XXXIII Congresso de Iniciação Científica da Unicamp

Estudo Exploratório de Algoritmos de Máquina para Classificação de Galáxias

Palavras-Chave: Classificação de Galáxias, Redes Neurais Convolucionais, Aprendizado Profundo

Autores(as):

Leonardo Basset Figueiredo Pereira, FT – UNICAMP

Prof. Dr. André Leon S. Gradvohl (orientador), FT – UNICAMP

INTRODUÇÃO:

O estudo das galáxias e suas estruturas é fundamental para compreender como a matéria se organiza em escalas cosmológicas. A classificação morfológica dessas galáxias possibilita a obtenção de informações importantes sobre sua evolução histórica e potencial para formar novas estrelas (NASA, 2024). Tradicionalmente, essa classificação era feita por meio de inspeção visual, empregando telescópios ópticos por astrônomos experientes. A seguir, são apresentados três principais tipos de galáxias ilustradas nas figuras 1, 2 e 3:

- **Galáxias Espirais:** são as mais comuns do universo, representando aproximadamente 66% de todas as galáxias conhecidas. Elas possuem formato de disco, com longos “braços” que se estendem a partir de uma região central brilhante chamada núcleo. Quando esses braços partem diretamente do núcleo, são classificados como espirais normais. Já quando os braços se originam de uma estrutura alongada no centro, são denominadas espirais barradas. Esse tipo de galáxia é caracterizado por uma intensa atividade de formação estelar em seu interior.
- **Galáxias Elípticas:** possuem uma forma circular e achatada, com aparência mais uniforme. Em comparação com outros tipos, apresentam menor quantidade de poeira e gás, o que limita a formação de novas estrelas. Como resultado, são compostas predominantemente por estrelas antigas e têm baixa atividade de formação estelar.
- **Galáxias Irregulares:** essas galáxias não apresentam um formato bem definido, possuindo estruturas assimétricas e desorganizadas. São consideradas formações muito antigas, surgidas antes dos tipos elípticos e espirais. Sua morfologia irregular geralmente é resultado da interação gravitacional com galáxias vizinhas, que distorcem sua estrutura original.

Figura 1: Galáxia Espiral.
Fonte: (National Geographic, 2024)

Figura 2: Galáxia Elíptica.
Fonte: (Lilith - UFMG, 2018)

Figura 3: Galáxia Irregular.
Fonte: (Space Today, 2018)

No entanto, com o avanço tecnológico e o crescente volume de dados obtidos, essa abordagem tornou-se impraticável, impulsionando a utilização de algoritmos de aprendizado de máquina, especialmente de aprendizado profundo (El Bouchefry & Souza, 2020).

O principal objetivo deste projeto foi identificar quais os problemas relativos à classificação de galáxias e os melhores algoritmos de aprendizado de máquina para realizar as tarefas de classificação para os problemas identificados. Para atingir esse objetivo, foram definidos os seguintes objetivos específicos:

- Levantar o estado da arte e identificar os potenciais problemas e soluções para a classificação de galáxias.
- Pesquisar a respeito dos melhores algoritmos de aprendizado de máquina utilizados atualmente para realizar a classificação de galáxias.
- Identificar as principais bases de dados com informações sobre a classificação de galáxias. Essas bases podem ser compostas por dados brutos (valores numéricos obtidos por radiotelescópios ou sensores embarcados em satélites) ou imagens.
- Implementar e testar alguns algoritmos de aprendizado de máquina mais promissores para o problema de classificação de galáxias.

- Propor a utilização de métricas para a análise de qualidade dos algoritmos de aprendizado de máquina implementados.

METODOLOGIA:

Inicialmente, foi realizada uma revisão bibliográfica abrangente sobre pesquisas existentes que abordam a classificação de galáxias e técnicas de aprendizado de máquina. Foram identificados e analisados artigos importantes que utilizavam diversas abordagens e técnicas, incluindo métodos tradicionais e mais recentes. Destacaram-se artigos como *"Galaxy Classification: a deep learning approach for classifying Sloan Digital Sky Survey images"* (Gharat & Dandawate, 2022), que propõe métodos baseados em aprendizado profundo para imagens obtidas pelo Sloan Digital Sky Survey, e *"Applying a Deep Learning Approach to Galaxy Classification with Galaxy Zoo"* (McRobie, 2023), que apresentou uma aplicação prática de redes neurais convolucionais (*CNNs*) utilizando a base Galaxy Zoo.

A etapa seguinte envolveu a implementação prática dos algoritmos selecionados, como as Redes Neurais Convolucionais (*CNNs*). Árvores de Decisão também foram testadas devido à sua capacidade interpretativa e facilidade em lidar com dados categóricos. Além disso, algoritmos de agrupamento baseados em densidade, como DBSCAN e HDBSCAN, foram incluídos para verificar sua eficácia na classificação de dados com ruído e distribuição irregular (Tramacere et al., 2016; Raja, 2024). Foi explorado também o *Extreme Gradient Boosting* (XGBoost), um algoritmo avançado que utilizava árvores de decisão com *boosting* de gradiente, reconhecido por sua eficiência computacional, boa capacidade de generalização e opções avançadas de regularização. (IBM, 2024).

Os algoritmos foram implementados utilizando a linguagem Python. As bibliotecas principais utilizadas incluem *pandas* e *numpy* para manipulação e pré-processamento dos dados, *tensorflow* e *keras* para construção e treinamento das *CNNs*, *sklearn* para implementação das Árvores de Decisão e algoritmos de *clustering*, e *matplotlib* para visualização dos resultados.

Na seleção da base de dados, optou-se inicialmente pela base Galaxy Zoo devido à sua popularidade e relevância na área de classificação morfológica. Contudo, dificuldades de acesso à versão atualizada levaram ao uso da versão de 2013, que já continha uma classificação detalhada em 37 classes. Na Figura 4 está ilustrada a separação das classes morfológicas. Tentativas foram feitas para reduzir a complexidade da base para sete classes principais, visando simplificar o problema. No entanto, esta redução provocou dificuldades adicionais na correta classificação das galáxias devido à perda de informações importantes das subclasses originais. Portanto, a decisão final foi manter as 37 classes iniciais para preservar a riqueza dos dados e garantir maior precisão analítica.

RESULTADOS E DISCUSSÃO:

Resultados do modelo utilizando apenas as *CNNs* apresentaram acurácia de 35%, precisão de 52% e AUC média de 85%, indicando claramente a presença de *overfitting*, fenômeno onde o modelo aprendeu excessivamente os dados de treinamento, prejudicando a generalização para novos dados (IBM, 2021). Para solucionar este problema, foram aplicados algoritmos de agrupamento DBSCAN e HDBSCAN (Tramacere, 2016; Raja 2024). Essas técnicas melhoraram moderadamente os resultados, elevando a acurácia para 45%, a precisão para 62% e mantendo a AUC estável em 84%, o que indicou que, apesar da melhora, o problema de *overfitting* persistia.

Diante desses problemas, foi realizada uma substituição da métrica de avaliação, passando-se da acurácia para *True Skill Statistic* (TSS), uma métrica mais robusta para dados desbalanceados (Yoon & Lee, 2023). O uso do TSS revelou um desempenho médio de apenas 35%, destacando a baixa capacidade de generalização do modelo treinado, especialmente em relação às classes minoritárias.

Na tentativa de aumentar o desempenho dos algoritmos, foi realizada uma combinação de *CNNs* com Árvores de Decisão conforme observado no trabalho de Dieleman et al. (2015). Contudo, os resultados não apresentaram melhorias significativas, obtendo-se TSS de 35% e AUC reduzida para 54%, indicando ainda maior dificuldade em classificar corretamente as galáxias. Técnicas adicionais como *early stopping*, método que interrompe o treinamento para evitar *overfitting* (Cyborg Codes, 2024), também não trouxeram resultados expressivos.

Esses experimentos sugeriram fortemente que o modelo não conseguiu capturar adequadamente as complexidades das classes minoritárias nem mitigar o *overfitting*, exigindo, portanto, uma abordagem ainda mais rigorosa e ajustes mais detalhados nos algoritmos e técnicas aplicadas.

CONCLUSÕES:

Os resultados obtidos neste estudo destacaram a complexidade inerente ao problema da classificação morfológica das galáxias. Um dos maiores desafios identificados foi o desbalanceamento entre classes, o que limitou significativamente a capacidade de generalização dos modelos implementados. A tentativa inicial de simplificação do conjunto de classes de 37 para sete, embora intuitivamente plausível para reduzir complexidade, revelou-se problemática ao provocar a perda de características importantes que distinguem claramente subclasses específicas.

Uma observação relevante foi o problema do *overfitting*, o modelo apresentou alta precisão no treino, mas baixo desempenho na validação, indicando dificuldade em generalizar. Para mitigar isso, técnicas como early stopping, algoritmos de *clustering* (DBSCAN e HDBSCAN) e a métrica *True Skill Statistic* (TSS) foram testadas.

A combinação de CNNs com Árvores de Decisão, apesar de promissora teoricamente, demonstrou dificuldades práticas devido à necessidade de ajuste minucioso de hiperparâmetros e abordagens de pré-processamento adequadas. Os resultados reforçaram a importância de abordagens multidimensionais, envolvendo não apenas algoritmos avançados, mas também estratégias eficazes como *data augmentation*, *weighted sampling* e classificação hierárquica, que poderiam ter melhorado o balanceamento dos dados, fortalecer a generalização dos modelos e contribuído para resultados mais consistentes e aplicáveis.

Em suma, o estudo concluiu que, apesar das dificuldades enfrentadas, a classificação automática de galáxias permaneceu uma área fértil para pesquisas futuras, exigindo abordagens ainda mais sofisticadas e refinadas.

BIBLIOGRAFIA

ELBOUCHEFRY,K.;SOUZA,R.S.de.Learning in Big Data:Introduction to Machine Learning. In: KNOWLEDGE Discovery in Big Data from Astronomy and Earth Observation. Amsterdam: Elsevier, 2020. P. 225–249. DOI: 10.1016/B978-0-12-819154-5.00023-0.

NASA. Galaxies- NASA Science. Disponível em: Galaxies - NASA Science. Acesso em: 14 fev 2025

GHARAT, Sarvesh; DANDAWATE, Yogesh. Galaxy classification: a deep learning approach for classifying Sloan Digital Sky Survey images. *Monthly Notices of the Royal Astronomical Society*, v. 511, n. 4, p. 5120–5124, 2022. DOI: 10.1093/mnras/stac457.

MCROBIE, Thomas. Applying a Deep Learning Approach to Galaxy Classification with Galaxy Zoo 2. *Medium*, 2023. Disponível em: https://medium.com/@thomas.mcrobie999/applying-a-deep-learning-approach-to-galaxy-classification-with-galaxy-zoo-2-afb51c81541f. Acesso em: 14 fev 2025

DIELEMAN, Sander; WILLETT, Kyle W.; DAMBRE, Joni. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, v. 450, n. 2, p. 1441–1459, 2015. DOI: 10.1093/mnras/stv632.

RAJA, Mudasir et al. Membership determination in open clusters using the DBSCAN Clustering Algorithm. *arXiv preprint arXiv:2404.10477*, 2024. Disponível em: https://arxiv.org/abs/2404.10477.

IBM. Overfitting. out 2021 Disponível em: https://www.ibm.com/think/topics/overfitting. Acesso em: 14 fev 2025

TRAMACERE, A. et al. ASTErIsM: application of topometric clustering algorithms in automatic galaxy detection and classification. *Monthly Notices of the Royal Astronomical Society*, v. 463, n. 3, p. 2939–2957, 2016. DOI: 10.1093/mnras/stw2103.

YOON, Sunhee; LEE, Wang-Hee. Application of true skill statistics as a practical method for quantitatively assessing CLIMEX performance. *Ecological Indicators*, v. 146, p. 109830, 2023. DOI: 10.1016/j.ecolind.2022.109830.

CYBORG CODES. What is Early Stopping in Deep Learning. *Medium*, 2021. abr 2024 Disponível em: https://cyborgcodes.medium.com/what-is-early-stopping-in-deep-learning-eeb1e710a3cf. Acesso em: 14 fev 2025

MA, X.; LI, X.; LUO, A.; ZHANG, J.; LI, H. Galaxy image classification using hierarchical data learning with weighted sampling and label smoothing. *Monthly Notices of the Royal Astronomical Society*, v. 519, n. 3, p. 4765–4779, 2023. DOI: 10.1093/mnras/stac3770.

IBM. XGBoost. mai 2024.

NATIONAL GEOGRAPHIC BRASIL. O que é uma galáxia? Elas são classificadas em 8 tipos, segundo a NASA. Disponível em: *https://www.nationalgeographicbrasil.com/espaco/2024/07/o-que-e-uma-galaxia-elas-sao-classificadas-em-8-tipos-segundo-a-nasa*.

SOARES, D. O'S.; Universidade Federal de Minas Gerais. Elíptica EM. Disponível em: *https://lilith.fisica.ufmg.br/dsoares/reino/eliptica-EM.htm*.

SACANI, Sérgio. A Galáxia Irregular IC 4710 fotografada pelo Hubble. SPACE TODAY, 26 fev. 2018. Disponível em: https://spacetoday.com.br/a-galaxia-irregular-ic-4710-fotografada-pelo-hubble/.