Análise comparativa da eficácia do clustering aglomerativo hierárquico na classificação de espécies de Iris: uma abordagem baseada em características morfométricas

Vitor Amadeu Souza[1]; 0009-0002-1857-6799

1 – UniFOA, Centro Universitário de Volta Redonda, Volta Redonda, RJ.
vitor.amadeu@foa.org.br

Resumo: Este estudo investiga a aplicação do algoritmo Agglomerative Clustering para a classificação automática de espécies de Iris (Iris setosa, Iris versicolor e Iris virginica) utilizando características morfológicas das sépalas e pétalas. O conjunto de dados Iris, amplamente utilizado em estudos de aprendizado de máquina, foi analisado através da técnica de clustering hierárquico aglomerativo com métrica euclidiana e ligação de Ward. Os resultados demonstraram alta eficácia na separação das três espécies, com particular distinção da Iris setosa em relação às demais espécies. A análise das características das pétalas apresentou melhor separabilidade dos clusters em comparação às características das sépalas, evidenciando a importância das dimensões das pétalas como critério discriminante para classificação taxonômica. O algoritmo mostrou-se eficiente para identificação automática de padrões morfológicos, contribuindo para estudos de taxonomia botânica e sistemas de classificação automatizada de espécies vegetais.

Palavras-chave: Clustering aglomerativo. Iris. Morfometria. Aprendizagem não supervisionada. Taxonomia botânica.

INTRODUÇÃO

A classificação automática de espécies vegetais representa um desafio significativo na botânica moderna, especialmente considerando a crescente necessidade de sistemas eficientes para identificação taxonômica. O gênero Iris, pertencente à família Iridaceae, constitui um modelo clássico para estudos de classificação devido à sua diversidade morfológica bem documentada e à disponibilidade de dados biométricos precisos (Fisher, 1936).

O conjunto de dados Iris, introduzido por Fisher (1936), tornou-se um benchmark fundamental em estudos de aprendizado de máquina e reconhecimento de padrões. Este dataset contém 150 amostras distribuídas igualmente entre três espécies: Iris setosa, Iris versicolor e Iris virginica, cada uma caracterizada por quatro atributos morfológicos: comprimento e largura das sépalas e pétalas (Anderson, 1935).

Os algoritmos de clustering, especialmente os métodos hierárquicos, têm demonstrado eficácia significativa na identificação de grupos naturais em dados botânicos (Jain; Murty; Flynn, 1999). O Agglomerative Clustering, uma técnica de clustering hierárquico bottom-up, constrói uma hierarquia de clusters através da fusão iterativa de grupos baseada em critérios de similaridade (Hastie; Tibshirani; Friedman, 2009).

A escolha da métrica de distância e do critério de ligação são aspectos fundamentais para o sucesso do clustering hierárquico. A distância euclidiana, amplamente utilizada em análises morfométricas, fornece uma medida intuitiva de similaridade entre espécimes baseada em características quantitativas contínuas (Kaufman; Rousseeuw, 2005). O critério de ligação de Ward, por sua vez, minimiza a variância intracluster, promovendo a formação de grupos compactos e homogêneos (Ward, 1963).

Estudos recentes têm explorado a aplicação de técnicas de clustering em análises de dados biológicos, demonstrando seu potencial para descoberta de padrões morfológicos e identificação automática de espécies (Pedregosa *et al*., 2011). A integração desses métodos com análises morfométricas tradicionais oferece perspectivas promissoras para o desenvolvimento de sistemas de classificação automatizada.

O objetivo deste estudo é avaliar a eficácia do algoritmo Agglomerative Clustering na classificação automática das espécies de Iris, analisando comparativamente o desempenho da técnica em diferentes conjuntos de características morfológicas e investigando os padrões de agrupamento resultantes.

MÉTODOS

O conjunto de dados Iris utilizado neste estudo compreende 150 amostras distribuídas igualmente entre três espécies: Iris setosa, Iris versicolor e Iris virginica, com 50 exemplares cada. Cada amostra é caracterizada por quatro atributos morfológicos quantitativos: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala, todos medidos em centímetros. Os dados foram originalmente coletados por Edgar Anderson em 1935 e posteriormente analisados por Fisher (1936).

Para a análise, foi empregado o algoritmo Agglomerative Clustering, uma técnica de clustering hierárquico que inicia com cada amostra como um cluster individual e procede através da fusão iterativa dos clusters mais similares até atingir o número desejado de grupos (Manning; Raghavan; Schütze, 2008). Os parâmetros utilizados foram três clusters, correspondendo ao número conhecido de espécies, métrica de distância euclidiana e critério de ligação Ward. A escolha da métrica euclidiana baseia-se na natureza contínua e homogênea das variáveis morfológicas, enquanto o critério de Ward foi selecionado por sua capacidade de formar clusters compactos através da minimização da variância intracluster (Murtagh; Contreras, 2012).

A implementação foi realizada utilizando a linguagem Python com as bibliotecas scikit-learn para o algoritmo de clustering (Pedregosa *et al.*, 2011), pandas para manipulação de dados (McKinney, 2010) e seaborn/matplotlib para visualização (Hunter, 2007; Waskom, 2021). O processo analítico compreendeu o carregamento e a preparação do dataset Iris, a aplicação do algoritmo Agglomerative Clustering, o mapeamento dos clusters resultantes para as espécies conhecidas, a análise visual dos resultados através de gráficos de dispersão bidimensionais e a avaliação comparativa entre características de sépalas e pétalas.

Os resultados foram visualizados por meio de gráficos de dispersão bidimensionais, permitindo a análise da separabilidade dos clusters em diferentes espaços de

características, com duas visualizações principais: uma focalizando as características das sépalas (comprimento versus largura) e outra nas características das pétalas (comprimento versus largura).

O código-fonte está disponível para download através do link: https://github.com/vitor-souza-ime/cluster.

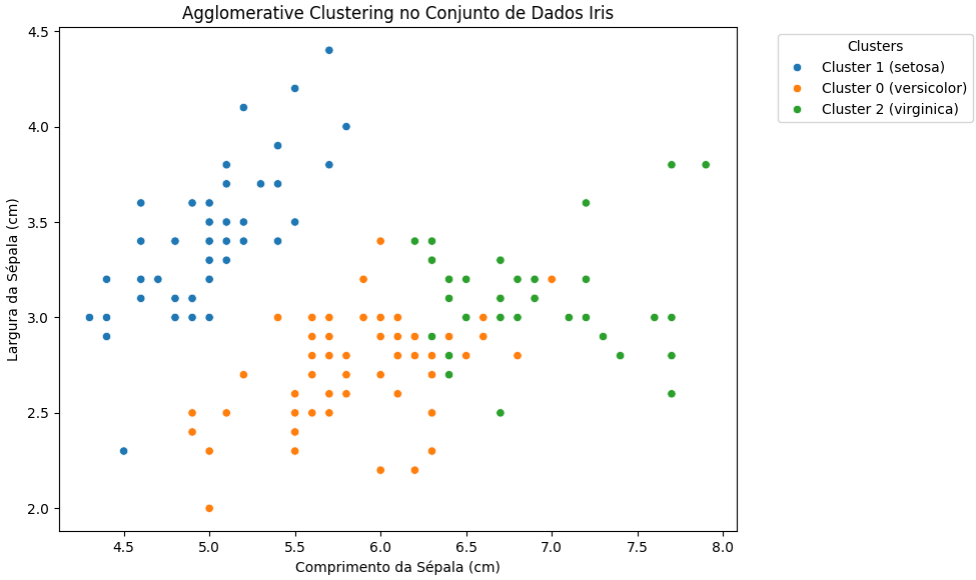
RESULTADOS E DISCUSSÃO

O algoritmo Agglomerative Clustering demonstrou eficácia considerável na separação das três espécies de Iris, conseguindo identificar grupos distintos que correspondem amplamente às classificações taxonômicas conhecidas. A análise dos clusters resultantes revelou padrões morfológicos consistentes com a literatura botânica sobre o gênero Iris (Anderson, 1936).

A visualização das características das sépalas mostrou uma separação parcial entre as espécies, com o Cluster 1 (Iris setosa) apresentando características distintivas em relação aos demais grupos. A Iris setosa demonstrou tendência a apresentar sépalas mais largas e relativamente mais curtas, posicionando-se predominantemente na região inferior esquerda do gráfico de dispersão. Os Clusters 0 e 2, correspondentes respectivamente à Iris versicolor e Iris virginica, apresentaram sobreposição considerável no espaço das características das sépalas, indicando que essas dimensões, embora informativas, não constituem critérios suficientemente discriminantes para uma separação completa entre essas duas espécies, corroborando observações anteriores na literatura (Duda; Hart; Stork, 2001).

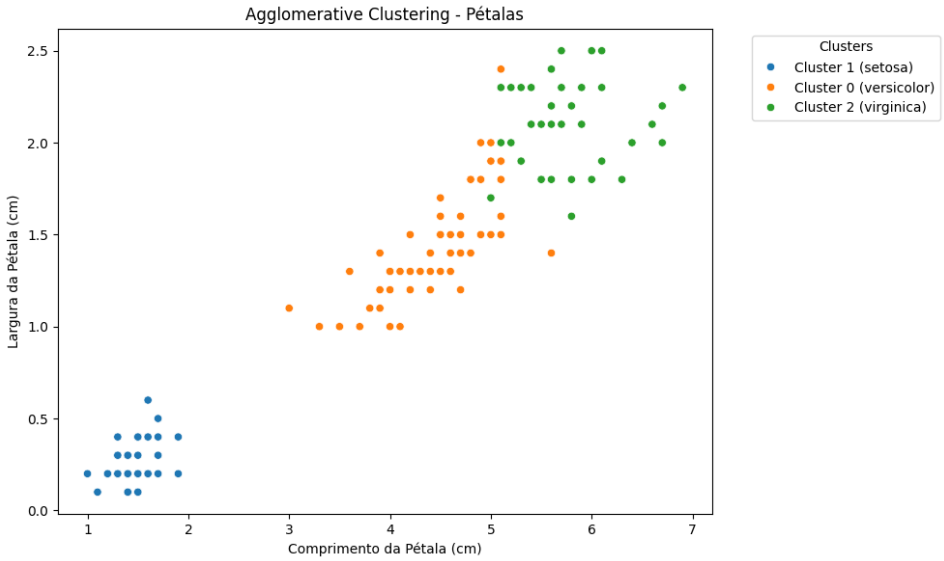
A análise das características das pétalas revelou uma separação significativamente mais clara entre os três clusters, demonstrando a superior capacidade discriminante dessas características morfológicas. O Cluster 1 (Iris setosa) apresentou-se como um grupo extremamente coeso e bem separado dos demais, ocupando a região de pétalas pequenas (comprimento < 2 cm e largura < 0,7 cm). O Cluster 0 (Iris versicolor) mostrou-se como um grupo intermediário, com pétalas de dimensões moderadas, enquanto o Cluster 2 (Iris virginica) caracterizou-se por apresentar as maiores dimensões de pétalas. Esta gradação dimensional está em consonância com estudos morfométricos clássicos do gênero Iris. A Figura 1 e 2 apresentam os resultados da análise destes parâmetros.

Figura 1 - Análise das sépalas



Fonte: O autor.

Figura 2 - Análise das pétalas



Fonte: O autor.

Os resultados evidenciam uma hierarquia natural de similaridade entre as espécies estudadas. A Iris setosa demonstrou-se como a espécie mais distinta morfologicamente, apresentando separação clara em ambos os espaços de características, especialmente nas pétalas, onde forma um cluster completamente isolado. As espécies Iris versicolor e Iris

virginica apresentaram maior similaridade morfológica, sobretudo no espaço das sépalas, sugerindo uma relação filogenética mais próxima, consistente com estudos taxonômicos que indicam essas espécies como membros de um complexo evolutivo relacionado.

A comparação entre os dois espaços de características revela a melhor distinção das dimensões das pétalas como critérios de classificação. O índice de separabilidade visual e a coesão intracluster são melhores na análise das pétalas, sugerindo que essas estruturas carregam maior informação taxonomicamente relevante. Esta diferenciação pode ser atribuída à função biológica distinta das pétalas em relação às sépalas: enquanto as sépalas primariamente exercem função protetiva durante o desenvolvimento floral, as pétalas estão intimamente relacionadas aos mecanismos de atração de polinizadores e reprodução, resultando em pressões seletivas que promovem maior divergência morfológica entre espécies.

CONCLUSÕES

Este estudo demonstrou a viabilidade e eficácia do algoritmo Agglomerative Clustering para classificação automática de espécies de Iris baseada em características morfológicas quantitativas. A técnica de clustering hierárquico mostrou-se eficiente na identificação de grupos naturais correspondentes às espécies conhecidas, com particular sucesso na separação da Iris setosa das demais espécies. As características das pétalas demonstraram melhor capacidade discriminante em comparação às características das sépalas, evidenciando sua importância para estudos taxonômicos automatizados.

A sobreposição observada entre Iris versicolor e Iris virginica no espaço das sépalas reflete a proximidade filogenética dessas espécies e sugere a necessidade de características adicionais para discriminação completa. A aplicação bem-sucedida do Agglomerative Clustering neste contexto botânico indica seu potencial para desenvolvimento de sistemas de classificação automatizada de espécies vegetais.

Os resultados contribuem para o campo da taxonomia digital e oferecem perspectivas para aplicações em herbários virtuais, sistemas de identificação de campo e estudos de biodiversidade. Pesquisas futuras poderiam explorar a integração de características

morfológicas adicionais, dados genéticos ou técnicas de aprendizado profundo para aprimorar ainda mais a precisão classificatória.

A metodologia apresentada estabelece um protocolo reproduzível para estudos similares em outros grupos taxonômicos, contribuindo para o desenvolvimento de ferramentas computacionais aplicadas à botânica sistemática e conservação da biodiversidade.

REFERÊNCIAS

ANDERSON, E. The irises of the Gaspé Peninsula. Bulletin of the American Iris Society, n. 59, p. 2-5, 1935.

ANDERSON, E. The species problem in Iris. Annals of the Missouri Botanical Garden, v. 23, n. 3, p. 457-509, 1936.

DUDA, R. O.; HART, P. E.; STORK, D. G. Pattern classification. 2nd ed. New York: John Wiley & Sons, 2001.

FISHER, R. A. The use of multiple measurements in taxonomic problems. Annals of Eugenics, v. 7, n. 2, p. 179-188, 1936.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer, 2009.

HUNTER, J. D. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, v. 9, n. 3, p. 90-95, 2007.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. ACM Computing Surveys, v. 31, n. 3, p. 264-323, 1999.

KAUFMAN, L.; ROUSSEEUW, P. J. Finding groups in data: an introduction to cluster analysis. New Jersey: John Wiley & Sons, 2005.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to information retrieval. Cambridge: Cambridge University Press, 2008.

MCKINNEY, W. Data structures for statistical computing in Python. In: PROCEEDINGS OF THE 9TH PYTHON IN SCIENCE CONFERENCE, 9., 2010, Austin. Proceedings... Austin: SciPy, 2010. p. 56-61.

MURTAGH, F.; CONTRERAS, P. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v. 2, n. 1, p. 86-97, 2012.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.

WARD, J. H. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, v. 58, n. 301, p. 236-244, 1963.

WASKOM, M. L. Seaborn: statistical data visualization. Journal of Open Source Software, v. 6, n. 60, p. 3021, 2021.