Análise Exploratória de Dados Astronômicos Utilizando Técnicas de Agrupamento Não Supervisionado

Samara Paloma Ribeiro[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Ouro Preto (UFOP)
Ouro Preto – MG – Brazil

samara.augusto@aluno.ufop.edu.br

Resumo. *Este trabalho apresenta uma análise exploratória de dados astronômicos utilizando técnicas de aprendizado não supervisionado, com o objetivo de investigar se algoritmos de agrupamento são capazes de identificar padrões naturais associados às classes espectrais de estrelas. O estudo foi conduzido sobre um conjunto de dados sintético contendo 1001 estrelas, descritas por atributos físicos como temperatura, luminosidade, raio e distância.*
Foram aplicados três algoritmos de clustering: Agrupamento Hierárquico, DBSCAN e Gaussian Mixture Model (GMM), selecionados por sua relevância em problemas de descoberta de padrões e por sua capacidade de modelar distribuições complexas. A avaliação dos agrupamentos foi realizada por meio de métricas internas de qualidade, como o coeficiente de silhueta, além da comparação dos clusters obtidos com as classes espectrais conhecidas.
Os resultados indicam que o Agrupamento Hierárquico e o GMM apresentaram melhor correspondência com as classes espectrais, especialmente na separação de estrelas quentes e frias, enquanto o DBSCAN mostrou sensibilidade à escolha de parâmetros e à densidade dos dados, identificando outliers relevantes. Conclui-se que métodos de aprendizado não supervisionado são ferramentas eficazes para a análise exploratória e a organização de dados astronômicos, sendo especialmente úteis em cenários onde rótulos confiáveis não estão disponíveis.

1. Introdução

A astronomia moderna produz volumes massivos de dados, provenientes de levantamentos celestes como o Sloan Digital Sky Survey (SDSS) e o Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST) [Yang et al. 2022]. Estes dados incluem informações espectroscópicas e fotométricas de milhões de objetos celestes, como estrelas, galáxias e quasares. A análise manual destes dados é inviável, tornando necessário o uso de técnicas de mineração de dados e aprendizado de máquina para extrair conhecimento significativo.

O aprendizado não supervisionado, em particular algoritmos de agrupamento (*clustering*), permite identificar estruturas naturais nos dados sem a necessidade de rótulos previamente conhecidos [Fotopoulou 2024]. Na astronomia, esses métodos são usados para:

- Identificar classes de estrelas com propriedades semelhantes, baseadas em temperatura, luminosidade, raio e abundância química;

- Detectar aglomerados estelares, nuvens moleculares e subestruturas de galáxias;
- Explorar relações complexas em dados de alta dimensionalidade, facilitando a visualização e interpretação científica [Yu and Hou 2022, Hunt and Reffert 2020].

Este trabalho foca em um conjunto sintético de 1001 estrelas [Contributors], com atributos como temperatura, luminosidade, raio, massa, magnitude absoluta, distância e classe espectral (B, A, F, G, K, M). O objetivo é investigar se algoritmos de clustering são capazes de identificar padrões correspondentes às classes espectrais, demonstrando a aplicabilidade de aprendizado não supervisionado em dados astronômicos.

2. Descrição do Problema

A astronomia moderna gera enormes volumes de dados observacionais e simulados, provenientes de levantamentos celestes como o Sloan Digital Sky Survey (SDSS) e o Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST) [Yang et al. 2022]. Esses dados contêm informações detalhadas sobre milhões de objetos celestes, incluindo estrelas, galáxias e quasares, com múltiplas características físicas e espectrais. A análise manual desses conjuntos é inviável, exigindo técnicas computacionais para explorar padrões e estruturas escondidas nos dados.

Um dos desafios centrais na astronomia é a classificação de estrelas em classes espectrais, que refletem sua temperatura, cor, composição química e estágio evolutivo. Tradicionalmente, essas classes são definidas por especialistas a partir de espectros ou propriedades fotométricas. Entretanto, com o crescimento do volume de dados, surge a necessidade de métodos automáticos capazes de identificar padrões naturais sem depender de rótulos pré-existentes.

Neste trabalho, o problema consiste em investigar se **algoritmos de aprendizado não supervisionado** podem identificar agrupamentos naturais de estrelas com base em características físicas como temperatura, luminosidade, raio, massa, magnitude absoluta e distância. O objetivo é determinar se os clusters encontrados correspondem, ou se aproximam, das classes espectrais conhecidas (B, A, F, G, K, M), permitindo validar a eficácia de diferentes métodos de clustering na análise exploratória de dados astronômicos.

O conjunto de dados utilizado é sintético, contendo informações de 1001 estrelas [Contributors], projetado para simular propriedades estelares reais. Este cenário permite experimentar com técnicas de mineração de dados sem limitações de disponibilidade ou qualidade de observações reais, ao mesmo tempo em que fornece um ambiente controlado para comparação entre algoritmos.

Os desafios específicos deste problema incluem:

- **Dados de alta dimensionalidade:** Cada estrela é representada por múltiplos atributos contínuos e categóricos, exigindo técnicas que consigam lidar com diferentes escalas e relações não lineares.
- **Clusters de formas irregulares:** As classes espectrais podem não estar distribuídas de forma linear ou esférica nos atributos disponíveis.
- **Ausência de rótulos supervisionados:** O aprendizado não supervisionado deve inferir agrupamentos a partir da estrutura intrínseca dos dados, tornando a escolha do algoritmo e de seus parâmetros críticos para a qualidade do resultado.

Portanto, o problema proposto consiste em aplicar algoritmos de clustering para explorar, analisar e validar a estrutura subjacente dos dados estelares, avaliando se os padrões identificados correspondem às classes espectrais conhecidas e fornecendo insights sobre a distribuição de propriedades físicas das estrelas.

3. Técnicas de Inteligência Artificial e Justificativa

Para abordar o problema de identificação de padrões em dados estelares, foram selecionados três algoritmos de agrupamento não supervisionado, escolhidos com base em sua aplicação comprovada em astronomia e nas características do conjunto de dados:

1. **Agrupamento Hierárquico (Hierarchical Clustering):**
   Este algoritmo constrói uma estrutura hierárquica entre os dados por meio de fusões sucessivas, representadas em um dendrograma. Essa abordagem permite analisar relações entre objetos em diferentes níveis de similaridade, sendo especialmente útil para investigar a organização hierárquica de sistemas estelares e populações astronômicas. Além disso, o método não requer a definição prévia do número de clusters, o que favorece análises exploratórias [Yu and Hou 2022].

2. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**
   O DBSCAN identifica agrupamentos com base na densidade dos dados, sendo capaz de detectar clusters de formato arbitrário e separar pontos considerados ruído. Essa característica é particularmente relevante em astronomia, onde os dados frequentemente apresentam regiões densas intercaladas com áreas esparsas. O algoritmo também dispensa a definição prévia do número de clusters, favorecendo a descoberta de padrões naturais nos dados [Yang et al. 2022, Hunt and Reffert 2020].

3. **Gaussian Mixture Model (GMM):**
   O GMM é um modelo probabilístico que assume que os dados são gerados a partir de uma combinação de distribuições gaussianas multivariadas. Cada componente do modelo representa um cluster, permitindo capturar estruturas mais complexas e sobrepostas nos dados. Em astronomia, essa abordagem é adequada para modelar transições suaves entre classes espectrais e lidar com incertezas inerentes às observações.

A utilização conjunta desses algoritmos permite uma análise complementar dos dados estelares: o Agrupamento Hierárquico revela relações estruturais globais, o DBSCAN destaca regiões densas e identifica ruídos, enquanto o GMM modela a distribuição probabilística dos grupos. Essa diversidade metodológica contribui para uma avaliação mais robusta dos padrões presentes no conjunto de dados analisado.

4. Implementação

O ambiente utilizado nesta análise foi **Python 3.12.3**, com bibliotecas especializadas em ciência de dados e aprendizado de máquina, como `pandas`, `numpy`, `scikit-learn`, `matplotlib` e `seaborn`. O dataset sintético contém informações físicas e espectrais de 1001 estrelas, incluindo temperatura, luminosidade, raio, distância e classe espectral (B, A, F, G, K, M).

4.1. Análise Exploratória

Antes do pré-processamento, realizamos uma análise exploratória para compreender a distribuição dos dados:

- **Visualização inicial:** inspeção das primeiras linhas e estatísticas descritivas.
- **Criação da classe espectral principal:** extração da primeira letra da coluna `Spectral Class`.
- **Distribuição das classes:** gráficos de barras mostraram a frequência de cada classe espectral.

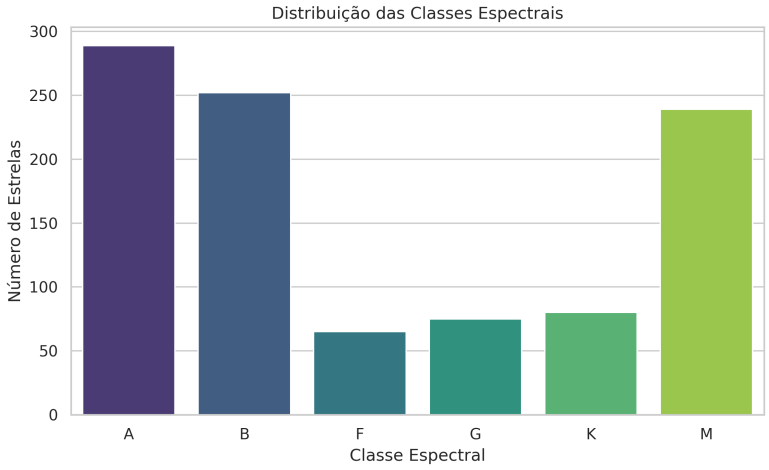

Figura 1. Distribuição das classes espectrais no conjunto de dados.

- **Diagrama HR:** scatter plot de Temperatura vs Luminosidade com eixo invertido para temperatura e escala logarítmica para luminosidade.

4.2. Pré-processamento

O pré-processamento preparou as variáveis numéricas para os algoritmos de clustering:

1. Carregamento do dataset e inspeção inicial.
2. Seleção das features numéricas: `Temperature (K)`, `Luminosity (L/Lo)`, `Radius (R/Ro)`, `Distance (ly)`.
3. Tratamento de valores inválidos: remoção de nulos e infinitos.
4. Normalização das variáveis com `StandardScaler` do `scikit-learn`.
5. Separação da classe espectral para análise comparativa posterior.
6. Salvamento dos dados processados em arquivos `.csv` para reuso.

4.3. Algoritmos de Clustering

Três técnicas foram aplicadas, cada uma com suas particularidades:

- **Hierarchical Clustering:** `AgglomerativeClustering` com `linkage='ward'` e 7 clusters.
- **DBSCAN:** `eps=1.2`, `min_samples=10`, identifica clusters densos e outliers.
- **Gaussian Mixture Model (GMM):** 7 componentes gaussianas, `covariance_type='full'`, modelo probabilístico.

4.4. Pipeline de Implementação

O fluxo do pipeline seguiu estas etapas:

1. **Carregamento e inspeção dos dados:** análise visual e estatística.
2. **Pré-processamento:** tratamento de valores inválidos, normalização e criação da coluna auxiliar de classe espectral.
3. **Aplicação dos algoritmos de clustering:** geração de clusters com Hierarchical, DBSCAN e GMM.
4. **Avaliação de desempenho:** cálculo do *Silhouette Score* e comparação dos clusters com classes espectrais conhecidas usando `crosstab`.
5. **Visualização:** redução de dimensionalidade via PCA, scatter plots, dendrogramas e gráficos de densidade/probabilidade.
6. **Salvamento de resultados:** clusters, métricas e visualizações foram exportados para arquivos `.csv` e gráficos.

4.5. Observações Técnicas

- A normalização é essencial para algoritmos baseados em distância.
- DBSCAN pode gerar outliers, exigindo cuidado na interpretação de métricas.
- GMM fornece probabilidades de pertencimento, permitindo analisar regiões de transição entre clusters.
- PCA permite visualização em 2D mantendo a interpretabilidade dos clusters.

5. Resultados Obtidos e Conclusão

A análise dos dados astronômicos por meio de técnicas de aprendizado não supervisionado evidenciou que os algoritmos Hierarchical Clustering, DBSCAN e Gaussian Mixture Model (GMM) são capazes de identificar padrões relevantes associados às propriedades físicas e às classes espectrais das estrelas. A avaliação do desempenho foi realizada combinando métricas internas de qualidade de cluster, como o *Silhouette Score*, e métricas externas derivadas da comparação entre os clusters obtidos e as classes espectrais reais, incluindo precisão, sensibilidade (*recall*) e F1-score.

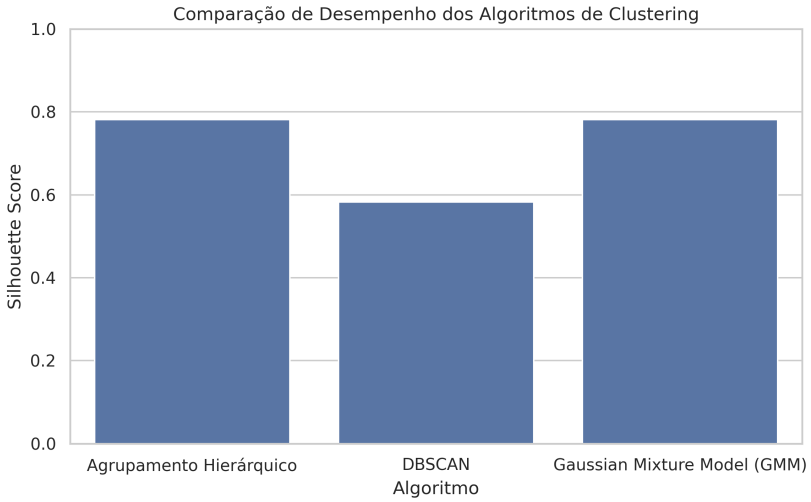

Figura 2. Comparação do desempenho dos algoritmos de clustering utilizando o *Silhouette Score*. Observa-se que o Hierarchical Clustering e o Gaussian Mixture Model apresentaram maior separação e coesão dos clusters, enquanto o DBSCAN obteve desempenho inferior devido à presença de ruídos e regiões de baixa densidade.

5.1. Hierarchical Clustering

O Agrupamento Hierárquico apresentou clusters bem definidos, especialmente nas classes espectrais extremas (B e M). O dendrograma permitiu observar a hierarquia natural das estrelas em subgrupos, refletindo similaridades físicas como temperatura e luminosidade. O valor do *Silhouette Score* igual a 0.781 indicou alta coesão intra-cluster e boa separação inter-cluster.

A Tabela 1 mostra a correspondência entre os clusters obtidos e as classes espectrais reais, evidenciando forte concentração de determinadas classes em clusters específicos, o que reforça a capacidade do método em capturar estruturas globais dos dados.

Tabela 1. Comparação entre clusters do Agrupamento Hierárquico e classes espectrais reais

Cluster	A	B	F	G	K	M
0	264	55	65	75	80	132
1	0	144	0	0	0	0
2	0	0	0	0	0	63
3	0	36	0	0	0	0
4	0	0	0	0	0	44
5	25	0	0	0	0	0
6	0	17	0	0	0	0

Além disso, o algoritmo apresentou precisão média ponderada de 0.611, sensibilidade de 0.593 e F1-score de 0.539, indicando um equilíbrio consistente entre a identificação correta das classes e a redução de erros de classificação.

5.2. DBSCAN

O algoritmo DBSCAN identificou clusters baseados na densidade dos dados, destacando regiões densas do espaço de características e isolando padrões menos frequentes. Conforme observado na Tabela 2, há uma forte predominância da classe espectral M em múltiplos clusters, além de uma concentração significativa das classes A e B em regiões densas.

Apesar de sua eficiência na detecção de outliers, o DBSCAN apresentou menor desempenho global, com *Silhouette Score* de 0.583. As métricas externas refletem essa limitação, com precisão média ponderada de 0.590, sensibilidade de 0.449 e F1-score de 0.383, indicando maior fragmentação dos clusters e sobreposição entre classes.

Tabela 2. Distribuição das classes espectrais por cluster obtido pelo DBSCAN

Cluster	A	B	F	G	K	M
0	264	199	65	75	80	132
1	25	0	0	0	0	0
2	0	0	0	0	0	44
3	0	0	0	0	0	37
4	0	0	0	0	0	26
5	0	36	0	0	0	0
6	0	17	0	0	0	0

5.3. Gaussian Mixture Model (GMM)

O Gaussian Mixture Model proporcionou uma segmentação probabilística das estrelas, capturando de forma eficiente a sobreposição natural entre classes espectrais. Os clusters obtidos refletem tanto agrupamentos bem definidos quanto regiões de transição, especialmente entre classes intermediárias como F, G e K.

O GMM apresentou desempenho semelhante ao Agrupamento Hierárquico, com *Silhouette Score* de 0.781. As métricas externas também indicam bom desempenho global, com precisão média ponderada de 0.611, sensibilidade de 0.593 e F1-score de 0.539, reforçando sua capacidade de modelar distribuições complexas e transições suaves entre classes.

Tabela 3. Distribuição das classes espectrais por cluster obtido pelo GMM

Cluster	A	B	F	G	K	M
0	264	55	65	75	80	132
1	0	0	0	0	0	63
2	0	36	0	0	0	0
3	25	0	0	0	0	0
4	0	0	0	0	0	44
5	0	144	0	0	0	0
6	0	17	0	0	0	0

5.4. Comparação Global das Métricas

A Tabela 4 apresenta um resumo das métricas de avaliação externas para os três algoritmos, permitindo uma comparação direta entre precisão, sensibilidade e F1-score.

Tabela 4. Resumo das métricas de avaliação dos algoritmos de clustering

Algoritmo	Precisão	Sensibilidade	F1-score
Hierarchical Clustering	0.611	0.593	0.539
DBSCAN	0.590	0.449	0.383
Gaussian Mixture Model	0.611	0.593	0.539

5.5. Conclusão Final

Os resultados confirmam que o aprendizado não supervisionado é uma abordagem eficaz para a identificação de padrões naturais em dados astronômicos. A análise demonstrou que o Agrupamento Hierárquico e o Gaussian Mixture Model apresentaram melhor desempenho global, tanto em métricas internas quanto externas, sendo particularmente adequados para dados com estrutura hierárquica ou distribuições probabilísticas complexas. Por outro lado, o DBSCAN mostrou-se mais sensível à densidade local dos dados, destacando-se na identificação de outliers, porém com menor capacidade de separação consistente entre classes espectrais. A combinação dessas técnicas fornece uma visão mais robusta e abrangente da distribuição estelar, reforçando a importância do uso de múltiplos métodos não supervisionados na análise de grandes volumes de dados astronômicos.

Referências

Contributors, K. Stars dataset. `https://https://www.kaggle.com/datasets/waqi786/stars-dataset?resource=download`.

Fotopoulou, S. (2024). A review of unsupervised learning in astronomy. *arXiv preprint arXiv:2406.17316 [astro-ph.IM]*.

Hunt, E. L. and Reffert, S. (2020). Improving the open cluster census. i. comparison of clustering algorithms applied to gaia dr2 data. *arXiv preprint arXiv:2012.04267 [astro-ph.GA]*.

Yang, H., Shi, C., Cai, J., Zhou, L., Yang, Y., Zhao, X., He, Y., and Hao, J. (2022). Data mining techniques on astronomical spectra data. i: Clustering analysis. *arXiv preprint arXiv:2212.08419 [astro-ph.IM]*.

Yu, H. and Hou, X. (2022). Hierarchical clustering in astronomy. *arXiv preprint arXiv:2211.06002 [astro-ph.IM]*.