Utilising an Isolation Forest Algorithm to Identify Unusual Galaxies in GAMA

Kieran Broadbelt (1st Year PhD candidate)
Supervisors: Kevin Pimbblet, Daniel Farrow
Contact: k.broadbelt-2018@hull.ac.uk

E.A. Milne Centre for Astrophysics

University of Hull

Abstract

We apply an Isolation Forest (iForest) algorithm to identify anomalous objects in two samples from the Galaxy and Mass Assembly fourth data release (GAMA DR4). The first sample is high signal-to-noise galaxies and the second is E+A galaxies. We apply the iForest using three approaches: spectroscopic data, photometric data and a combination of both. Across the six subsets we identify 102 of the most unusual galaxies. Among the E+A sample are galaxies that are unexpectedly star-forming, showing strong absorption lines and minimal [OII] or [OIII] emission. For the high S/N galaxies we recover extreme emitters, rare 'Green Bean' galaxies, and red spiral galaxies. Additionally, the model detects pipeline and data reduction errors. We explore possible explanations for the star-forming nature of the anomalous E+A galaxies and place these findings in the broader context of E+A selection.

Introduction

Large datasets pose increasing problems to robustly analyze the wealth of data (e.g., LSST Rubin, DESI, Euclid). Automated and robust algorithms have been created to classify and characterize galaxies
(e.g., Lochner et al. 2016; Clarke et al. 2020; Reza 2021; Chang et al. 2021), but there remains the issue of how to detect novel or unusual objects in these surveys. While unsupervised machine learning has been used before to find such outliers (e.g., Baron & Poznanski 2016), here we present an iForest algorithm applied to the GAMA dataset.

We apply it to two main galaxy GAMA samples: E+A galaxies and high signal-to-noise (S/N) galaxies. We apply the iForest to these two datasets using three parameter subsets – spectroscopic, photometric and combined – producing six subsamples in total. Anomalies are identified in all six using the iForest.

The E+A sample is selected by applying the cuts from Wilkinson et al. (2017): $\mathrm{EW}_{\mathrm{H}\delta} > 3\text{Å}$ and $\mathrm{EW}_{[\mathrm{OII}]} > -2.5\text{Å}$. The S/N sample is cut using an equivalent width S/N=3.

Data

We take our data from the GAMA DR4 (Driver et al. 2022) and the necessary DMUs (Liske et al. 2015; Gordon et al. 2017; Bellstedt et al. 2020). GAMA DR4 compiles nearly 250,000 spectra and a large number of data management units that contain a wide variety of additional galaxy properties. We use the spectroscopic DMU SPECLINESSFRV05 (Gordon et al. 2017), the photometric DMU GKVINPUTCATV02 (Bellstedt et al. 2020) and the spectra DMU SPECCATV27 (Liske et al. 2015).

Methodology

Isolation forests work by constructing a tree like structure – an iTree – from the data they are applied to. Anomalous objects in the data are more likely to be isolated within this structure and therefore more likely to be at the root of the structure.

To analyze the morphology of the anomalous galaxies detected by the iForest, we determine 8 morphometric parameters: concentration, asymmetry and smoothness (CAS Conselice 2003; Bershady et al. 2000), and Gini and M20 (Abraham et al. 2003; Lotz et al. 2004). We add 3 further parameters: multimode, intensity and deviation (MID Freeman et al. 2013; Peth et al. 2015) to complement this analysis using the python package STATMORPH (Rodriquez-Gomez et al. 2019).

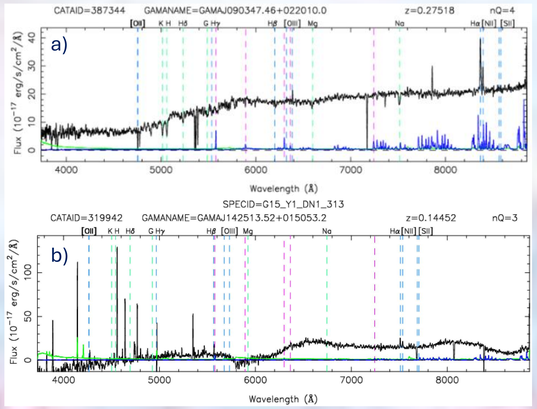

Figure 1a. example star-forming E+A spectra **b.** example false emission line spectra extracted from the isolation forest applied to the E+A spectroscopic data.

Results

The iForest is calibrated to select 25 anomalous galaxies per sample, resulting in 150 objects identified by the algorithm. In total, we find 121 unique objects due to duplicates , 'bad spectra' (13 from data reduction pipeline issues) and shredded galaxies. We find 13 extreme emission line galaxies (EELGs, $\mathrm{EW} \geq 300\text{Å}$), one of which is a rare 'Green Bean' identified by Prescott and Sanderson (2019), and 21 red spirals are found, all of low masses ($< 10^{10}\mathrm{M}_{\odot}$).

The morphometrics indicate that the anomalous galaxies are typically "normal" in morphology, with few extremes. We find 75% of the anomalies are found in the Sb/Sc/Irr region, 20% lie in the merger region and the remaining 5% in the E/S0 region of the G-M20 plot shown in fig. 4. Within the E+A sample we find 20 of the 49 E+A galaxies have strong Hα EW measures. This is suggestive of current star-formation or AGN activity. Further, 13 of the E+As have low [OIII] ($\mathrm{EW}_{[\mathrm{OIII}]} < 6\text{Å}$) and have narrow Hα lines. We use BPT and WHAN diagrams (WHAN; fig. 3) and find 1 AGN candidate in the weak region, and a possible candidate on the border of strong AGN. Most 30 E+As lie in the star-forming region when they should typically lie in the passive region. Ruling out AGN activity, we look at possible dust obscuration. By analyzing the infra-red and optical bands we find that none of the anomalous E+A galaxies are anomalously red.

We propose that E+A galaxies isolated here are still star-forming, but the spectra suggests they are not forming high mass stars of O and B class that would result in prominent [OII] or [OIII] lines. The presence of Hα strongly suggest star formation is occurring and the presence of Balmer absorption suggests a population of A class or older stars. This star-formation could be happening due to high metallicity and low temperatures reducing the Jeans mass so low that it would be impossible for higher mass O and B class stars to form. The spectra of these galaxies appears like E+A galaxies, with the lack of [OII] and strong Hδ and Hα absorption.

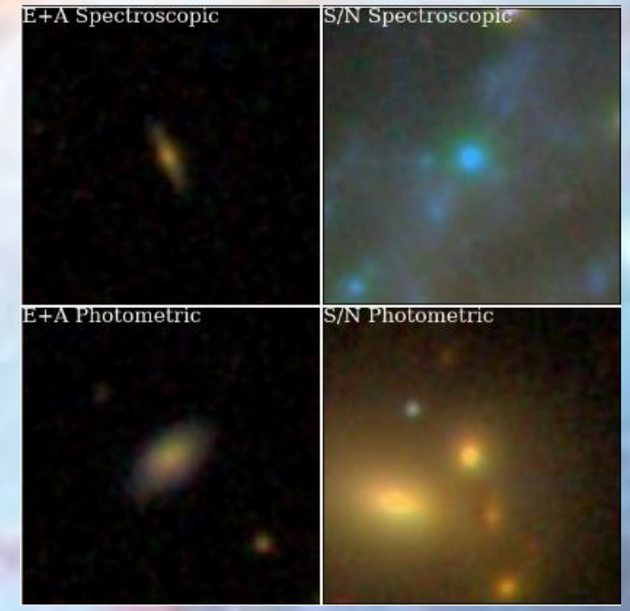

Figure 2. SDSS imaging of the 4 most anomalous galaxies based on anomaly scores. In each corner shows the galaxy sample and data used for isolation

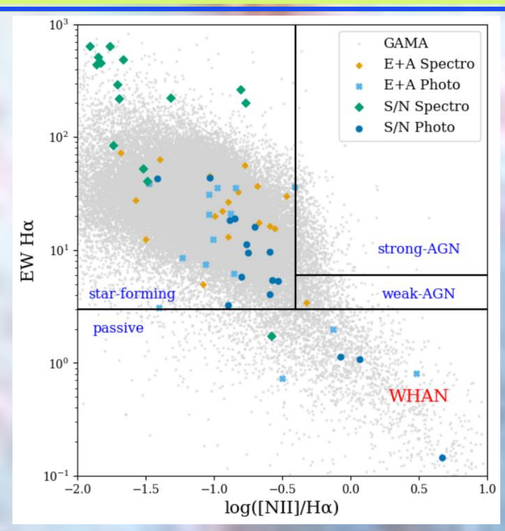

Figure 3. WHAN plot of the galaxies in GAMA and the anomalous objects. Orange are the high S/N sample, blue to E+A sample. 1 E+A lies in the weak AGN region, majority of E+A lie in the SF region.

Conclusions

We have utlised an iForest algorithm to isolate anomalous galaxies in two core samples of E+A galaxies and high signal-to-noise galaxies. We isolated a total of 101 unique objects and find 13 EELGs, 1 Green Bean, 21 red spirals, 13 star-forming E+As and several data reduction/pipeline errors.

The SF-E+A galaxies are still star-forming but are unable to form O and B class stars that would result in a stronger [OII] line. They are however still forming A and lower mass stars as indicated by the presence of both Hδ and Hα. Two possible hypotheses we propose are,

i. Small scale perturbations in the ISM make it more difficult for higher mass stars like O and B class to form but they precipitate the formation of lower mass stars like A class.

ii. The galaxies are still efficiently star-forming but due to high metallicity and low temperatures they are unable to form O and B class stars.

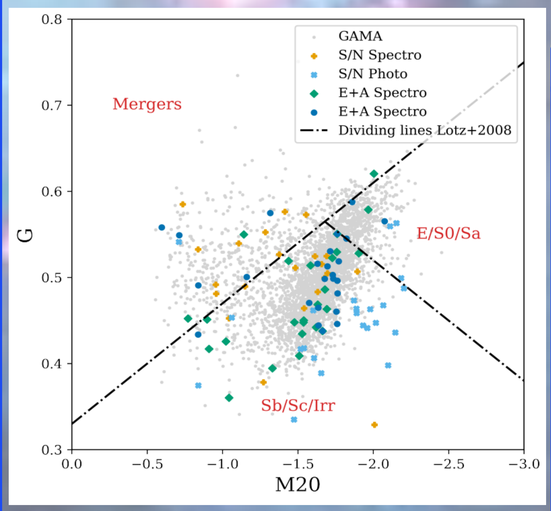

Figure 4. Gini-M20 plot with dividing regions from Lotz+2008. The figure shows that most of the anomalous E+A galaxies don't lie in the E/S0/Sa region like expected but instead in the Sb/Sc/Irr region.

References

Abraham et al., 2003, ApJ, 588, 218
Baron & Poznanski, 2016, MNRAS, 465, 4530
Bellstedt et al., 2020, MNRAS, 496, 3235
Bershady et al., 2000, AJ, 119, 2645
Chang et al., 2021, ApJ, 920, 68
Clarke et al, 2020, A&A, 639, A84
Conselice, 2003, ApJS, 147, 1
Dressler & Gunn, 1983, ApJ, 270, 7
Driver et al., 2022, MNRAS, 513, 439
Freeman et al., 2013, MNRAS, 434, 282
Gordon et al., 2017, MNRAS, 465, 2671
Liske et al., 2015, MNRAS, 452, 2087
Lochner et al., 2016, ApJS, 225, 31
Liu et al., 2008, in 2008 Eighth IEE International Conference on Data Mining. pp 413–422
Lotz et al., 2004, AJ, 128, 163
Peth et al., 2015, MNRAS, 458, 963
Prescott & Sanderson, 2019, ApJ, 885, 40
Rodriquez-Gomez et al., 2019, MNRAS, 483, 4140
Wilkinson et al., 2017, MNRAS, 472, 1447