SARAO 2022 MSc Project Proposal

Section A

1. Project Title

Unsupervised Machine Learning Applied to Radio Data

2. Research Area

Science

3. Academic Level

MSc

4. Abstract

Telescopes such as MeerKAT, LOFAR, ASKAP and culminating in the SKA, are pushing the limits of data analysis. Machine learning is well recognised as being critical to handling this data deluge. However most commonly used algorithms are supervised, requiring a (often very large) training set which can be expensive and time-consuming to obtain. Unsupervised machine learning techniques, such as clustering algorithms, can be incredibly powerful when working with large, unlabeled datasets.

This project proposes to develop a new unsupervised learning framework for radio data, applying clustering techniques coupled with active learning to collect human input. An unsupervised framework will allow rapid construction of training sets, as well as the detection of patterns in the data that humans may not notice. The student will develop this framework by extending the existing tool Astronomaly [1], and apply it to existing MeerKAT data such as the MeerKAT Galaxy Clusters Legacy Survey.

5. Primary Supervisor's Details

a. **Full name:** Dr. Michelle Lochner
b. **Email address:** mlochner@uwc.ac.za
c. **University:** University of the Western Cape/ South African Radio Astronomy Observatory

Section B

1. Scientific Merit

Machine learning has emerged as a critical tool for managing large datasets. However, most commonly used are supervised learning algorithms, such as convolutional neural networks (see e.g. [2]) that need large training sets which may take a significant investment of expert hours or may simply not be available.

Unsupervised learning offers an alternative approach that is especially useful when most of the data are unlabeled. One branch of unsupervised learning is clustering, which automatically sorts the data into similar clusters. An approach has been applied to LOFAR data [3] using self-organising maps implemented in PINK [4] and also using a more traditional clustering algorithm [5], showing the potential of this approach. The advantage of unsupervised learning is that it allows the data to speak for itself, without applying human biases, allowing for the possibility of detecting patterns in the data.

In this project, the student will extend the existing framework of Astronomaly to incorporate a clustering component. Initially, we will be guided by the framework developed by S. Webb and M. Lochner [6] which was applied to fast optical transients. Although we will focus on radio images, the framework is general to any type of data.

The aim is to not simply cluster the data, but also to use the clusters to organise the data into a training set in far less time than it would take to manually label the data. The student will apply this framework to existing MeerKAT data, such as the MeerKAT Galaxy Clusters Legacy Survey (MGCLS, soon to be released), and implement several algorithms including traditional clustering and deep learning approaches.

The student will add an implementation of the popular HDBSCAN [7] algorithm to Astronomaly as well as a version of a self-organising network [8]. These two algorithms will allow a comparison between more traditional machine learning approaches and deep learning to determine which is more appropriate for radio data.

The student will then develop a framework to use the detected clusters to rapidly build a training set. The aim is to construct deep learning ready training sets using only a small amount of expert labeling time. This part of the project may require a larger dataset than MGCLS, in which case we will turn to the publicly available LOFAR Two-Metre Sky Survey [9], which has already been analysed making it a good reference set.

This framework has the potential to be extremely useful not just for MeerKAT data, but for other large datasets from telescopes such as LOFAR, ASKAP and eventually the SKA.

2. Feasibility

Data collection is not an issue as the project is to be run on existing, public data. We will focus on the MGCLS data, which will be public by the time the student starts and the LOFAR Two-Metre Sky Survey, which is public. We will make use of the existing framework of Astronomaly, meaning the student will be able to focus on their core project rather than infrastructure work to work with the data.

Rough timeline:

- 2022 - Develop a clustering framework to include in Astronomaly, including implementing various unsupervised algorithms
- 2023 - Run framework on existing MeerKAT and LOFAR data, investigating the classes of objects automatically found

3. Link to SARAO research priority areas

This project is closely linked to SARAO research priority areas with a focus on MeerKAT. It will develop critical data science and radio data manipulation skills and pave the way for use of these techniques with the SKA.

4. Student Requirements

Good programming skills are critical for this project, primarily in python. Experience with machine learning is advantageous, but not essential.

References

[1] https://arxiv.org/abs/2010.11202 (M. Lochner & B. Bassett, 2020)
[2] https://arxiv.org/abs/2102.03780 (B. Becker et al., 2021)
[3] https://arxiv.org/abs/2011.06001 (R. Mostert et al., 2020)
[4] https://ui.adsabs.harvard.edu/abs/2015ASPC..495...81P/abstract (K. Polsterer et al., 2015)
[5] https://arxiv.org/abs/2102.00843 (K. Ntwaetsile & J. E. Geach, 2021)
[6] https://arxiv.org/abs/2008.04666 (S. Webb, M. Lochner et al. 2020)
[7] http://dx.doi.org/10.21105/joss.00205 (L. McInnes & J. Healy, 2017)
[8] https://ui.adsabs.harvard.edu/abs/2007SchpJ...2.1568K (T. Kohonen & T. Honkela, 2007)
[9] https://arxiv.org/abs/1811.07926 (T. W. Shimwell et al. 2018)