UNIVERSIDADE FEDERAL DE SÃO CARLOS
PRÓ-REITORIA DE PESQUISA
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

RELATÓRIO FINAL

AGRUPAMENTO DE DADOS PARTICIONAL UTILIZANDO CAMINHOS MÍNIMOS EM GRAFOS

Outubro, 2024

Fernando Borges
ICTSR
Estatística – DEs

UNIVERSIDADE FEDERAL DE SÃO CARLOS
PRÓ-REITORIA DE PESQUISA
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

AGRUPAMENTO DE DADOS PARTICIONAL UTILIZANDO CAMINHOS MÍNIMOS EM GRAFOS

Relatório Final apresentado à Universidade Federal de São Carlos, Pró-Reitoria de Pesquisa, Programa Institucional de Voluntariado de Iniciação Científica, sob orientação do Prof. Alexandre Luís Magalhães Levada, Departamento de Computação.

Outubro, 2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS
PRÓ-REITORIA DE PESQUISA
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

SUMÁRIO

UNIVERSIDADE FEDERAL DE SÃO CARLOS
PRÓ-REITORIA DE PESQUISA
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

RESUMO

O agrupamento de dados é uma tarefa extremamente relevante para a área de aprendizado de máquina e um dos algoritmos mais conhecidos e utilizados é o k-médias. Assim como diversos outros algoritmos que visam reduzir problemas intrínsecos a essa tarefa, o algoritmo k-médias surgiu com o objetivo de minimizar a variância intra-grupo e maximizar a variância entre-grupos. Propomos uma extensão a esse algoritmo, o k-médias topológico, que substitui a distância euclidiana por comprimentos de caminhos mínimos em grafos como medida de dissimilaridade. Dado que o algoritmo proposto é baseado em grafos, sua complexidade computacional é menor do que diversos métodos modernos de agrupamento de dados e, além disso, apresenta diversas características desejáveis que o k-médias tradicional não é capaz de fornecer. Resultados empíricos em dados reais indicam que o k-médias topológico é capaz de melhorar a qualidade dos agrupamentos quando comparado ao k-médias tradicional, comparando-se à performance de outros métodos modernos, como o algoritmo HDBSCAN, por apresentar uma maior capacidade de lidar com dados de alta dimensão e agrupamentos não-lineares.

1. INTRODUÇÃO

Os algoritmos de agrupamento de dados são técnicas utilizadas na área de aprendizado de máquina com grande relevância na análise e reconhecimento de padrões em dados [1]. Em geral, o objetivo desses métodos é agrupar observações similares com base nas características intrínsecas ao conjunto de dados, sem a necessidade de rótulos predefinidos.

Entre os diversos métodos que existem, o k-médias é um dos mais conhecidos e utilizados devido à sua simplicidade e eficiência computacional [2]. Ao utilizar a distância euclidiana para medir a proximidade entre as observações e os k centroides definidos, o algoritmo realoca cada ponto ao agrupamento com centroide mais próximo e repete o processo até convergência [3]. Dessa forma, o k-médias é considerado um método particional que visa dividir o conjunto de dados em k grupos ao minimizar a soma do erro quadrático das distâncias entre os pontos e o centroide do grupo a qual estão associados [4]. No entanto, sua versão tradicional enfrenta diversas limitações em dados de maior dimensionalidade ou na presença de *outliers*, o que reduz a qualidade de seus agrupamentos e restringe sua capacidade de detectar agrupamentos não lineares [5][6].

Alguns métodos não supervisionados modernos são capazes de mitigar algumas dessas limitações presentes no k-médias. O algoritmo HDBSCAN, por exemplo, torna possível extrair agrupamentos de formas e densidades variadas na presença de ruído e, assim, permite uma alta flexibilidade aos dados em diferentes cenários [7][8]. Apesar disso, o HDBSCAN possui algumas limitações relacionadas a sua complexidade computacional e sua sensibilidade aos hiperparâmetros, o que o torna menos eficiente em conjuntos de dados grandes [9].

A dificuldade desses métodos de agrupamento em lidar com dados de alta dimensão, isto é, cenários em que o número de parâmetros é muito maior que o de observações presentes no conjunto de dados, é um problema comum que recebe o nome de “maldição da dimensionalidade” [10]. Esse tipo de cenário é suscetível a esse fenômeno uma vez que a distância entre objetos se torna menos discriminativa em altas dimensões e a complexidade geométrica dos agrupamentos pode aumentar [11][12].

O presente projeto propõe uma variante do algoritmo k-médias, denominada k-médias topológico, que utiliza distâncias geodésicas calculadas por meio do algoritmo de Dijkstra na representação dos dados em um grafo de k vizinhos mais próximos. O grafo é gerado por meio do cálculo da distância euclidiana entre uma observação e as demais para que seja criada uma aresta entre uma observação e as k observações mais próximas a ele [13][14]. Espera-se que o grafo gerado seja um único componente. A ideia é que a distância geodésica, ao ser utilizada para definir a proximidade entre os pontos nesse grafo de adjacências induzido pelos dados, permita a detecção de agrupamentos com maior complexidade geométrica e seja uma alternativa à distância euclidiana por mitigar os efeitos negativos das limitações em cenários de alta dimensionalidade e por diminuir a restrição de agrupamentos circularmente simétricos.

2. OBJETIVOS E METODOLOGIA

O principal objetivo desse projeto é comparar a performance do algoritmo proposto, o k-médias topológico, com o algoritmo k-médias tradicional e o HDBSCAN, um método moderno que é capaz de detectar agrupamentos de qualquer formato na presença de ruído. Para a realização da análise comparativa, o método de trabalho envolveu a seleção de dois grupos de conjuntos de dados a serem utilizados nos experimentos e três métricas de avaliação dos agrupamentos. Todos os conjuntos de dados utilizados estão disponíveis no repositório www.openml.org.

O primeiro grupo de conjunto de dados foi utilizado nos experimentos que envolveram a comparação dos algoritmos k-médias tradicional e topológico. A alta dimensionalidade e a presença de poucas observações são aspectos em comum a todos os conjuntos de dados que formam esse grupo, tendo em vista que o interesse é comparar a performance dos algoritmos na influência da “maldição da dimensionalidade”. A Tabela 1 descreve as características dos conjuntos de dados do primeiro grupo.

O segundo grupo de conjunto de dados foi utilizado nos experimentos que envolveram a comparação do algoritmo k-médias topológico e o algoritmo HDBSCAN. Com o intuito de comparar a performance dos dois métodos, os conjuntos de dados escolhidos nesse grupo

apresentam agrupamentos de alta densidade com um alto valor de observações e classes. A Tabela 2 descreve as características dos conjuntos de dados do segundo grupo.

Tabela 1: Número de observações, parâmetros e classes de cada conjunto de dados selecionado para os experimentos em alta dimensão (grupo 1).

Número	Conjunto de dados	Observações	Parâmetros	Classes
1	AP_Colon_Kidney	546	10935	2
2	AP_Prostate_Kidney	329	10935	2
3	AP_Breast_Colon	630	10935	2
4	tr12.wc	313	5804	8
5	tr31.wc	927	10128	7
6	tr45.wc	690	8261	10
7	SRBCT	83	2308	4
8	pasture	36	22	2
9	leukemia	72	7129	2
10	GCM	190	16063	14

Tabela 2: Número de observações, parâmetros e classes de cada conjunto de dados selecionado para os experimentos em agrupamentos de alta densidade (grupo 2).

Número	Conjunto de dados	Observações	Parâmetros	Classes
1	digits	1797	64	10
2	JapaneseVowels	9961	14	9
3	optdigits	5620	64	10
4	satimage	6430	36	6
5	waveform-5000	5000	40	3
6	abalone	4177	8	28
7	semeion	1593	256	10
8	cnae-9	1080	856	9
9	mfeat-pixel	2000	240	10
10	micro-mass	360	1300	10

As métricas escolhidas para medir o desempenho dos métodos com relação à qualidade de seus agrupamentos foram o índice de Rand [15], a Informação Mútua Ajustada (em inglês, Adjusted Mutual Information) [16] e o índice de Fowlkes-Mallows [17]. Além disso, realizou-se um teste de Wilcoxon [18] com nível de significância de 5% para verificar se houve uma diferença significativa de desempenho em cada grupo de experimentos.

3. RESULTADOS E DISCUSSÃO

Para a obtenção dos resultados nos dois grupos de experimentos, foi calculado a média de 30 execuções de cada algoritmo para cada uma das três métricas de avaliação com uma escolha aleatória de centroides iniciais. O parâmetro k (número de agrupamentos) foi definido como o número distinto de classes no conjunto de dados. Para o k-médias topológico, o número de vizinhos do grafo de k vizinhos mais próximos foi calculado pela raiz quadrada do número de observações, isto é, foi definido como $\lfloor \sqrt{n} \rfloor$ em um conjunto de dados com n observações. Como existe a necessidade de que não sejam gerados grafos desconexos, o número de vizinhos deve ser incrementado até que o grafo gerado esteja conectado.

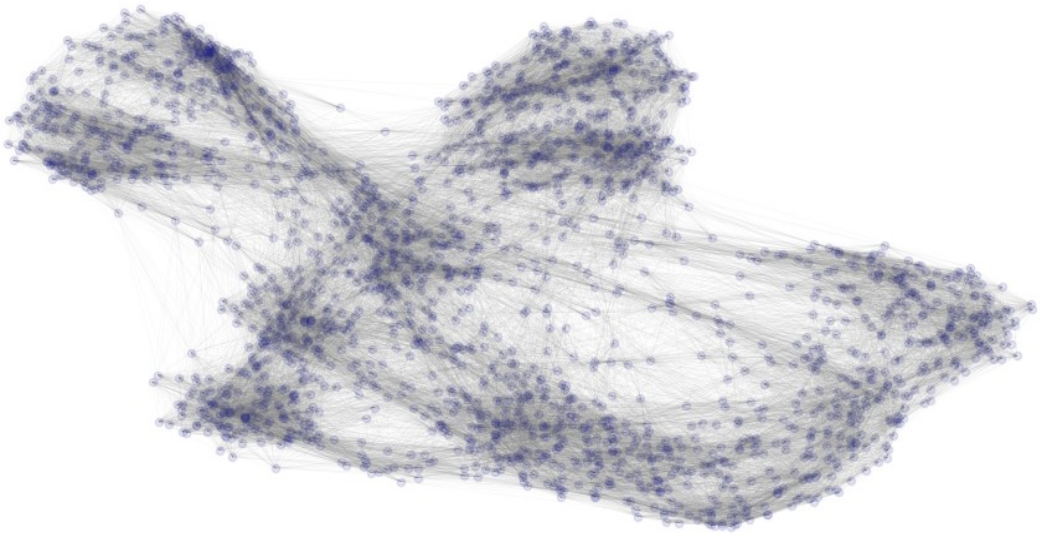

Figura 1: Grafo de k vizinhos mais próximos para o conjunto de dados mfeat-pixel com $k = 45$ *vizinhos.*

Tabela 3: Médias para cada métrica de avaliação dos agrupamentos após 30 execuções dos algoritmos k-médias tradicional e topológico em dados de alta dimensão (grupo 1).

Número	K-médias tradicional			K-médias topológico		
	Rand	**AMI**	**Fowlkes-Mallows**	**Rand**	**AMI**	**Fowlkes-Mallows**
1	0.574	0.133	0.582	**0.691**	**0.341**	**0.715**
2	0.634	0.252	0.702	**0.678**	**0.296**	**0.756**
3	0.509	0.011	0.514	**0.582**	**0.156**	**0.627**
4	0.307	0.039	0.347	**0.510**	**0.154**	**0.329**
5	0.375	0.078	**0.466**	**0.484**	**0.144**	0.393
6	0.387	0.131	**0.315**	**0.533**	**0.185**	0.294
7	0.600	0.143	0.362	**0.669**	**0.283**	**0.436**
8	**0.711**	**0.372**	0.565	0.677	0.350	**0.569**
9	0.556	0.108	0.610	**0.576**	**0.124**	**0.656**
10	0.871	**0.365**	**0.331**	**0.880**	0.332	0.300

Os resultados para a comparação de performance entre os algoritmos k-médias tradicional e topológico em dados de alta dimensão estão dispostos na Tabela 3. Observa-se que o algoritmo k-médias topológico apresentou uma performance próxima ou superior ao k-médias tradicional para as três métricas de avaliação em todos os conjuntos de dados selecionados. Para checar a significância das diferenças de performance entre os dois métodos, o teste não paramétrico de Wilcoxon para amostras pareadas foi utilizado. De acordo com os testes, ao nível de significância de 5%, há evidências de que o k-médias topológico obteve performances superiores ao k-médias tradicional ao levar em consideração o índice de Rand (p-valor de 0.01) e AMI (p-valor de 0.02). Entretanto, ao mesmo nível de significância, não rejeitamos a hipótese nula em termos do índice de Fowlkes-Mallows (p-valor de 0.27) e, dessa forma, não há evidências de que o algoritmo proposto apresentou uma melhoria com relação ao método tradicional para essa medida de qualidade.

Com relação ao algoritmo HDBSCAN, os parâmetros padrões de sua implementação em Python da biblioteca scikit-learn foram utilizados, com exceção ao parâmetro relacionado ao tamanho mínimo dos grupos, que foi definido como 10 para todos os experimentos.

Tabela 4: Médias para cada métrica de avaliação dos agrupamentos após 30 execuções dos algoritmos k-médias tradicional e HDBSCAN em dados de alta densidade (grupo 2).

Número	HDBSCAN			K-médias topológico		
	Rand	AMI	Fowlkes-Mallows	Rand	AMI	Fowlkes-Mallows
1	0.789	**0.636**	0.422	**0.884**	0.543	**0.474**
2	0.192	0.064	**0.319**	**0.781**	**0.067**	0.156
3	0.803	0.570	0.413	**0.881**	**0.582**	**0.477**
4	0.404	0.265	0.462	**0.797**	**0.476**	**0.506**
5	0.341	0.002	0.509	**0.625**	**0.257**	**0.570**
6	0.645	0.087	**0.225**	**0.856**	**0.140**	0.122
7	0.502	0.372	0.300	**0.846**	**0.389**	**0.306**
8	0.206	0.038	**0.313**	**0.659**	**0.271**	0.290
9	0.830	0.574	0.450	**0.879**	**0.597**	**0.456**
10	0.595	0.317	0.334	**0.814**	**0.538**	**0.402**

Os resultados para a comparação de performance entre os algoritmos k-médias topológico e HDBSCAN em dados de alta densidade estão dispostos na Tabela 4. De acordo com o teste não paramétrico de Wilcoxon ao nível de significância de 5% para amostras pareadas em cada uma das métricas, há evidências de que o k-médias topológico obteve performances superiores ao HDBSCAN ao levar em consideração o índice de Rand (p-valor de 0.002) e AMI (p-valor de 0.03). Entretanto, não rejeitamos a hipótese nula em termos do índice de Fowlkes-Mallows (p-valor de 0.61) e, dessa forma, não há evidências de que o algoritmo proposto apresentou uma melhoria com relação ao método tradicional para essa medida de qualidade.

Os experimentos indicaram que a performance obtida no algoritmo proposto ficou próxima ou superior ao k-médias tradicional e ao HDBSCAN nos conjuntos de dados

selecionados. Por se tratar de uma métrica de avaliação mais robusta, não tivemos evidências de que o k-médias topológico obteve performances superiores aos outros dois métodos com base no índice de Fowlkes-Mallows. Além disso, vale ressaltar que, por se tratar de um algoritmo extremamente sensível à escolha dos parâmetros, o HDBSCAN poderia obter resultados muito melhores com um ajuste de seus hiperparâmetros a um maior custo computacional. Similarmente, o k-médias topológico se beneficiaria com melhores métodos de inicialização de seus centroides e métodos mais eficientes de seleção do parâmetro k na geração do grafo de k vizinhos mais próximos, além da utilização de uma medida diferente da distância euclidiana em sua construção.

4. CONCLUSÃO

Técnicas de agrupamento de dados, apesar de extremamente úteis na análise e reconhecimento de padrões em dados, possuem diversas limitações que ainda dificultam essa tarefa em dados reais, principalmente em cenários em que há uma escassez de observações e uma quantidade grande de parâmetros. Dados de alta dimensionalidade propiciam o fenômeno da "maldição da dimensionalidade" e fazem com que esses métodos produzam agrupamentos de menor qualidade, principalmente quando o algoritmo está fundamentado na utilização da distância euclidiana.

Com o intuito de minimizar essa limitação no método particional k-médias, a sua versão topológica proposta nesse projeto fez uso da distância geodésica calculada por meio de uma representação em grafo dos dados. Pelos resultados obtidos, observou-se que o algoritmo proposto foi capaz de mitigar a limitação imposta pela distância euclidiana que tornava o método incapaz de detectar agrupamentos não esféricos, além de melhorar sua performance em cenários de alta dimensionalidade. Dessa forma, o k-médias topológico se mostra como uma alternativa viável e eficiente em cenários que a distância euclidiana pode afetar negativamente a performance do método, equiparando-se a algoritmos modernos, como o HDBSCAN, que possuem maior complexidade computacional.

Apesar dos pontos positivos, o k-médias topológico ainda lida com algumas limitações relacionadas à escolha de inicialização de seus centroides e escolha do número de vizinhos na construção de sua representação em grafo dos dados. Como o grafo de k vizinhos mais próximos é construído com base na distância euclidiana entre os pontos e seus centroides, a utilização de uma outra medida de distância pode também ser interessante em certos casos para aumentar a performance do algoritmo na detecção de agrupamentos de formatos diferentes.

5. DIFICULDADES ENCONTRADAS

Como uma das suposições para a utilização do grafo de k vizinhos mais próximos é de que ele esteja totalmente conectado, isto é, não seja criado grafos desconexos, não foi possível utilizar $\log_2 n$ como uma estimativa para o número de vizinhos. Ao contrário da raiz quadrada do número de observações, essa estimativa produziu grafos desconexos para alguns experimentos em alta dimensão, apesar de ser um estimador ótimo na construção desses grafos [19]. Além disso, a dificuldade associada à escolha dos centroides iniciais poderia ser resolvida com a implementação da estratégia do k-médias++, em que os centroides são inicializados como uma das amostras de treinamento [20].

6. REFERÊNCIAS BIBLIOGRÁFICAS

[1] EZUGWU, A. E. et al. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. **Engineering Applications of Artificial Intelligence**, v. 110, p. 104743, abr. 2022.

[2] HARTIGAN, J. A.; WONG, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. **Applied Statistics**, v. 28, n. 1, p. 100, 1979.

[3] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The elements of statistical learning**: data mining, inference, and prediction. New York: Springer, 2004.

[4] SINAGA, K. P.; YANG, M.-S. Unsupervised K-Means Clustering Algorithm. **IEEE Access**, v. 8, p. 80716-80727, 2020.

[5] IKOTUN, A. M. et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. **Information Sciences,** v. 622, p. 178-210, dez. 2022.

[6] JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651-666, jun. 2010.

[7] MCINNES, L.; HEALY, J.; ASTELS, S. hdbscan: Hierarchical density based clustering. **The Journal of Open Source Software**, v. 2, n. 11, p. 205, 21 mar. 2017.

[8] STEWART , G.; AL-KHASSAWENEH, M. An Implementation of the HDBSCAN* Clustering Algorithm. **Applied Sciences**, v. 12, n. 5, p. 2405, 25 fev. 2022.

[9] BUSHRA, A. A.; YI, G. Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms. **IEEE Access**, v. 9, p. 87918-87935, 2021.

[10] STEINLEY D. K-means clustering: A half-century synthesis. **British Journal of Mathematical and Statistical Psychology**, v. 70, n. 1, p. 1-25, 2017.

[11] DONOHO, D. High-dimensional data analysis: The curses and blessings of dimensionality. **Proceedings of the AMS Conference on Math Challenges of the 21st Century**, p. 32-56, 2000.

[12] HOULE, M. E. et al. Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? **Lecture Notes in Computer Science**, p. 482-500, 1 jan. 2010.

[13] DONG, W.; MOSES, C.; LI, K. Efficient k-nearest neighbor graph construction for generic similarity measures. **Proceedings of the 20th international conference on World wide web - WWW '11**, 2011.

[14] FEFFERMAN, C.; MITTER, S.; NARAYANAN, H. Testing the Manifold **Hypothesis. Journal of the American Mathematical Society**, v. 29, p. 983-1049, 1 out. 2013.

[15] RAND, W. M. Objective Criteria for the Evaluation of Clustering Methods. **Journal of the American Statistical Association**, v. 66, n. 336, p. 846-850, dez. 1971.

[16] VINH, N. X.; EPPS, J.; BAILEY, J. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? **Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09**, 2009.

[17] FOWLKES, E. B.; MALLOWS, C. L. A Method for Comparing Two Hierarchical Clusterings. **Journal of the American Statistical Association**, v. 78, n. 383, p. 553-553, 1 set. 1983.

[18] WILCOXON, F. Individual Comparisons by Ranking Methods. **Biometrics Bulletin**, v. 1, n. 6, p. 80-83, 1945.

[19] MAIER, M.; HEIN, M.; VON LUXBURG, U. Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters. **Theoretical Computer Science**, v. 410, n. 19, p. 1749-1764, 28 abr. 2009.

[20] ARTHUR, D.; VASSILVITSKII, S. k-means++: the advantages of careful seeding. **Symposium on Discrete Algorithms**, p. 1027-1035, 7 jan. 2007.