Draft version April 2, 2024
Typeset using LaTeX default style in AASTeX631

A Binary Classification of SFGs and AGNs Employing Different Clustering Techniques: BAWG2023 Hackathon

BAWG participants[1]

[1] *BRICS*

Keywords: Hackathons — Astronomy — Machine Learning — Clustering — Binary Classification

1. INTRODUCTION

The BRICS Astronomy Working Group (BAWG) Hackathon took place in Cape Town, South Africa, on October 18-19, 2023, subsequent to two days of BAWG science conference meetings. This two-day hackathon offered participants the opportunity to apply machine learning techniques to tackle a data-intensive astronomical challenge.

The primary task was to develop an optimal unsupervised learning pipeline for binary clustering between Active Galactic Nuclei (AGN) and Star-Forming Galaxies (SFGs). Participants, mainly postgraduate astronomy students from BRICS countries with varying levels of machine learning experience, engaged in this challenge.

Pre-hackathon preparation included Jupyter notebooks provided a week in advance, showcasing an unsupervised learning approach. This preparatory material aimed to equip participants with the necessary insights to undertake the challenge effectively.

The hackathon's goal was to equip postgrad students with the necessary data science skills to derive new insights from a provided dataset within the two-day event timeframe, fostering a hands-on experience in applying data science to real-world astronomical data.

2. DATA SET

In this hackathon, we use the early science radio continuum data from the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE, Jarvis et al. (2016)) survey. MIGHTEE is an extragalactic project undertaken by a South African-led international collaboration of researchers to explore star forming galaxies (SFG) and active galactic nuclei (AGN) evolution over cosmic time with the MeerKAT telescope. The survey focuses on four well-studied extragalactic deep fields; COSMOS, XMM-LSS, ELAIS-S1 and E-CDFS. Upon completion, the survey will cover up to 20 deg2 at μJy sensitivity at Giga-Hertz frequencies. MIGHTEE combines excellent multiwavelegth data from other deep surveys to provide our understanding of galaxy evolution.

The MIGHTEE-COSMOS multiwavelength catalogue comprised radio and matched measurements from optical, near-, mid- and far-infrared, and Xray information for the radio sources in the central part of the MIGHTEE Early Science Data in the COSMOS field and is assembled as follows Whittam et al. (2024). MIGHTEE-COSMOS radio catalogue was produced by Heywood et al. (2022). The radio catalogue was cross-matched with the optical and near-infrared counterparts adopted from Bowler et al. (2020) catalogue by Whittam et al. (2024). The optical and near-infrared catalogue comprised the near-infrared imaging in the *YJHKs* band; the optical measurements in the *grizy* bands from Hyper Suprime-Cam Subaru Strategic Program (HSC SSP); and deep optical imaging from CFHTLS's u^*griz bands. The host galaxy of 5,224 radio sources were identified and thus a sample of 5,224 radio sources had an optical and near-infrared counterpart. This sample was further cross-matched with measurements from other surveys by Whittam et al. (2022) as follows. Whittam et al. (2022) used the positions of the optical host galaxies to find X-ray counterparts to the MIGHTEE-COSMOS radio sources using the optical and infrared counterpart of the Chandra COSMOS-Legacy survey catalogue presented in Marchesi et al. (2016). Of 5,224 radio sources, 572 ($\sim$ 10%) were detected in X-ray observations.

The Mid-Infrared (MIR) counterparts were also added, taken from the COSMOS2015 catalogue Laigle et al. (2016). This catalogue provides the 3.6 μm, 4.5 μm, 5.8 μm, 8.0 μm fluxes respectively.

The Herschel Extra-galactic Legacy Project (HELP; Vaccari (2015)) provided the far-infrared data. The observations come from the Multiband Imaging Photometer (MIPS) instrument on the Spitzer Space Telescope, the Photodetector

Array Camera and Spectrometer (PACS) on Herschel, and the Spectral and Photometric Imaging Receiver (SPIRE) on Herschel. The MIPS provide 24 μm data, PACS produce 100 μm and 160 μm data, and SPIRE provides the 250 μm, 350 μm, and 500 μm data. Four thousand five hundred forty-one radio sources are identified in the MIPS and PACS data, and 4,958 in the SPIRE data.

The sources in this catalogue were also classified as SFGs or AGN by Whittam et al. (2022) using conventional astronomy techniques and catalogue article is summarised in Appendix A of their study.

3. BINARY CLUSTERING USING UNSUPERVISED LEARNING

The hackathon featured six groups, each comprising approximately four members. A diverse range of methods was employed for the clustering task. In the subsequent sections, we will offer brief descriptions of the pipelines utilized by each group, with their respective accuracies detailed.

3.1. *Group 1*

Group one commenced with data preprocessing, which included outlier replacement with the mean, normalization of the data, and feature removal based on a correlation analysis with a threshold of 0.95, thus reducing the features from 144 to 72. They concluded the preprocessing with brute-force feature selection, settling on the features {qir, PEAKFLUX, L14}. They employed Gaussian Mixture Models (GMMs) for clustering, achieving an accuracy of 87% on the testing set.

3.2. *Group 2*

Group two initiated their process with a log transformation on large-scale data, followed by normalization using Min-max scaling. They selected features associated with galaxy properties, such as luminosity, star formation rates, and mass: {Lbbdered(0.1-1), Lga(0.1-1), Ltor(1-30), SFR_IR, L14, qir, Mstar}. K-means clustering was utilized, resulting in a 79% accuracy on the testing set.

3.3. *Group 3*

Group 3's approach included a correlation analysis on the 144 features, applying a 0.95% threshold and reducing the feature count to 69. This was followed by Min-max scaling normalization. They used K-means for clustering and reported a 75% accuracy on the testing set.

3.4. *Group 4*

Group 4 addressed missing values by removing the affected features, then standardized the data and applied PCA, reducing it to 13 components representing 84% of the variance. They opted for GMMs for clustering, attaining a 70% accuracy on the testing set.

3.5. *Group 5*

Group 4 initiated their analysis by conducting a correlation assessment on the 144 features, setting a threshold at 0.80% which reduced the features to 43. After further optimization, they identified 9 key features: {E_S_INT, IM_MAJ, IM_MIN, COS_best_z_v5, BB, EBVbbb, EBVgal, Mstar, qir}. Without any additional preprocessing, they utilized GMMs and achieved an accuracy of 78% on the testing set.

3.6. *Group 6*

Group 6 standardized the data using a standard scaler and then engaged in a multi-step feature selection process. This process involved eliminating features with a correlation greater than 80%. They utilized ChatGPT and academic papers to discern the significance of the features, consequently removing those related to position, orientation, and interstellar dust and gas. The focus was then shifted to features associated with Luminosity, Mass, and Infrared Luminosities. The final selection included ten features {fSF, fAGN, qir, Mstar, Lbb(0.1-1), NU_EFF, SB, Lga(0.1-1), BB, Nh, age}, chosen after evaluating various combinations on the validation set.

Employing these ten features, the group explored several clustering approaches, including k-means, PCA, and Bayesian Gaussian Mixture Model (BGMM). However they manged to achieve their highest accuracy of 83% on the testing set with GMMs.

4. DISCUSSIONS

We detail the methodologies employed by six groups during the hackathon, where the objective was binary clustering through unsupervised learning. Each group utilized unique strategies, with notable commonalities and differences in their approaches. We also discuss and compare groups that achieved accuracy. The discussion concludes with a comparison to other supervised learning approaches.

4.1. *Common Methods*

A prevalent theme across the groups was the emphasis on data preprocessing, a crucial step in unsupervised learning. Most groups opted for normalization (Min-max scaling or standardization) and feature selection, which was pivotal in enhancing the clustering performance. Specifically, Groups 1, 3, 4, and 6 conducted a form of feature reduction, either through correlation analysis or principal component analysis (PCA), demonstrating the importance of eliminating redundant features to improve model efficiency and accuracy.

Another common strategy was the use of GMMs, favored by Groups 1, 4, 5, and 6. This preference underscores GMMs' flexibility and effectiveness in identifying latent structures within the data, especially when the clusters are not distinctly separable.

4.2. *Differences*

The main differences lay in the specific preprocessing techniques and the features selected for the clustering task. For instance, Group 1 used brute-force feature selection to identify their final features, whereas Group 6 combined insights from ChatGPT, academic literature, and a rigorous feature evaluation process. The diversity in feature selection approaches, from correlation thresholds to leveraging external knowledge, illustrates the varied strategies that can be employed in unsupervised learning tasks.

Groups also differed in their clustering algorithms, with Group 2 utilizing K-means, a contrast to the GMMs preferred by the majority. This variance in algorithm choice highlights the experimentation with different methodologies to determine the most suitable for the given data and task.

4.3. *Analysis of Accuracies*

The groups' accuracies in the hackathon showed a range from 70% to 87%, with Group 1 achieving the highest at 87%. This variation highlights the impact of different data preprocessing and clustering strategies on performance. The success of Group 1, using Gaussian Mixture Models (GMMs), underscores the effectiveness of thorough preprocessing combined with a sophisticated clustering algorithm. The results emphasize the significance of careful method selection and data handling in optimizing unsupervised learning outcomes.

5. CONCLUSION

the BAWG2023 Hackathon offered a unique platform for postgraduate astronomy students from BRICS countries to apply machine learning techniques to the challenge of binary clustering Active Galactic Nuclei (AGN) and Star-Forming Galaxies (SFGs) using data from the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) survey.

The hackathon featured diverse methodologies employed by six groups, emphasizing the importance of data preprocessing, feature selection, and the use of Gaussian Mixture Models (GMMs) for some groups, highlighting the flexibility and effectiveness of these models in clustering tasks. The range of accuracies achieved (70% to 87%) underscored the impact of different preprocessing and clustering strategies on performance, with Group 1 achieving the highest accuracy through comprehensive preprocessing and sophisticated clustering algorithms.

This achievement highlights the valuable role hackathons can play in discovering usable pipelines for labelling astronomical data, thereby broadening our comprehension of the cosmos.

REFERENCES

Bowler, R. A. A., Jarvis, M. J., Dunlop, J. S., et al. 2020, MNRAS, 493, 2059, doi: 10.1093/mnras/staa313

Heywood, I., Jarvis, M. J., Hale, C. L., et al. 2022, MNRAS, 509, 2150, doi: 10.1093/mnras/stab3021

Jarvis, M., Taylor, R., Agudo, I., et al. 2016, in MeerKAT Science: On the Pathway to the SKA, 6, doi: 10.22323/1.277.0006

Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24, doi: 10.3847/0067-0049/224/2/24

Marchesi, S., Civano, F., Elvis, M., et al. 2016, ApJ, 817, 34, doi: 10.3847/0004-637X/817/1/34

Vaccari, M. 2015, in The Many Facets of Extragalactic Radio Surveys: Towards New Scientific Challenges, 27, doi: 10.22323/1.267.0027

Whittam, I. H., Jarvis, M. J., Hale, C. L., et al. 2022, MNRAS, 516, 245, doi: 10.1093/mnras/stac2140

Whittam, I. H., Prescott, M., Hale, C. L., et al. 2024, MNRAS, 527, 3231, doi: 10.1093/mnras/stad3307