Detection of Pulsars from Radio Telescope Data using Linear and Non-Linear Machine Learning Models

Federico Boscolo
s294908@studenti.polito.it

Antonio De Cinque
s303503@studenti.polito.it

27 June 2022

Abstract

Pulsars are a kind of Neutron star that, in the first phases of their formation, rapidly rotate emitting radio frequencies at regular intervals, detectable from Earth using large radio telescopes. Finding a pulsar involves looking for periodic radio signals, but the vast majority of such signals are generated by radio interference, making pulsars very hard to find. This report aims to find an effective Machine Learning approach to the classification of Pulsars among periodic signals captured by telescopes. Different models are examined for the purpose of classification, and linear models prove to be the most effective.

Contents

1 Introduction

1.1 A brief introduction to Pulsars

Pulsars are a rare type of Neutron star that produce radio emission detectable here on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter.

The first pulsar was discovered by Jocelyn Bell in 1967, while examining data from a newly built radio telescope. Initially, the data was dismissed as radio interference but was then measured with another telescope, confirming the existence of a rotating radio source in the universe, later confirmed to be a neutron star. Since then, many pulsars have been found throughout the universe, which led to scientist being able to study neutron stars for the first time. This allowed researchers to get a glimpse at the behavior of matter at nuclear density. Other applications for pulsars include maps and clocks thanks to their very precise periods of rotation.

1.2 Data Set Information

HTRU2 is a dataset which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey (South).

The data set contains 16,259 spurious examples caused by RFI/noise, and 1,639 real pulsar examples, for a total of 17,898 total examples. These examples have all been checked by human annotators.

The dataset has been split into Training and Evaluation (Test) data. The training set used for this application contains 8108 spurious examples and 821 real pulsar examples. The evaluation set contains 8151 spurious examples and 818 real pulsar examples. The samples are encoded as follows: the list of features is stored on a single line of a file separated by a comma, and the class label lies at the end of the line.

1.3 Attribute Information

Each candidate is described by 8 continuous variables. The first four are simple statistics obtained from the integrated pulse profile. This is an array of continuous variables that describe a longitude-resolved version of the signal that has been averaged in both time and frequency. The remaining four variables are similarly obtained from the DM-SNR curve. These are summarised below:

Feature Name	Range	Type
Mean of the integrated profile	6.1796875 - 186.023437	Real
Standard deviation of the integrated profile	24.77204176 - 98.77891067	Real
Excess kurtosis of the integrated profile	-1.730781724 - 8.069522046	Real
Skewness of the integrated profile	-1.791885981 - 68.10162173	Real
Mean of the DM-SNR curve	0.213210702 - 209.3001672	Real
Standard deviation of the DM-SNR curve	7.370432165 - 110.6422106	Real
Excess kurtosis of the DM-SNR curve	-2.812353306 - 34.53984419	Real
Skewness of the DM-SNR curve	-1.976975603 - 1191.000837	Real

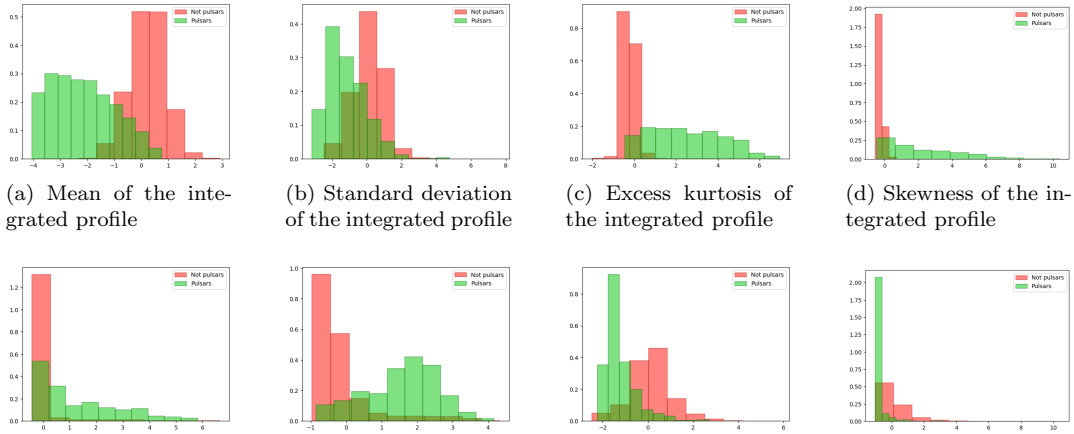
Table 1: Feature Description in training set

The numerical distribution of features turns out to be extremely unbalanced. The variation between the eight features is too large. Such distributions are not optimal for our purposes. Without normalization, training is difficult to converge and overflow problems may arise.

Therefore, training data has been pre-processed with Z-normalization

$$Z = \frac{x - \mu}{\sigma}$$

where x the original feature vector, μ is the mean of that feature vector, and σ is the standard deviation.



(a) Mean of the integrated profile

(b) Standard deviation of the integrated profile

(c) Excess kurtosis of the integrated profile

(d) Skewness of the integrated profile

(e) Mean of the DM-SNR curve

(f) Standard deviation of the DM-SNR curve

(g) Excess kurtosis of the DM-SNR curve

(h) Skewness of the DM-SNR curve

Figure 1: Z-normalized HTRU2 training set features.

In Figure 1, red histograms refer to "Not pulsars" and green histograms refer to "Pulsars". A preliminary analysis of the training data shows that the normalized features seem well distributed, without evident outliers. Therefore, no further pre-processing has been considered.

1.4 Feature distribution

Second paragraph

References

[1] R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles, *Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach*, MNRAS, 2016.

[2] R. J. Lyon, *"PulsarFeatureLab"*, 2015,
https://dx.doi.org/10.6084/m9.figshare.1536472.v1.

[3] Nora Roberts, D. R. Lorimer, M. Kramer, *Handbook of Pulsar Astronomy (illustrated, herdruk ed.)*, Cambridge University Press, 2015.