Radio Galaxy Zoo Classification Pipeline

Matthew Alger
The Australian National University

May 14, 2016

In this document, I will describe the Radio Galaxy Zoo (RGZ) classification pipeline that I have implemented.

1 Definitions

A *Radio Galaxy Zoo subject* is a representation of one radio source. It consists of a location in the sky (specified in right ascension/declination coordinates), an image of the sky at this location in radio wavelengths, and an image of the sky at this location in infrared wavelengths. The subject may contain other nearby radio sources.

The Radio Galaxy Zoo *crowd classifications* are crowdsourced solutions to the classification task. Each crowd classification contains the combination of radio sources in a subject that a volunteer associates with the same active galactic nucleus (AGN), as well as the location where the volunteer believes the host galaxy is located. There are multiple crowd classifications for each RGZ subject.

2 The Classification Task

The goal of the classification task is to locate the host galaxy of the subject.

2.1 Pipeline Inputs and Outputs

As input to the classification pipeline we take a RGZ subject and (for training) a set of associated crowd classifications. The output of the pipeline is the location of the host galaxy associated with the subject.

2.2 Assumptions and Limitations

I am ignoring the fact that a RGZ subject may contain multiple host galaxies, and instead assuming that there is only one host galaxy per subject.

I am exclusively working with the Australia Telescope Large-Area Survey (ATLAS)[5] data set for now, though I expect my results to generalise to both Faint Images of the Radio Sky at Twenty-Centimeters (FIRST)[3] and the upcoming Evolutionary Map of the Universe (EMU)[6]. This is important as the majority of RGZ subjects are from FIRST, and the vast majority of subjects to be classified in future will be from EMU. The reason for this limitation is twofold: the ATLAS data set is small and well-known (containing 2443 radio subjects), and thus provides a good data set for exploring machine learning techniques; and the ATLAS data set is similar to the data that will be collected in EMU[1].

I am assuming that radio sources associated with a host galaxy will be "small", i.e., that they are less than 2 arcminutes in diameter. 2 arcminutes is the width of an image presented to RGZ volunteers. This assumption does not hold in general, as some radio sources can be spread over a very large area and these are known to be present in the RGZ data[2].

3 Collating Crowd Classifications

Raw crowd classifications are not immediately useful. There are multiple classifications for the same subject, and these may not agree. The first step in the pipeline is thus to collate the crowd classifications into labels for training. There are two components to collation. The first is collating the radio components

associated with the same AGN, and the second is collating the locations of the host galaxies associated with each set AGN. The collated radio components are called the *consensus radio combination* and the collated host galaxy locations are called the *consensus host galaxy locations*. [**TODO:** *Detail the method of collating radio components. My new method differs from Kyle's.*]

Collating the radio components is straightforward and I loosely follow the method of Banfield et al. [1]. I count the occurrences of each unique radio combination, and then the most popular radio combination is considered the consensus radio combination. [**TODO:** *Elaborate.*]

Collating the locations of each radio combination is more complicated. Banfield et al. [1] use kernel density estimation to find the most common location chosen by volunteers, however this is not robust and does not allow us to find which galaxy was intended to be chosen by each volunteer (which is useful if we want to estimate the uncertainty in the consensus). Instead, I cluster the volunteers' locations using PG-means[4] and choose the cluster with the most members as the consensus host galaxy location. [**TODO:** *Elaborate on PG-means. Compare to existing method. Have a diagram.*]

After collation, the crowd classifications provide us with a map between RGZ subjects and host galaxies.

4 Locating Host Galaxies as Binary Classification

Each subject contains a number of potential host galaxies. We can cast the problem of finding the true host galaxy as binary classification by labelling each potential host galaxy with 0 if it is not the true host and 1 if it is.

To find potential host galaxies, we use the Spitzer Wide-Area Infrared Extragalactic Survey (SWIRE) Chandra Deep Field South (CDFS) Region Fall '05Spitzer Catalog[**TODO:** *Cite.*] and the SWIRE European Large Area ISO Survey — South 1 (ELAIS-S1) Region Fall '05 Spitzer Catalog[**TODO:** *Cite.*], available through the Infrared Science Archive's GATOR interface[1]. These catalogues contain all infrared galaxies detected in the CDFS and ELAIS-S1 regions, which are the regions covered by ATLAS. [**TODO:** *Show a diagram.*]

Finding the labels for each potential host amounts to finding the nearest SWIRE potential host for each true host identified by the crowd classifications. For the location in each crowd classification, the nearest SWIRE potential host is found and its label is set to 1. All other labels are set to 0.

References

[1] J. Banfield, O. Wong, K. Willett, R. Norris, L. Rudnick, S. Shabala, B. Simmons, C. Snyder, A. Garon, N. Seymour, et al. Radio Galaxy Zoo: host galaxies and radio morphologies derived from visual inspection. *Monthly Notices of the Royal Astronomical Society*, 453(3):2326–2340, 2015.

[2] J. Banfield, H. Andernach, A. Kapińska, L. Rudnick, M. Hardcastle, G. Cotter, S. Vaughan, T. Jones, I. Heywood, J. Wing, O. Wong, T. Matorny, I. Terentev, . R. López-Sánchez, R. Norris, N. Seymour, S. Shabala, and K. Willett. Radio Galaxy Zoo: discovery of a poor cluster through a giant wide-angle tail radio galaxy. *Monthly Notices of the Royal Astronomical Society*, 2016. doi: 10.1093/mnras/stw1067. URL http://mnras.oxfordjournals.org/content/early/2016/05/05/mnras.stw1067.abstract.

[3] R. H. Becker, R. L. White, and D. J. Helfand. The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters. *Astrophysical Journal*, 450:559, Sept. 1995. doi: 10.1086/176166.

[4] Y. F. G. Hamerly. PG-means: learning the number of clusters in data. *Advances in neural information processing systems*, 19:393–400, 2007.

[5] R. P. Norris, J. Afonso, P. N. Appleton, B. J. Boyle, P. Ciliegi, S. M. Croom, M. T. Huynh, C. A. Jackson, A. M. Koekemoer, C. J. Lonsdale, et al. Deep atlas radio observations of the chandra deep field-south/spitzer wide-area infrared extragalactic field. *The Astronomical Journal*, 132(6):2409, 2006.

[6] R. P. Norris, A. M. Hopkins, J. Afonso, S. Brown, J. J. Condon, L. Dunne, I. Feain, R. Hollow, M. Jarvis, M. Johnston-Hollitt, E. Lenc, E. Middelberg, P. Padovani, I. Prandoni, L. Rudnick,

[1] http://irsa.ipac.caltech.edu/applications/Gator/

N. Seymour, G. Umana, H. Andernach, D. M. Alexander, P. N. Appleton, D. Bacon, J. Banfield, W. Becker, M. J. I. Brown, P. Ciliegi, C. Jackson, S. Eales, A. C. Edge, B. M. Gaensler, G. Giovannini, C. A. Hales, P. Hancock, M. T. Huynh, E. Ibar, R. J. Ivison, R. Kennicutt, A. E. Kimball, A. M. Koekemoer, B. S. Koribalski, Á. R. López-Sánchez, M. Y. Mao, T. Murphy, H. Messias, K. A. Pimbblet, A. Raccanelli, K. E. Randall, T. H. Reiprich, I. G. Roseboom, H. Röttgering, D. J. Saikia, R. G. Sharp, O. B. Slee, I. Smail, M. A. Thompson, J. S. Urquhart, J. V. Wall, and G.-B. Zhao. EMU: Evolutionary Map of the Universe. *PASA*, 28:215–248, Aug. 2011. doi: 10.1071/AS11021.