Finding New Pulsars Using Machine Learning

Jacob Ian Matthews

Draft 1

May 25, 2020

Abstract

Contents

1 Introduction

1.1 Aim

This project consists of three aims:

i. Investigate the use of Machine Learning (ML) techniques in surveying pulsars;

ii. Create a training dataset for a Machine Learning algorithm to find pulsars in data obtained by the Murchison Widefield Array (MWA); and

iii. Evaluate the utility of the Machine Learning algorithm used by the LOFAR Telescope for the Murchison Widefield Array, and adjust the algorithm as necessary to achieve optimum pulsar candidate generation.

1.2 Structure of this Report

In this report, I will first explain in *Section 1.3* how Pulsars, Radio Astronomy, and Machine learning work, and then explain what has been dubbed the "Candidate Selection Problem" (Lyon et al. 2016) and why Machine Learning is necessary in completing future pulsar surveys.

In *Section 2* I discuss the methods undertook in: (i) developing the machine learning training dataset for the Murchison Widefield Array algorithm, and (ii) evaluating the machine learning algorithm used by the LOFAR surveys for use with the Murchison Widefield Array.

In *Section 3* I analyse the results and findings obtained by the methods described in *Section 2*, and in *Section 4* I will discuss (i) the efficacy of the training datasets, and (ii) the usefulness of the LOFAR machine learning algorithm with the Murchison Widefield Array and why changes were made to the algorithm.

This report will end with my final conclusions on the use of machine learning in discovering new pulsars (*Section 5*), and my recommendations to future researchers undertaking a similar project (*Section 6*).

1.3 Background Theory

To answer the question of ”what is a pulsar?” we must first investigate the evolution and death of stars.

A star can form when a cloud of hydrogen gas in the interstellar medium (ISM) collects mass over millions of years; as the mass of the gas cloud increases, its gravitational pull to gather more mass also increases (Maoz 2016). This proto-star will eventually reach a critical mass in which the pressure of gravity within the gas causes enough friction between the gas particles to generate the required heat (thermal pressure) to begin fusing the hydrogen atoms into helium (Maoz 2016). This marks the beginning of the star's main sequence lifetime. Once the star has fused all of the available hydrogen gas in its core, the star will begin fusing helium into carbon and its outer envelope will expand, moving the star into its ”red-giant” phase (Maoz 2016). If the initial mass of the star was greater than 8 times the mass of the Sun (i.e. $8M_\odot$), the star will continue to fuse the elements in its core until it reaches a core of iron. At this point phenomena called nuclear photodisintegration and neutronization occurs, the latter of which causes electrons and photons to combine and form neutrons and anti-neutrinos (Maoz 2016). Neutronization can be shown as:

$$e^- + p \rightarrow n + \nu_e$$

This process removes the electron degeneracy pressure in the core of the star (a pressure which balances the star's gravitational pressure), causing the star to collapse under its own gravity in a timeframe of 0.1 seconds (Maoz 2016). The gravitational collapse stops once the gravitational pressure of the star is balanced by the neutron degeneracy pressure, i.e. the pressure from pushing neutrons together. The remaining star is incredibly dense, with a mass of approximately $1.4M_\odot$ and a radius of around 11km. This is called a neutron star (Maoz 2016).

Prior to the collapse of the star, we can imagine the star to be rotating at an angular velocity of ω_1. We know from the conservation of angular momentum that when the radius of a rotating object decreases, the angular velocity will increase (a spinning ice skater pulling their arms in close increases the speed of their spinning). We can therefore show that the angular velocity of the star after the gravitational collapse, ω_2, is much greater than the prior angular velocity:

$$L_1 = L_2$$

where L is the angular momentum, $L_1 = I_1\omega_1$ and $L_2 = I_2\omega_2$. Therefore:

$$I_1\omega_1 = I_2\omega_2$$

$$\omega_2 = \frac{I_1}{I_2}\omega_1$$

Assuming the star is a sphere, its moment of inertia, I is:

$$I = \frac{2}{5}MR^2$$

where M is the mass of the star and R is the radius of the star. We can thus show:

$$\omega_2 = \frac{\frac{2}{5}MR_1^2}{\frac{2}{5}MR_2^2}\omega_1$$

$$\omega_2 = \left(\frac{R_1}{R_2}\right)^2\omega_1$$

where $R_1 \gg R_2$. We are left with a neutron star with a very large angular velocity. Analogous to the angular velocity of the star, the magnetic field of the star is also amplified. The ionised gas in the iron core of the star, which generates a magnetic field, is compressed by the gravitational collapse, forcing the flux of the magnetic field to be amplified such that the field strength is approximately 10^{10} times stronger in the neutron star compared to during the star's main sequence lifetime (Maoz 2016).

If the rotation of the neutron star is misaligned with the axis of the magnetic field by an angle θ, the spinning magnetic dipole will radiate electromagnetic waves (Maoz 2016). As the neutron star rotates, the radiated electromagnetic waves will periodically sweep across the line of sight of an observer, creating a pulse of light. See Figure 1. We can therefore define a pulsar as a rapidly rotating neutron star that appears to periodically emit electromagnetic waves (Maoz 2016; Lorimer and Kramer 2005; Swainston 2020).

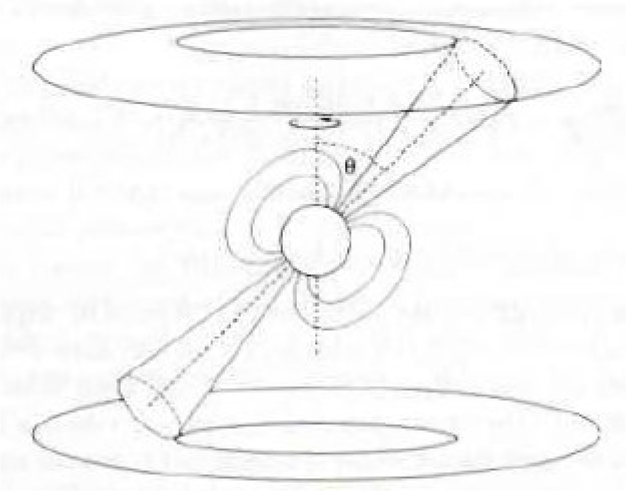

Figure 1: A pulsar (Maoz 2016).

While some pulsars, like the Crab Pulsar (Maoz 2016), emit electromagnetic waves in the visible spectrum and can therefore be detected by an optical telescope, the majority of pulsar emission is invisible to the human eye and requires a radio telescope to be detected.

1.3.1 What is a Pulsar Profile?

1.3.2 What is multi-path scattering?

1.3.3 What is a Pulsar Dispersion Measure?

1.3.4 What is a Radio Telescope and how do they work?

1.3.5 What is the Murchison Widefield Array?

1.3.6 Why are we conducting sky surveys?

1.3.7 How is a sky survey conducted with the Murchison Widefield Array?

1.3.8 What is tied-array beamforming?

1.3.9 How many Pulsar candidates are found in a Murchison Widefield Array sky survey?

1.3.10 How long do sky surveys take?

1.3.11 How much data is created from a sky survey?

1.3.12 What is Radio Frequency Interference?

1.3.13 What is a Signal-to-Noise Ratio?

1.3.14 What is PRESTO and a .PFD file?

1.3.15 What is Machine Learning?

1.3.16 How does Machine Learning work?

1.3.17 Why do we need to use Machine Learning in finding new Pulsars?

1.3.18 Why is this particular Machine Learning Classifier used?

2 Methods

2.1 Developing the Machine Learning Training Dataset

Before a machine learning algorithm can make predictions and classify candidates as a pulsar or a non-pulsar, it must first build a classification model from a training dataset which contains similar data with known positive and negative classifications (Tan et al. 2017; Lyon et al. 2016). For the use case of pulsar classification, the training dataset must contain examples of data from both pulsars and from non-pulsars so that the algorithm can learn how to distinguish between the two classes.

To maximise the accuracy of the machine learning algorithm, the input data (including the training dataset) must be composed of a common group of features that can be determined for each candidate that maximises the differences between a pulsar and a non-pulsar. The candidate features used by Tan et al. 2017 to maximise the differences between pulsar and non-pulsar candidates are:

$$Prof_{\mu}, Prof_{\sigma}, Prof_{S}, Prof_{k} \tag{1}$$

$$DM_{\mu}, DM_{\sigma}, DM_{S}, DM_{k}, DM_{\mu'}, DM_{\sigma'}, DM_{|S'|}, DM_{k'} \tag{2}$$

$$Subband_{\mu}, Subband_{\sigma}, Subband_{S}, Subband_{k} \tag{3}$$

$$Subint_{\mu}, Subint_{\sigma}, Subint_{S}, Subint_{k} \tag{4}$$

Where candidate features are calculated from the: (1) Integrated Pulsar Profile, (2) the Dispersion Measure – Signal-to-Noise Ratio Curve (DM-S/N), (3) the correlation coefficients between each sub-band and the integrated pulsar profile, and from the (4) correlation coefficients between each sub-integration and the integrated pulsar profile. See *Appendix 2A* for formulae to calculate each feature.

To extract the 20 above features from each classification candidate, we can use the software `PulsarFeatureLab` (Lyon et al. 2016).

2.1.1 Pulsar Candidate Feature Extraction

The Python software tool `PulsarFeatureLab` can be used to consume pulsar candidate files of the PRESTO Prepfold `PFD` filetype and output the 20 above features for each candidate into a single file of WEKA Data Mining `ARFF` filetype (Lyon et al. 2016).

To create a closed software environment in which the dependencies of the `PulsarFeatureLab` software are unaffected by the host operating system, a containerised virtual operating system can be created using the free software Docker (`https://docker.com`).

First, a directory to store the Dockerfile and pulsar candidate data is created by completing the following commands in a UNIX terminal:

```
$ mkdir ~/pulsars
$ cd ~/pulsars
$ touch Dockerfile
```

To create the Docker image, the contents of the `Dockerfile` can be edited to contain:

Dockerfile

```
FROM alpine/git:latest as builder
WORKDIR /root/
RUN cd /root/ && git clone --single-branch --branch V1.3.2
    https://github.com/scienceguyrob/PulsarFeatureLab.git &&
    mkdir PulsarFeatureLab/PulsarFeatureLab/Data/IO

FROM python:2.7
WORKDIR /usr/src/app
COPY --from=builder /root/PulsarFeatureLab .
RUN pip install numpy scipy matplotlib astropy
ENTRYPOINT ["python", "./PulsarFeatureLab/Src/
    PulsarFeatureLab.py"]
```

The above `Dockerfile` instructs Docker to:

i. use an image of Alpine Linux with `git` preinstalled to download the `PulsarFeatureLab` software from GitHub (`https://github.com/scienceguyrob/PulsarFeatureLab`);

ii. create a directory inside the downloaded software to store the input and output data;

iii. create a Docker image based on Python 2.7;

iv. transfer the `PulsarFeatureLab` software into the Python 2.7 image; and

v. install `PulsarFeatureLab`'s library dependencies (`NumPy, SciPy, matplotlib` and `astropy`).

The above Docker image can now be built into a container (a virtual operating system) and a directory to hold the input data can be created by running the following commands on a UNIX terminal:

```
$ docker build -t jacobianm/pulsarfeaturelab:1.3.2 .
$ mkdir ~/pulsars/data/pfd
```

Candidate `PFD` files of known pulsars and non-pulsars detected by the Murchison Widefield Array (MWA) provided by N. Swainston can now populate the above created directory, and the following command can be ran to extract the features from the candidates:

```
$ docker run --rm -v ~/pulsars/data/pfd:/usr/src/app/
    PulsarFeatureLab/Data/IO jacobianm/pulsarfeaturelab:1.3.2
     -d "/usr/src/app/PulsarFeatureLab/Data/IO" -c 3 -t 6 -f
    "/usr/src/app/PulsarFeatureLab/Data/IO/output.arff" --
    arff --meta
```

This function instructs Docker to connect the directory containing the `PFD` files to the `PulsarFeatureLab` container's input/output directory and then run the `PulsarFeatureLab` software with arguments stating where the input files are, what filetype they are (`PFD`), which set of features to extract, and where to place the output file.

2.1.2 Creating the Training Dataset

The `output.arff` file created by `PulsarFeatureLab`, contains a set of comma-separated features for each candidate, with an appended '?' character, per file line. Since the class of each candidate is already known, the '?' character on each line can be replaced by a '1' if the candidate is a pulsar, or a '0' if the candidate is a non-pulsar. This signals to the Machine Learning algorithm what a pulsar and a non-pulsar candidate's feature set may appear like.

The edited file can now be renamed and moved with the following command:

```
$ mv ~/pulsars/data/pfd/output.arff ~/pulsars/data/
    trainingSet.arff
```

The Machine Learning Training Dataset has now been created using Murchison Widefield Array data.

2.2 Evaluating the LOTAASClassifier Machine Learning Classification Tool for use with the Murchison Widefield Array

3 Results and Outputs

3.1 Machine Learning Training Dataset

The machine learning training dataset created with candidates detected by the Murchison Widefield Array is as follows:

trainingSet.arff

3.2 Output from LOTAASClassifier with Murchison Widefield Array Data

3.3 Output from PulsarClassifier Machine Learning Classification Tool

4 Discussion

4.1 How effective are the training datasets?

4.2 Why didn't the LOFAR Machine Learning Algorithm work with the MWA?

4.3 What changes to the algorithm were necessary for it to successfully classify Pulsars from the MWA?

5 Conclusions

6 Recommendations

7 References

Lorimer, D. R., and M. Kramer. 2005. *Handbook of Pulsar Astronomy.* Vol. 4. Cambridge, United Kingdom: Cambridge University Press.

Lyon, R. J., B. W. Stappers, S. Cooper, J. M. Brooke, and J. D. Knowles. 2016. "Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach." *Monthly Notices of the Royal Astronomical Society* 459, no. 1 (April): 1104–1123. https://doi.org/10.1093/mnras/stw656.

Maoz, Dan. 2016. *Astrophysics in a Nutshell.* 2nd ed. 64–94. Princeton, New Jersey: Princeton University Press.

Swainston, Nick. 2020. "Investigating the Low-Frequency Population of Southern Pulsars with the MWA: Milestone 2," International Centre for Radio Astronomy Research.

Tan, C. M., R. J. Lyon, B. W. Stappers, S. Cooper, J. W. T. Hessels, V. I. Kondratiev, D. Michilli, and S. Sanidas. 2017. "Ensemble candidate classification for the LOTAAS pulsar survey." *Monthly Notices of the Royal Astronomical Society* 474 (4): 4571–4583. https://doi.org/10.1093/mnras/stx3047.

8 Appendices

8.1 Pulsar Feature Lab

8.2 Pulsar Classifier