PULSARS NEUTRON STARS CLASSIFIER

CASPTONE REPORT

MACHINE LEARNING ENGINEER NANODEGREE

RAFAEL BARRETO

1. DEFINITION

1.1 PROJECT OVERVIEW

1.1.1 PULSARS

According to reference [1]:

"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.

As pulsars rotate, their emission beam sweeps across the sky, and when this crosses our line of sight, produces a detectable pattern of broadband radio emission. As pulsars rotate rapidly, this pattern repeats periodically. Thus pulsar search involves looking for periodic radio signals with large radio telescopes.

Each pulsar produces a slightly different emission pattern, which varies slightly with each rotation. Thus a potential signal detection known as a 'candidate', is averaged over many rotations of the pulsar, as determined by the length of an observation. In the absence of additional info, each candidate could potentially describe a real pulsar. However in practice almost all detections are caused by radio frequency interference (RFI) and noise, making legitimate signals hard to find."

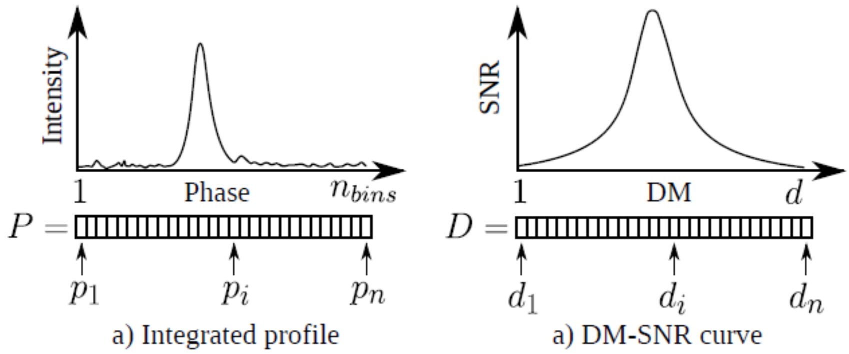
1.1.2 PROPOSED FEATURES

There are two components of the typical pulsar candidate and were extracted eight features, four feature for each component. These components features are simple statistics obtained respectively from the **Integrated Pulse Profile** and the **DM-SNR curve** (Dispersion Measure Signal-to-noise ratio).

These features are in agreement with the feature design criteria adopted [3]:

- Minimize biases & selection effects.
- Be survey-independent for data interoperability.
- Be implementation-independent, with concise mathematical definitions allowing for reproducibility.
- Be evaluated using a statistical framework that enables comparison and reproducibility.
- Guard against high dimensionality.

– Be accompanied by public feature generation code, to facilitate co-operation and feature improvement.
– Be supplied in a standard data format.
– Be evaluated on multiple data sets to ensure robustness.



t

Fig 1: Diagram showing which components of the standard pulsar candidate new features were extracted from. Plot a) shows the integrated pulse profile, plotted from the vector P. Plot b) shows the DM-SNR curve obtained from the vector D. **[3]**.

Feature	Description	Definition
$Prof_\mu$	Mean of the integrated profile P.	$\frac{1}{n}\sum_{i=1}^{n} p_i$
$Prof_\sigma$	Standard deviation of the integrated profile P.	$\sqrt{\frac{\sum_{i=1}^{n}(p_i-\bar{P})^2}{n-1}}$
$Prof_k$	Excess kurtosis of the integrated profile P.	$\frac{\frac{1}{n}(\sum_{i=1}^{n}(p_i-\bar{P})^4)}{(\frac{1}{n}(\sum_{i=1}^{n}(p_i-\bar{P})^2))^2}-3$
$Prof_s$	Skewness of the integrated profile P.	$\frac{\frac{1}{n}\sum_{i=1}^{n}(p_i-\bar{P})^3}{(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i-\bar{P})^2})^3}$
DM_μ	Mean of the DM-SNR curve D.	$\frac{1}{n}\sum_{i=1}^{n} d_i$
DM_σ	Standard deviation of the DM-SNR curve D.	$\sqrt{\frac{\sum_{i=1}^{n}(d_i-\bar{D})^2}{n-1}}$
DM_k	Excess kurtosis of the DM-SNR curve D.	$\frac{\frac{1}{n}(\sum_{i=1}^{n}(d_i-\bar{D})^4)}{(\frac{1}{n}(\sum_{i=1}^{n}(d_i-\bar{D})^2))^2}-3$
DM_s	Skewness of the DM-SNR curve D.	$\frac{\frac{1}{n}\sum_{i=1}^{n}(d_i-\bar{D})^3}{(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i-\bar{D})^2})^3}$

Fig. 2: Table 8.1: The eight features derived from the integrated pulse profile (folded profile) $P=\{p_1,\ldots,p_n\}$ and the DM-SNR curve $D=\{d_1,\ldots,d_n\}$. For both

P and D, all p_i and $d_i \in \mathbb{N}$ for $= 1, \dots, n$. $\tilde{P}$ and $\tilde{D}$ are the means of the integrated profile and DM-SNR curve respectively.

1.2 PROBLEM STATEMENT

The problem that is to be solved is classifying Pulsar candidates collected during the HTRU survey. Pulsars are a type of star, of considerable scientific interest. Candidates must be classified in to pulsar and non-pulsar classes to aid discovery.

A model should be trained to classify Pulsar candidates from HTRU2 Dataset. The algorithm must do a binary classification task.

1.3 EVALUATION METRICS

The evaluation of the algorithm will be made by the metrics: Classification Accuracy and F- beta score. These two metrics are very common on binary class classification projects.

The accuracy (ACC) is the proportion of the total number of predictions that were correct and F-beta score is a statistical method for determining accuracy accounting for both precision and recall.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

The overall accuracy of the system is:

$$ACC = \frac{\mathrm{T}}{\mathrm{T} + \mathrm{F}}$$

Where:

T – True values

F – False values

2. ANALYSIS

2.1 DATA EXPLORATION AND VIZUALIZATION

This project will use the HTRU2 Dataset that is available on [1] and [2].The dataset contains 16,259 spurious examples caused by RFI/noise, and 1,639 real pulsar examples. These examples have all been checked by human annotators. Each candidate is described by 8 continuous variables. The first four are simple statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables that describe a longituderesolved version of the signal that has been averaged in both time and frequency [3]. The remaining four variables are similarly obtained from the DM-SNR curve. These are summarized below:

1. Mean of the integrated profile.

2. Standard deviation of the integrated profile.

3. Excess kurtosis of the integrated profile.

4. Skewness of the integrated profile.

5. Mean of the DM-SNR curve.

6. Standard deviation of the DM-SNR curve.

7. Excess kurtosis of the DM-SNR curve.

8. Skewness of the DM-SNR curve.

HTRU 2 Summary

- 17,898 total examples.
- 1,639 positive examples.
- 16,259 negative examples.

The data used is presented in CSV format. Candidates are stored in both files in separate rows. Each row lists the variables first, and the class label is the final entry. The class labels used are 0 (negative) and 1 (positive).

The data contains no positional information or other astronomical details. It is simply feature data extracted from candidate files using the PulsarFeatureLab tool [2].

FEATURE'S DISTRIBUTION

The graphs of all features distribution for every feature were generated. Fig. 3 shows four examples:

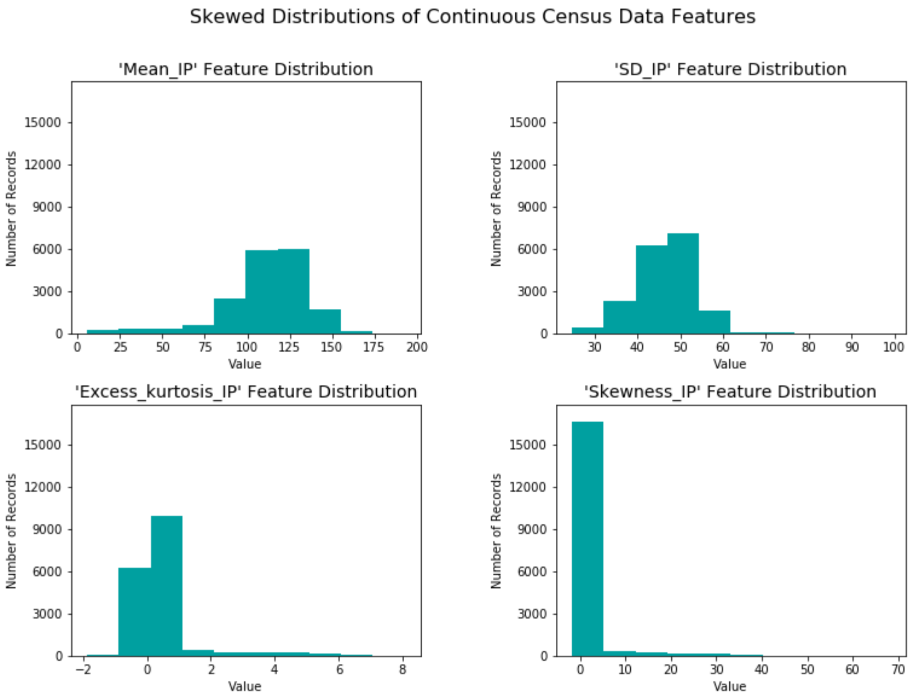


Fig. 3: Examples of Features Distribution

RELATION BETWEEN PAIRS OF FEATURES

For every possible pair of feature (total of 28 pairs) were plotted a graph. For almost all 28 graphs is easily observed pattern between the pairs of the features about the positive and negatives class. Below is showed two examples:

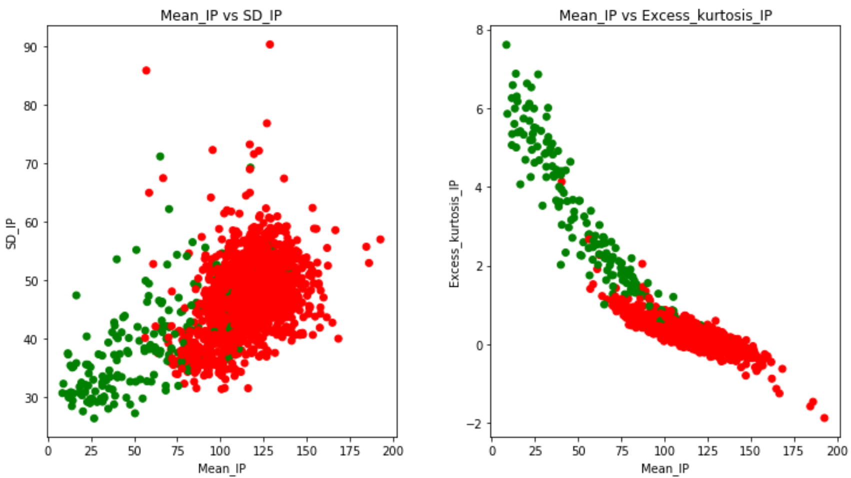


Fig. 4: Examples of relation between pairs of features

2.2 ALGORITHMS AND TECHNIQUES

Approaches capable of doing a binary classification is based on different algorithms as Support Vector Machines (SVM), Decision Trees, Gaussian Naive Bayes (GaussianNB), Logistic Regression, Ensemble Methods, etc.

In this project was tested three algorithms for training and evaluation: SVM, Decision Trees and GaussianNB. In the final the Random Forest method was used to extracting feature Importance to optimize even more the model.

2.3 BENCHMARK MODEL

The Benchmark used in this project will be the results achieved in thesis: "WHY ARE PULSARS HARD TO FIND?" written by James Robert Lyon in 2016. This thesis was submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Engineering and Physical Sciences. Below is showed a picture of the Table 8.6 extracted:

Dataset	Algorithm	G-Mean	F-Score	Recall	Precision	Specificity	FPR	Accuracy
HTRU 1	C4.5	0.962*	0.839*	0.961	0.748	0.962	0.038	0.962
	MLP	**0.976**	0.891	**0.976**	0.820	0.975	0.025*	0.975
	NB	0.925	0.837*	0.877	0.801	0.975	0.025*	0.965
	SVM	0.967	0.922	0.947	0.898	0.988	0.012	0.984
	GH-VFDT	0.961*	**0.941**	0.928	**0.955**	**0.995**	**0.005**	**0.988**
HTRU 2	C4.5	0.926	0.740	**0.904**	0.635*	0.949*	0.051*	0.946*
	MLP	**0.931**	0.752	0.913	0.650*	0.950*	0.050*	0.947*
	NB	0.902	0.692	0.863	0.579	0.943	0.057	0.937
	SVM	0.919	0.789	0.871	0.723	0.969	0.031	0.961
	GH-VFDT	0.907	**0.862**	0.829	**0.899**	**0.992**	**0.008**	**0.978**
LOTAAS 1	C4.5	0.969	0.623	0.948	0.494	0.991	0.009	0.990
	MLP	**0.988**	0.846*	**0.979**	0.753	0.998	0.002	0.997*
	NB	0.977	0.782	0.959	0.673	0.996	0.004	0.996
	SVM	0.949	**0.932**	0.901	**0.966**	**0.999***	**0.001***	**0.999**
	GH-VFDT	0.888	0.830*	0.789	0.875	**0.999***	**0.001***	0.998*

Table 8.6: Results obtained on the three test data sets. Bold type indicates the best performance observed. Results with an asterisk indicate no statistically significant difference between the algorithms at the $\alpha = 0.01$ level.

Fig 5. Image of the Table 8.6, Chapter 8: "New Candidate Features", page 232 of [3].

3. METHODOLOGY

3.1 DATA PREPROCESSING

The HTRU2 Dataset is composed only by numerical and continuous features. Therefore the only preprocessing applied was the numerical normalization between 0 and 1.

3.2 EVALUATING MODEL PERFORMANCE

Three supervised learning models were chosen: **Gaussian Naive Bayes, Support Vector Machines** and **Decision Trees.** For each model the data was trained with different samples sizes(1%, 10%, and 100%). Fig. 6 shows the results of time of processing and the metrics evalutation:

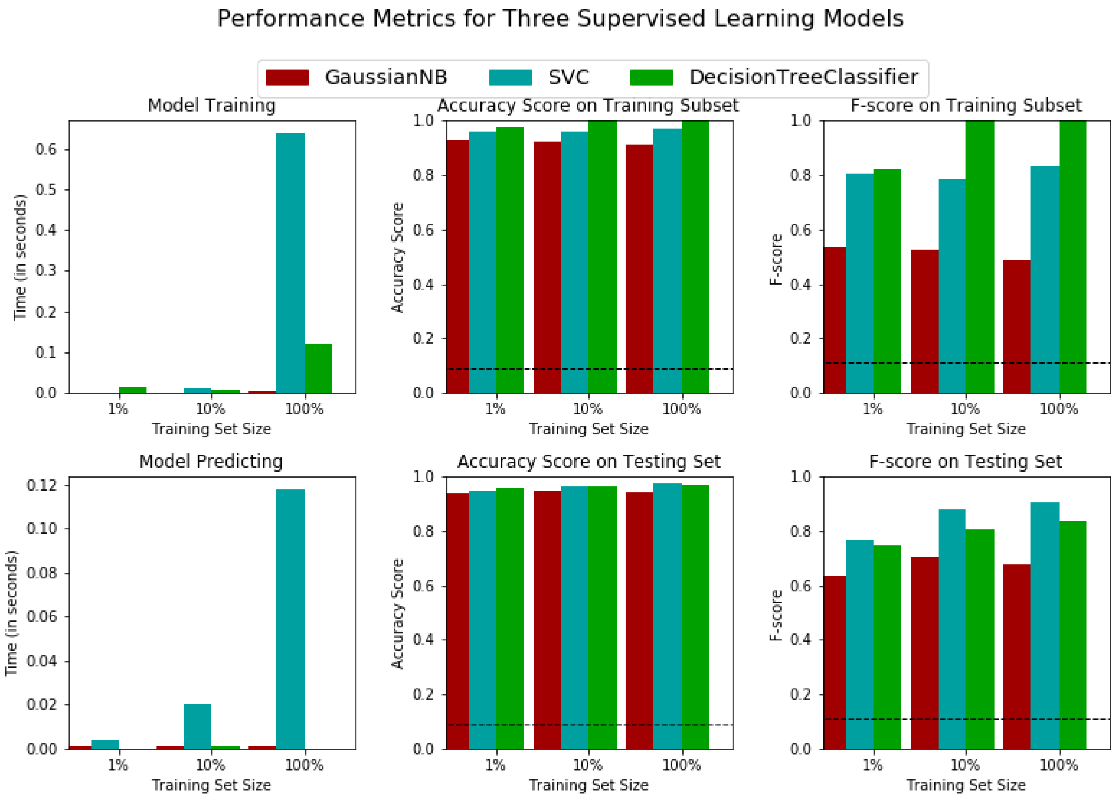


Fig. 6: Performance Metrics for Three Supervised Learning Models

GRID SEARCH

After this initial result, an exhaustive search was applied (GridSearchCV) over specified parameter values for the SVM estimator, with the parameters “**C**” and “**degree**”. C Is the penalty parameter of the error term and “degree” is the degree of the polynomial kernel function (‘poly’).

FEATURE IMPORTANCE

The classifier Random Forest with attribute "feature_importance" was chosen and fit to the training set to determine the top 5 most important features for the dataset.

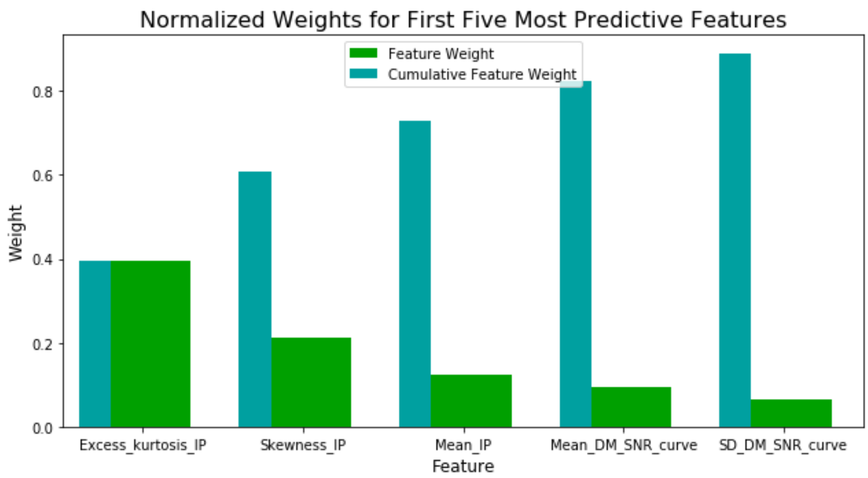


Fig. 7: Graph of the normalized Weights for First Five Most Predictive Features

4. RESULTS

All results are shown below:

Metric	Benchmark Predictor	Unoptimized Model	Optimized Model	Optimized Model with Feature Selection
Accuracy Score	0.0916	0.9701	0.9721	0.9732
F-score	0.2917	0.8893	0.8964	0.9003

Final result in comparison to the benchmark adopted:

Metric	Benchmark Dr. James Robert Lyon PHD Thesis	Optimized Model with Feature Selection
Accuracy Score	0.961	0.9732
F-score	0.789	0.9003

This work achieved better results for both metrics chosen in comparison the benchmark used. The Accuracy Score achieved 0.9732 and the F-score 0.9003. The optimized model and the reduction of the data due the feature relevance improved a little the metrics, with better effect on F-score.

4. CONCLUSION

In this project was shown that typical pulsar candidates can be predicted using machine learning techniques with very good results. Between the supervised learning models used the Support Vector Machine model achieved the better results in binary classification task without overfitting.

5. REFERENCES

[1] **HTRU2 Data Set.** Available at: https://archive.ics.uci.edu/ml/datasets/HTRU2

[2] **HTRU2 Data Set.** Available at: https://figshare.com/articles/HTRU2/3080389

[3] R. J. Lyon, **"Why Are Pulsars Hard To Find?"**, *PhD Thesis, University of Manchester*, 2016.

[4] M. J. Keith et al., **"The High Time Resolution Universe Pulsar Survey - I. System Configuration and Initial Discoveries"**, *Monthly Notices of the Royal Astronomical Society, vol. 409, pp. 619-627. DOI: 10.1111/j.1365-2966.2010.17325.x* , 2010.