Assignment [6]: a mathematical essay on support vector machine

Mustefa Abrahim
Deprt. Data Science
Indian Institute of Technology Madras (IITM)
ge22m014@smail.iitm.ac.in

***Abstract*—The main goal of this paper is to predict whether a star is a pulsar or not based on the statistical properties of the integrated profile and DM-SNR curve. A support vector classifier is used to model these factors' importance and predict the star's label.**

Index Terms—Support Vector Classifier, Visualization, Pulsar, prediction

I. Introduction

Pulsars are a rare type of Neutron star that emits radio emissions that can be detected on Earth. They hold great scientific promise as probes of space-time, the interstellar medium, and states of matter. To facilitate rapid analysis, machine learning tools are now being used to automatically label pulsar candidates. SVM, or Support Vector Machine, is a linear model that can be used to solve classification and regression problems. It can solve linear and nonlinear problems and is useful for a wide range of practical applications. The concept of SVM is straightforward: The algorithm draws a line or a hyperplane to divide the data into classes. Support Vector Machines are used in this study to model the category of stars based on the statistical properties of the integrated profile and DM-SNR curve. We begin by gathering, cleaning, and preparing the data before performing exploratory analysis. Finally, we create statistical models and visualisations to provide quantitative and visual evidence of the observed relationships. In the following section, we will discuss the key principles underlying the Support Vector classifier. Section 3 discusses the insights and observations derived from the data and models.

II. Support Vector Machine

Support Vector Machine, or SVM, is a popular Supervised Learning algorithm for Classification and Regression problems. However, it is primarily used for Classification problems in Machine Learning.

The SVM algorithm's goal is to find the best line or decision boundary that can divide n-dimensional space into classes so that we can easily place new data points in the correct category in the future. This best decision boundary is referred to as a hyperplane.

SVM selects the extreme points/vectors that aid in the formation of the hyperplane. These extreme cases are known as support vectors, and the algorithm is known as the Support Vector Machine.

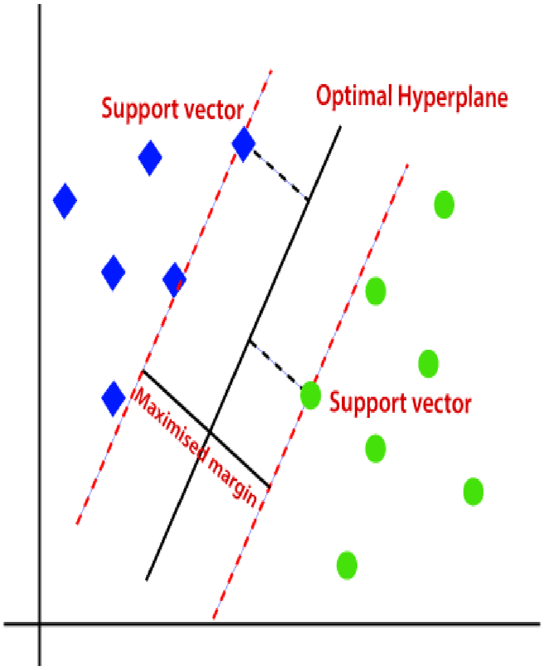
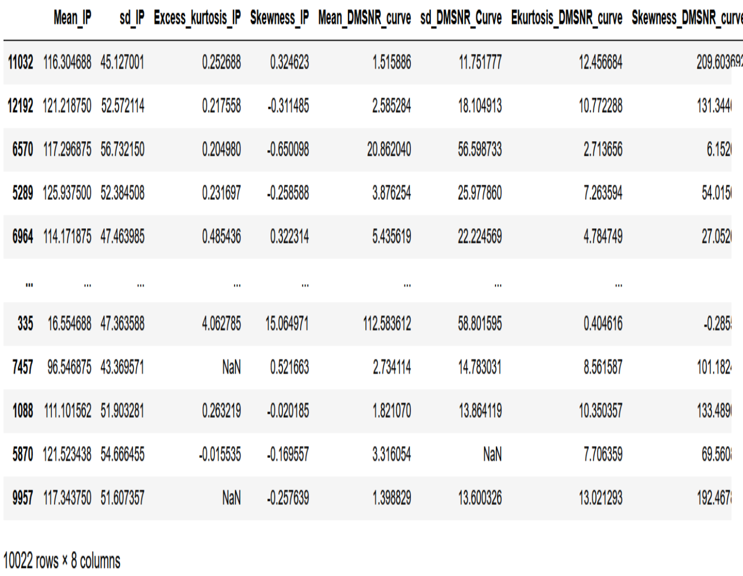


Figure 1: SVM models

In below Table, we have a description of the features and the count, min, mean, median, etc. of the first few features. In figure 4, we plot the distributions of a features based on the target class. In each of these, we observe that the distributions of "not pulsar" and "pulsar" have different means and variances. The pulsar stars tend to have a higher variance and smaller peaks.



	Mean_IP	sd_IP	Excess_kurtosis_IP	Skewness_IP	Mean_DMSNR_curve	sd_DMSNR_Curve	Ekurtosis_DMSNR_curve	Skewness_DMSNR_curve
11032	116.304688	45.127001	0.252688	0.324623	1.515886	11.751777	12.456684	209.603692
12192	121.218750	52.572114	0.217558	-0.311485	2.585284	18.104913	10.772288	131.344
6570	117.296875	56.732150	0.204980	-0.650098	20.862040	56.598733	2.713656	6.152
5289	125.937500	52.384508	0.231697	-0.258588	3.876254	25.977860	7.263594	54.015
6964	114.171875	47.463985	0.485436	0.322314	5.435619	22.224569	4.784749	27.052
...	...	...	...	...	...	...	...	
335	16.554688	47.363588	4.062785	15.064971	112.583612	58.801595	0.404616	-0.285
7457	96.546875	43.369571	NaN	0.521663	2.734114	14.783031	8.561587	101.182
1088	111.101562	51.903281	0.263219	-0.020185	1.821070	13.864119	10.350357	133.489
5870	121.523438	54.666455	-0.015535	-0.169557	3.316054	NaN	7.706359	69.560
9957	117.343750	51.607357	NaN	-0.257639	1.398829	13.600326	13.021293	192.467

10022 rows × 8 columns

Figure 1: SVM models

III. Data

Pulsars are a rare type of Neutron star that emits radio emissions that can be detected on Earth. They hold great scientific promise as probes of space-time, the interstellar medium, and states of matter. To facilitate rapid analysis, machine learning tools are now being used to automatically label pulsar candidates. The main task is to predict whether a star will start a pulsar or not. Eight continuous variables and one class variable are used to describe each candidate. The first four are straightforward statistics derived from the integrated pulse profile (folded profile). This is an array of continuous variables that describes a longitude-resolved version of the signal that has been time and frequency averaged. The remaining four variables are derived in the same way from the DM-SNR curve.

A. Attribute Details:

Eight continuous variables and one class variable are used to describe each candidate. The first four are straightforward statistics derived from the integrated pulse profile (folded profile). This is a set of continuous variables that describe a longitude-resolved version of the signal that has been time and frequency averaged (see [3] for more details). The remaining four variables are derived in the same way from the DM-SNR curve.

- The integrated profile's mean.
- The integrated profile's standard deviation.
- The integrated profile has excessive kurtosis.
- The integrated profile's skewness.
- The DM-SNR curve's mean.
- The DM-SNR curve's standard deviation.
- The DM-SNR curve has excessive kurtosis.
- The DM-SNR curve is skewed.
- Class

IV. Problem

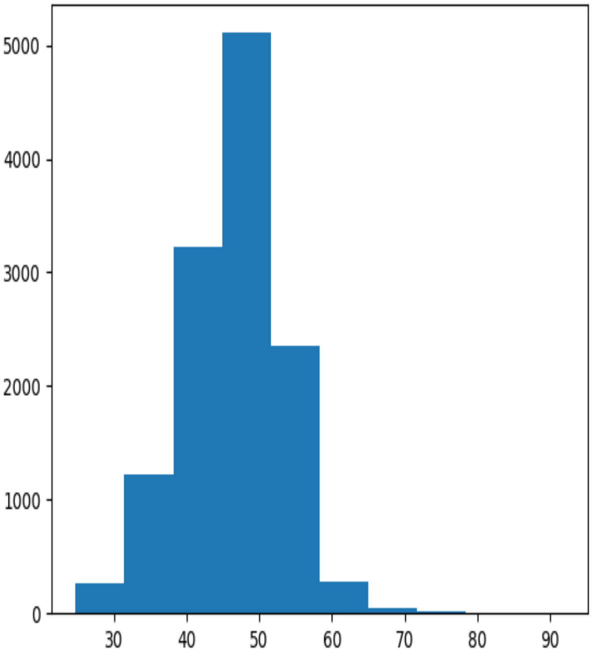
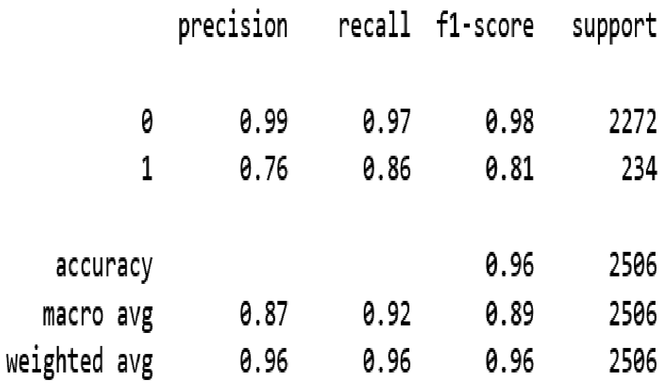


Figure 2: Standard deviation of IP



```
              precision    recall  f1-score   support

           0       0.99      0.97      0.98      2272
           1       0.76      0.86      0.81       234

    accuracy                           0.96      2506
   macro avg       0.87      0.92      0.89      2506
weighted avg       0.96      0.96      0.96      2506
```

Figure 3: Classification Report

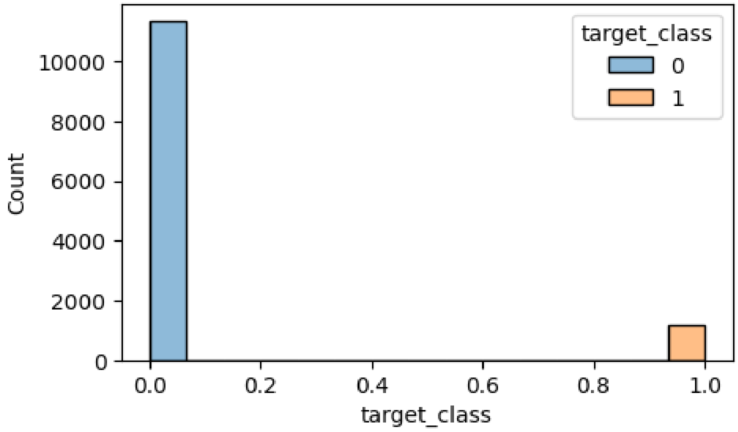


Figure 4: Target_class

V. Conclusions

In this study, we observe the factors that decide whether a star is a pulsar or not. We observed that in the case of the integrated profile as well as the DM-SNR curve, pulsars tend to have higher skew and kurtosis and lower mean and standard deviation. In the future, class imbalance handling techniques could be implemented.

References

[1] . Sorower MS. A literature survey on algorithms for multi-label learning. Oregon State University, Corvallis. 2010 Dec;18.

[2] . Utku A, Hacer (Uke) Karacan, Yildiz O, Akcayol MA. Implementation of a New Recommendation System Based on a Decision Tree Using Implicit Relevance Feedback. JSW. 2015 Dec 1;10(12):1367-74.

[3] . Gershman A, Meisels A, Lüke KH, Rokach L, Schclar A, Sturm A. A Decision Tree Based Recommender System. InIICS 2010 Jun 3 (pp. 170-179).