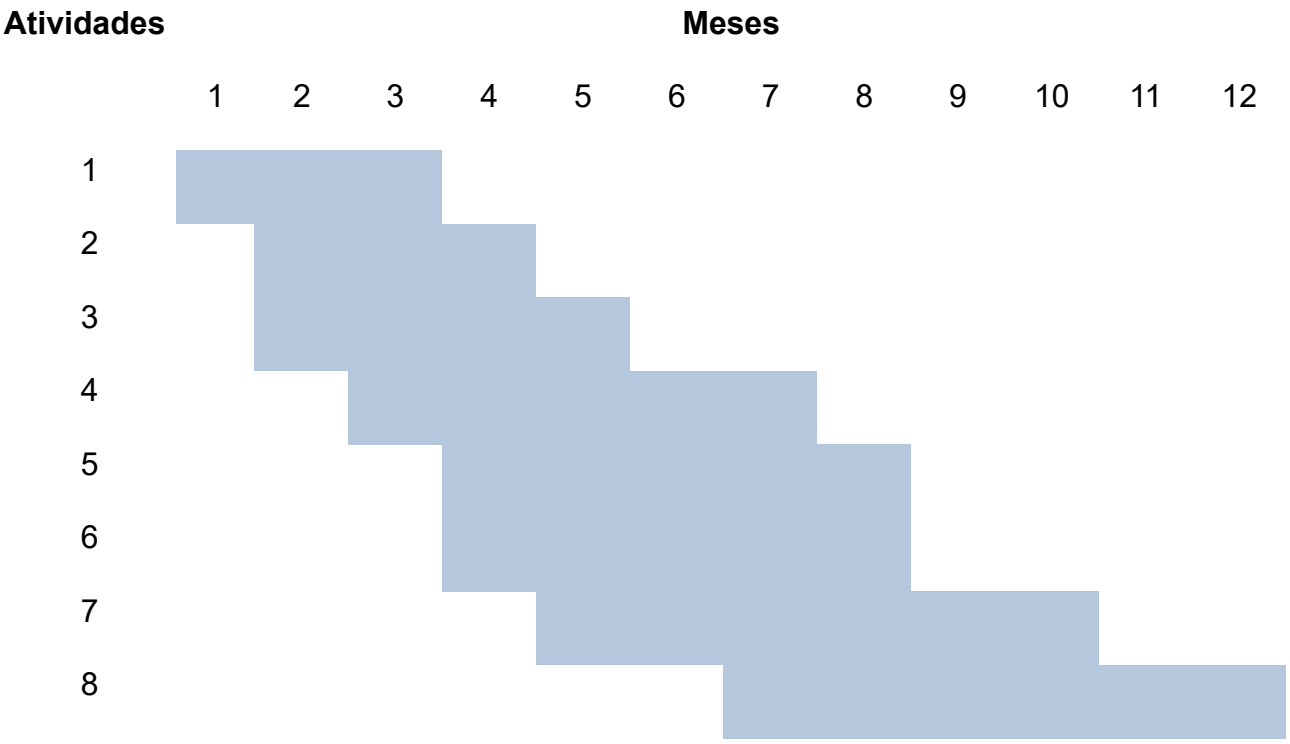
4. Desenvolver e implementar o algoritmo K-médias geodésico proposto em linguagem Python.

5. Realizar teste, validação e comparação do método proposto com o K-médias padrão e eventualmente outros métodos similares.

6. Aquisição e busca de conjuntos de dados provenientes de repositórios online.

7. Realizar testes com o método implementado em conjuntos de dados reais.

8. Escrita do relatório técnico final do projeto de pesquisa.



Referências

[1] D. Arthur and S. Vassilvitskii, “k-means++: the advantages of careful seeding,” in SODA ’07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[2] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 63, no. 2, pp. 411–423, 2001.

[3] A. K. Jain, “Data clustering: 50 years beyond k-means,” Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010.

[4] C. C. Aggarwal, Data Mining: The Textbook. Springer, 2015.

[5] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Elsevier, 2011.

[6] J. C. Bezdek, “Some new indexes of cluster validity,” IEEE Transactions on systems, man, and cybernetics, vol. 28, no. 3, pp. 301–315, 1998.

[7] U. Luxburg, “A tutorial on spectral clustering,” Statistics and computing, vol. 17, no. 4, pp. 395–416, 2007.

[8] D. Steinley, “K-means clustering: A half-century synthesis,” British Journal of Mathematical and Statistical Psychology, vol. 70, no. 1, pp. 1–25, 2017.

[9] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, “On the surprising behavior of distance metrics in high dimensional space,” in International Conference on Database Theory, pp. 420–434, Springer, 2001.

[10] D. L. Donoho, “High-dimensional data analysis: The curses and blessings of dimensionality,” in Proceedings of the AMS Conference on Math Challenges of the 21st Century, pp. 32–56, 2000.

[11] S. Dasgupta, “The hardness of approximation: The concentration of measure phenomenon,” in Proceedings of the 35th ACM Symposium on Theory of Computing (STOC), pp. 534–543, 2003.

[12] P. O. Brown, M. C. Chiang, S. Guo, Y. Jin, C. K. Leung, E. L. Murray, A. G. M. Pazdor and A. Cuzzocrea. "Mahalanobis Distance Based K-Means Clustering". In: Wrembel, R., Gamper, J., Kotsis, G., Tjoa, A.M., Khalil, I. (eds) Big Data Analytics and Knowledge Discovery. DaWaK 2022. Lecture Notes in Computer Science, vol 13428. Springer, Cham. *https://doi.org/10.1007/978-3-031-12670-3_23*.

[13] M. K. Khan, S. Sarker, S. M. Ahmed and M. H. A. Khan, "K-Cosine-Means Clustering Algorithm," *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, Khulna, Bangladesh, 2021, pp. 1-4, doi: 10.1109/ICECIT54077.2021.9641480.

[14] M. E. Celebi, H. A. Kingravi and P. A. Vela. "A comparative study of efficient initialization methods for the k-means clustering algorithm", Expert Systems with Applications, 40 (1), pp. 200–210, 2013, doi:10.1016/j.eswa.2012.07.021

[15] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data", Information Sciences, volume 622, 2023, pp. 178-210.

[16] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, Introduction to Algorithms, 4a edição, MIT Press, 2022.

[17] A. X. Falcão, J. Stolfi, R. A. Lotufo, "The image foresting transform: theory, algorithms, and applications", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, v. 26, n. 1, 2004, pp. 19-29.

[18] J. P. Papa, A. X. Falcão, V. H. C. de Albuquerque, J. M. R.S. Tavares, "Efficient supervised optimum-path forest classification for large datasets", Pattern Recognition, vol. 45, n. 1, 2012, pp. 512-520, *https://doi.org/10.1016/j.patcog.2011.07.013*.

[19] W. M. Rand, "Objective criteria for the evaluation of clustering methods". Journal of the American Statistical Association. American Statistical Association. 66 (336), 1971, pp. 846–850.

[20] E. B. Fowlkes, C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings". Journal of the American Statistical Association. 78 (383), 1983, pp. 553–569.

[21] P. J. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis", Computational and Applied Mathematics. v. 20, 1987, pp. 53–65.

[22] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis." Communications in Statistics-Simulation and Computation, v. 3 , no. 1, 1974, pp. 1--27.

[23] D. L. Davies, D. W. Bouldin, "A Cluster Separation Measure". IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1, vol. 2, 1979, pp. 224–227.

[24]. F. Wilcoxon, Individual comparisons by ranking methods. Biometrics Bulletin. 1 (6), pp. 80-83, 1945.