Agrupamento de dados particional utilizando caminhos mínimos em grafos

Fernando Borges[1], **Alexandre Luís Magalhães Levada**[2]

[1]*Departamento de Estatística, Universidade Federal de São Carlos*
[2]*Departamento de Computação, Universidade Federal de São Carlos*

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. 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.

O algoritmo k-médias é um dos mais conhecidos e utilizados devido à sua simplicidade e eficiência computacional, pois realoca cada ponto ao agrupamento com centroide mais próximo e repete o processo até convergência. No entanto, sua versão tradicional enfrenta diversas limitações em dados de maior dimensionalidade por utilizar a distância euclidiana, que restringe sua capacidade de detectar agrupamentos não lineares [1, 2].

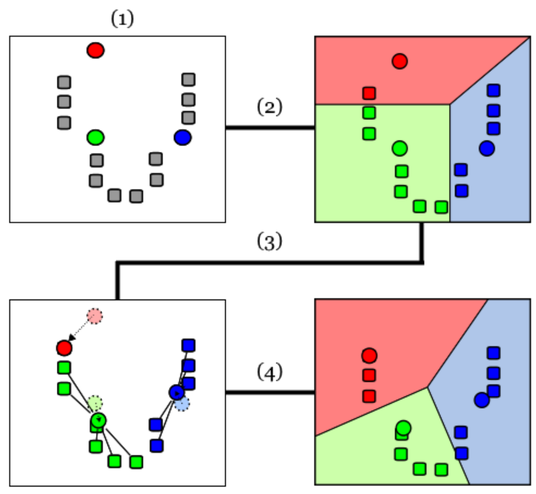

Figura 1: Representação de uma iteração do algoritmo k-médias.

A dificuldade em lidar com dados de alta dimensão e cenários em que há um número de parâmetros muito maior que o de observações presentes no conjunto de dados é um problema comum a esses métodos que recebe o nome de "maldição da dimensionalidade" [3].

Por meio de uma representação dos dados em um grafo de k vizinhos mais próximos, o algoritmo de Dijkstra pode ser utilizado para calcular as distâncias geodésicas entre as observações no algoritmo k-médias. 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 em cenários de alta dimensionalidade por diminuir a restrição de agrupamentos circularmente simétricos da distância euclidiana.

Objetivo

O objetivo desse projeto é implementar o algoritmo k-médias topológico, que substitui a distância euclidiana no k-médias por comprimentos de caminhos mínimos em grafos, e compará-lo a sua versão tradicional e ao algoritmo HDBSCAN, um método moderno capaz de detectar agrupamentos de qualquer formato na presença de ruído, com relação às qualidades de seus agrupamentos.

Materiais e métodos

A análise comparativa entre a performance do algoritmo k-médias topológico e a performance dos algoritmos k-médias tradicional e HDBSCAN foi realizada em dois grupos de experimentos contendo (i) dados de alta dimensionalidade e poucas observações e (ii) dados de alta densidade. Os dois grupos de experimentos utilizaram conjuntos de dados provenientes do repositório OpenML, que os disponibiliza gratuitamente para o desenvolvimento de pesquisas na área de ciência de dados.

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

Resultados e discussões

O primeiro grupo de experimentos utilizou conjuntos de dados de alta dimensionalidade com poucas observações com o intuito de comparar a performance dos algoritmos k-médias tradicional e topológico na influência da "maldição da dimensionalidade".

Tabela 1: Média das métricas de qualidade de agrupamento após 30 execuções no primeiro grupo.

	K-médias tradicional			K-médias topológico		
Conjunto de dados	**Rand**	**AMI**	**FM**	**Rand**	**AMI**	**FM**
AP_Colon_Kidney	0.574	0.133	0.582	**0.691**	**0.341**	**0.715**
AP_Prostate_Kidney	0.634	0.252	0.702	**0.678**	**0.306**	**0.756**
AP_Breast_Colon	0.509	0.011	0.514	**0.582**	**0.296**	**0.627**
tr12.wc	0.307	0.039	**0.347**	**0.510**	**0.154**	0.329
tr31.wc	0.366	0.078	**0.472**	**0.488**	**0.167**	0.398
tr45.wc	0.380	0.125	**0.315**	**0.548**	**0.195**	0.299
SRBCT	0.590	0.143	0.362	**0.669**	**0.283**	**0.436**
pasture	**0.711**	0.372	0.565	0.677	**0.380**	**0.569**
leukemia	0.556	0.108	0.610	**0.576**	**0.124**	**0.656**
GCM	0.871	**0.365**	**0.331**	**0.880**	0.332	0.301
Média	0.552	0.163	0.479	**0.628**	**0.236**	**0.508**

O segundo grupo de experimentos envolvou 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 possuem agrupamentos de alta densidade com um alto valor de observações e classes.

Tabela 2: Média das métricas de qualidade de agrupamento após 30 execuções no segundo grupo.

	HDBSCAN			K-médias topológico		
Conjunto de dados	**Rand**	**AMI**	**FM**	**Rand**	**AMI**	**FM**
digits	0.789	**0.636**	0.422	**0.884**	0.543	**0.474**
JapaneseVowels	0.192	0.064	**0.319**	**0.781**	**0.067**	0.156
optdigits	0.803	0.570	0.413	**0.881**	**0.582**	**0.477**
satimage	0.404	0.265	0.462	**0.797**	**0.476**	**0.506**
waveform-5000	0.341	0.002	0.509	**0.625**	**0.257**	**0.570**
abalone	0.645	0.087	**0.225**	**0.856**	**0.140**	0.122
semeion	0.502	0.372	0.300	**0.846**	**0.389**	**0.306**
cnae-9	0.206	0.038	**0.313**	**0.659**	**0.271**	0.290
mfeat-pixel	0.830	0.574	0.450	**0.879**	**0.597**	**0.456**
micro-mass	0.595	0.317	0.334	**0.814**	**0.538**	**0.402**
Média	0.531	0.293	0.375	**0.812**	**0.386**	**0.376**

De acordo com os testes de Wilcoxon pareados, ao nível de significância de 5%, há evidências de que o k-médias topológico obteve performances superiores ao k-médias tradicional e ao HDBSCAN em termos do índice de Rand e AMI. Vale ressaltar que o HDBSCAN poderia obter resultados 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 de seleção do parâmetro na geração do grafo de k vizinhos mais próximos.

Conclusões

Com o intuito de minimizar as limitações presentes no k-médias, a sua versão topológica 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. Além de se mostrar como uma alternativa viável e eficiente em cenários nos quais a distância euclidiana pode afetar negativamente a performance do método, o k-médias topológico se equipara a algoritmos modernos, como o HDBSCAN.

Referências

[1] 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.

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

[3] 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.