Project Proposal: Machine Learning on the Radio Galaxy Zoo

Matthew Alger Supervisor: Cheng Soon Ong

April 15, 2016

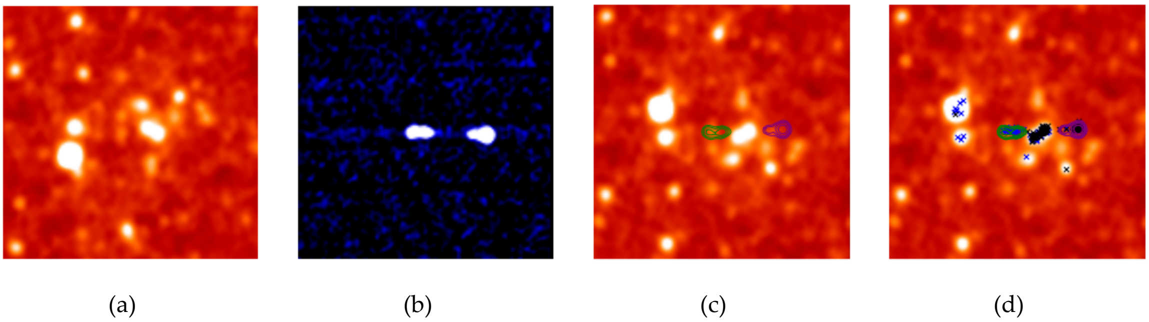
Black holes can be detected by looking for the jets of matter they emit. For distant, supermassive black holes, these jets emit radio waves, so we can see them in radio surveys such as the Australia Telescope Large Area Survey (ATLAS)[7] and Faint Images of the Radio Sky at Twenty-Centimeters (FIRST)[3] surveys. While we can learn a lot from observing these black holes, we can't interpret the observations until we know what galaxy the black hole is in — for example, we can't tell how large a black hole is unless we know how far away it is, and we can only determine this by observing the galaxy containing the black hole[1]. Matching a black hole (observed in radio images) with its host galaxy (observed in infrared images) is called "cross-identification", and a radio-emitting black hole is called a "radio source". Radio sources can be very complex, and cross-identification is very difficult — as a result, cross-identification is usually done by hand[2]. Upcoming radio surveys such as the Evolutionary Map of the Universe[6] and the WODAN survey[9] are expected to detect over 100 million new radio sources. Cross-identifying these by hand is infeasible, but around 10% of these new radio sources will be too complex for existing cross-identification algorithms[2, 6]. Radio Galaxy Zoo (RGZ)[2] is a crowdsourced citizen science project where volunteers manually perform the cross-identification process for over 175000 radio sources. The RGZ project had resulted in over 30000 cross-identifications by mid 2015, compared to around 6000 cross-identifications completed by experts[2].

My project aims to use the volunteers' cross-identifications from RGZ as training data for machine learning algorithms so that we can learn to automatically perform the cross-identification task. To my knowledge, machine learning has not yet been used on this cross-identification task — existing approaches tend to be based on astrophysical models, e.g. Fan et al. 2015[4].

The three main parts of this project are understanding and manipulating the RGZ data, casting the cross-identification problem as a classification task, and finding a good machine learning algorithm for this classification task.

RGZ data

The RGZ data are a set of subjects and associated volunteer classifications. A subject is an image of a section of the sky in both infrared and radio wavelengths. The classifications are, for each volunteer and each subject, which radio emissions in the radio image the volunteer believes are from the same radio source, and which location in the infrared image the volunteer believes each radio source is located at.



(a) (b) (c) (d)

Figure 1: (1a) Infrared image of subject. (1b) Radio image of subject. (1c) Volunteers decide which emissions are from the same source. (1d) Volunteers decide where each radio source is located.

These are stored in a MongoDB database. The raw locations provided by volunteers don't necessarily agree and need to be processed to find a single location for each radio source; Banfield et al.[2] provide code to do this, but their code doesn't find the uncertainty associated with the resulting location and we may want to make use of this uncertainty while learning. It may also not be possible to simply choose the most agreed upon location for a given radio source — for example, 49% of the volunteers may believe that location 1 is where the radio source is located, and 51% may believe that the location is actually location 2. It could be misleading to say that location 2 is the "true" location when there is this much disagreement, and so I will have to determine the extent of this problem and subsequently find solutions.

Forming a classification task

The cross-identification problem can be cast as a binary classification problem. The naïve way to do this is to slide a window over the infrared and radio images, where each window has the label 1 when the window is centred on the radio source and 0 otherwise. However, this is expensive, especially for larger images. A potentially better way is to identify locations that may host the radio source, and then label these as 1 if they are the true radio host and 0 otherwise. I will attempt to identify such potential hosts and classify them in this way. I will also need to find useful features of the potential hosts to use as inputs for the classifier — examples may include the infrared brightness of the potential hosts, the distance from the potential hosts to the centre of the radio emissions, and so on.

Training a classifier

The final part of the problem is to train a classifier and subsequently perform the classification task. It isn't obvious what kind of classifier to use, but preliminary results suggest that a logistic regression-based approach may be suitable. A logistic regression classifier can be trained on the potential host features, and configured to output the probability that each potential host should have a 1 label. The potential host with the highest probability is then considered to be the host galaxy. It may also be possible to treat the problem as a computer vision object detection problem on the radio image. If so, then a region-based convolutional neural network may be used to detect the host galaxy. These approaches could even be combined.

Future work

This proposal describes the first semester of my honours project. The second semester will focus on some facet of crowdsourced active learning with the Radio Galaxy Zoo, though the specific details of this have not yet been determined and may depend on the results of the first semester. As a result of this, I will need to be thinking about the active learning problem while working on the first semester of the project, and this is reflected in my project plan.

Project plan

- February:
  - Familiarise myself with the dataset.
  - Cast the problem as a binary classification problem.
  - Understand Kyle Willett's RGZ code accompanying the Banfield et al. 2015 paper.
- March:
  - Convert raw RGZ data into training sets.
  - Run RGZ training sets through simple classifiers.
  - Read papers on crowdsourced active learning (e.g. Yan et al. 2011[10] and Mozafari et al. 2012[5]).
  - Read papers on other astronomical classification problems (e.g. Richards et al. 2012[8]).
  - Determine the extent of the label disagreement problem. Can we just use majority vote?
- April:
  - Begin writing up problem observations.
  - Investigate using metadata and RGZ image structure as features.
  - Implement a CNN and understand CNN features with help from computer vision researchers.
  - Experiment with different classifiers to find the best classifier to use for the classification task.

- Summarise the crowdsourced active learning papers.
  - List weaknesses of crowdsourced active learning papers.

- May:
  - Make a software package for the RGZ classification problem.
  - Implement the active learning algorithm from Yan et al. 2011.
  - Decide on future work for semester 2.
  - Work on making software package user-friendly for astrophysicists.
  - Present results (final report and presentation).

References

[1] Radio galaxy zoo. *http://radio.galaxyzoo.org*, 2015.

[2] JK Banfield, OI Wong, KW Willett, RP Norris, L Rudnick, SS Shabala, BD 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.

[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, September 1995.

[4] Dongwei Fan, Tamás Budavári, Ray P. Norris, and Andrew M. Hopkins. Matching radio catalogues with realistic geometry: application to swire and atlas. *Monthly Notices of the Royal Astronomical Society*, 451(2):1299–1305, 2015.

[5] Barzan Mozafari, Purnamrita Sarkar, Michael J. Franklin, Michael I. Jordan, and Samuel Madden. Active learning for crowd-sourced databases. *CoRR*, abs/1209.3686, 2012.

[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, 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, August 2011.

[7] Ray P Norris, José Afonso, Phil N Appleton, Brian J Boyle, Paolo Ciliegi, Scott M Croom, Minh T Huynh, Carole A Jackson, Anton M Koekemoer, Carol 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.

[8] Joseph W. Richards, Dan L. Starr, Henrik Brink, Adam A. Miller, Joshua S. Bloom, Nathaniel R. Butler, J. Berian James, James P. Long, and John Rice. Active learning to overcome sample selection bias: Application to photometric variable star classification. *The Astrophysical Journal*, 744(2):192, 2012.

[9] Huub Röttgering, Jose Afonso, Peter Barthel, Fabien Batejat, Philip Best, Annalisa Bonafede, Marcus Brüggen, Gianfranco Brunetti, Krzysztof Chyży, John Conway, Francesco De Gasperin, Chiara Ferrari, Marijke Haverkorn, George Heald, Matthias Hoeft, Neal Jackson, Matt Jarvis, Louise Ker, Matt Lehnert, Giulia Macario, John McKean, George Miley, Raffaella Morganti, Tom Oosterloo, Emanuela Orrù, Roberto Pizzo, David Rafferty, Alexander Shulevski, Cyril Tasse, Ilse van Bemmel, Bas Tol, Reinout Weeren, Marc Verheijen, Glenn White, and Michael Wise. Lofar and apertif surveys of the radio sky: Probing shocks and magnetic fields in galaxy clusters. *Journal of Astrophysics and Astronomy*, 32(4):557–566, 2012.

[10] Yan Yan, Rómer Rosales, Glenn Fung, and Jennifer G. Dy. Active learning from crowds. *Proceedings of the 28th International Conference on Machine Learning*, pages 1161–1168, 2011.