Minimal Feature Pulsar classification using Machine learning

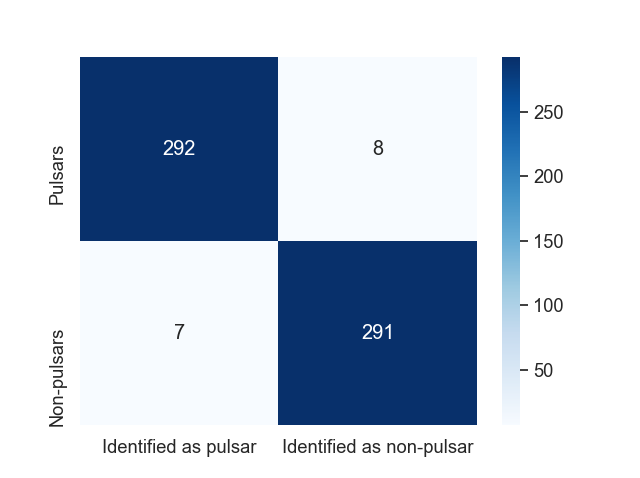
*- M Abhijith Bhuvandas*

The objective of this project is to create a classifier that can read the data files of stars and predict whether they are pulsars or not with acceptable errors using the least amount of features.

For this, an AdaBoost classifier was used. AdaBoost is an ensemble learning method created to increase the efficiency of binary classifiers. It uses an iterative approach to learn from the mistakes of weak classifiers, and turn them into strong ones.

For features, *Log of the signal-to-noise ratio* and *Skew of the folded proﬁle* were used.

* SNR is a measure of the purity of the signal.
* The pulse profile of the star can be folded and it’s skew can be calculated to give a numerical measure of the radio pulse profile.

Even on a dataset of only 2392 stars, of which only half are pulsars, the classifier is able to obtain a recall rate of 97.7% with a low False Positive Rate of 2.7%.

Problem

Create a machine learning model which can distinguish pulsars from non-pulsars using the least number of parameters and still give high accuracy.

Methods

1. Selecting parameters

There are two types of data usually recorded in a stars file.

* The empirical data recorded directly from the stars, such as Signal to Noise ratio and Intrinsic equivalent duty cycle of the profile.
* The statistical data based on statistical analysis of the data, such as the mean, standard deviation and kurtosis of the folded profile.

From these, a mix of both were found to be most representative.

1. Selecting Model

Ensemble learning has been shown to be superior in classification problems, so types of ensemble learning algorithms were studied. The boosting algorithms were used to iteratively adjust the weights of an observation based on the last classification.

AdaBoost is implemented by sequentially learning a series of weak learners and adaptively adjust the weights of the samples based on the performance of the weak learners. By trial and error, this was shown to be the best algorithm for fast learning and high accuracy for low parameter counts.

Results and Discussion

By trial and error, the ideal combination of parameters and model was found.

Model: AdaBoost classifier with Decision Tree base estimator, 1 layer max depth and 250 estimators

Parameters: *Logarithm of Signal to noise ratio* and *Skew of the folded profile*

Only two parameters were required for high accuracy.

From the test set of 299 pulsars and 298 non-pulsars, the classifier could reach a recall rate of 97.659% with a False Positive Rate of only 2.676%.

This was with training on a database of 1795 stars, with only 897 pulsars.

For a larger dataset, higher accuracy and lower False Positive Rate are expected.

A paper[[1]](https://arxiv.org/abs/2002.08519), submitted on 20th of February could reach a recall of 99% and a FPR of 0.548% on the same data and model by using SMOTE algorithm[[2]](https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879) to make up for the lack in number of pulsar data compared to non-pulsars.

References

[1] : Lin, Haitao, Xiangru Li, and Ziying Luo. “Pulsars Detection by Machine Learning with Very Few Features.” Monthly Notices of the Royal Astronomical Society 493.2 (2020): 1842–1854. Crossref. Web.

[2] : N. S. Altman (1992) An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, The American Statistician, 46:3, 175-185, DOI: 10.1080/00031305.1992.10475879

[3] : Jia, Peng, Qiang Liu, and Yongyang Sun. “Detection and Classification of Astronomical Targets with Deep Neural Networks in Wide-Field Small Aperture Telescopes.” The Astronomical Journal 159.5 (2020): 212. Crossref. Web.