Multiescala, Machine Learning y Métodos QSAR aplicados a biomoléculas

Master Studies in Theoretical Chemistry and Computational Modelling

Label of Pulsar Star Candidates using Support Vector Machine Algorithms

Report of Prof. Francesco Santini's homework

Author:
José Antonio Quiñonero Gris

November 15, 2023

1 Introduction

Support Vector Machine (SVM), or support-vector network, is a supervised machine learning algorithm used for two-group classification and regression problems, which conceptually implement the following idea: input vectors are non linearly mapped to a very high-dimension feature space, in which a linear decision surface is constructed [1].

It was originally implemented for the restricted case where the training data can be separated without errors, but it has been extended to non-separable training data, as is the case of the data used in this report. After testing linear and non-linear kernels [2, 3], the Gaussian Radial Basis Function (RBF) [4] kernel was used to construct the SVM.

2 Data

The chosen dataset is HTRU2 [5], which can be found in the UCI Machine Learning Repository and in kaggle. HTRU2 is a data set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey (South) [6].

Pulsars are a rare type of Neutron star that, as they (rapidly) rotate, produce a detectable pattern of broadband radio emission, which repeats periodically. 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. 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 [6–8]. Here is when machine learning algorithms come in hand to automatically label pulsar candidates, which is the purpose of this report.

The dataset contains a total of 17898 samples, where 16259 are spurious examples caused by RFI/noise, and 1639 are real pulsar examples. Each candidate is described by 8 continuous variables, which correspond to simple statistic obtained from the integrated pulse profile and from the dispersion signal-to-noise ratio (DM-SNR) curve. In the last entry, the class labels are 0 (negative) and 1 (positive):

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.

Summary of the HTRU2 dataset samples:

- 17898 total.
- 1639 (9.16%) positive.
- 16259 (90.84%) negative.

In order to visualize the data, the pairplot of the dataset is shown in 1.

Looking at the positive ($\approx 10\%$) *vs* negative ($\approx 90\%$) ratio and the pairplot in fig. 1, the raw data is imbalanced. It can be confirmed simply by eye from the Principal Component Analysis (PCA) plotted

in fig. 2.

3 Program

4 Conclusions

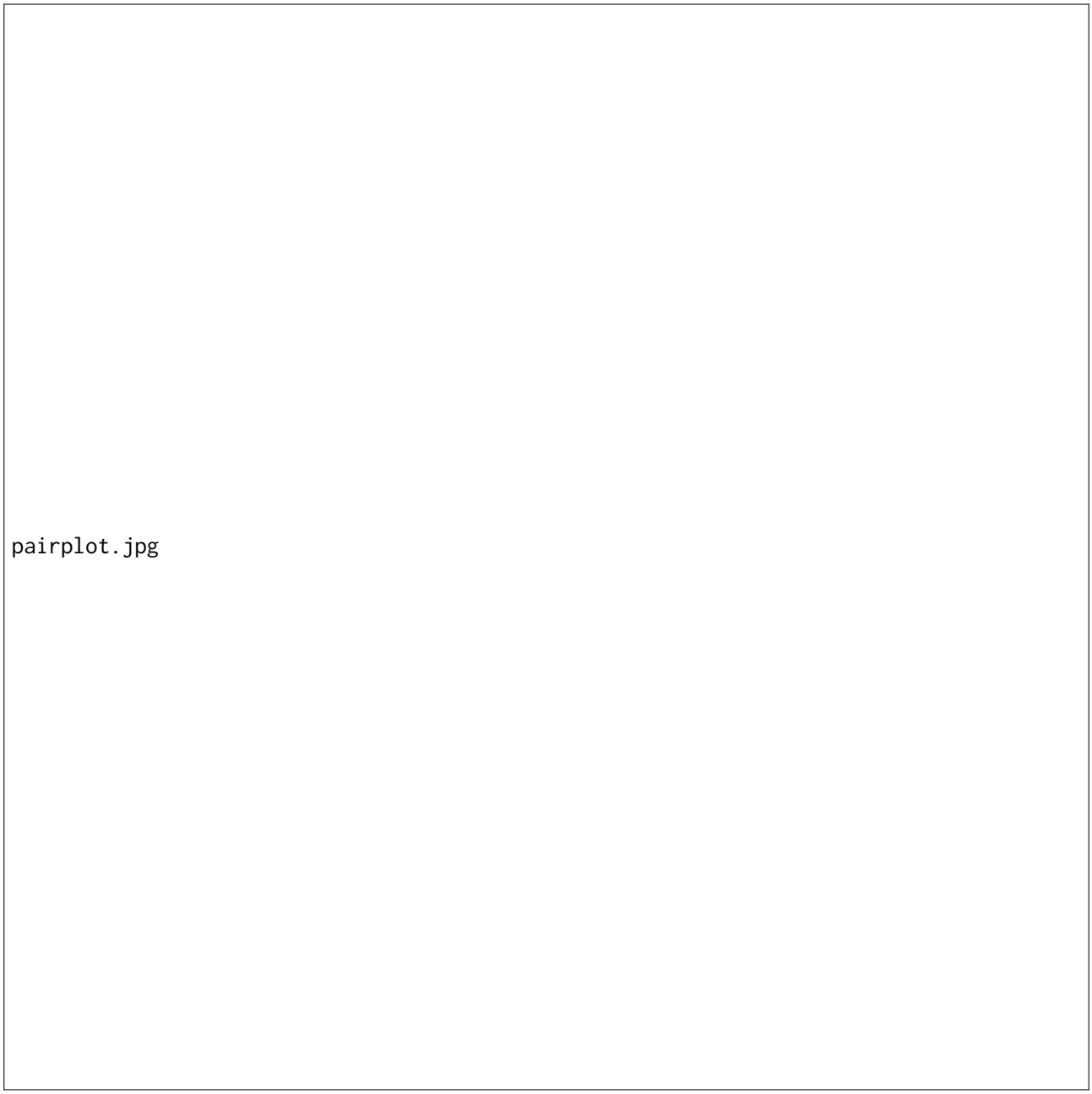

Figure 1: Pairplot of the HTRU2 dataset.

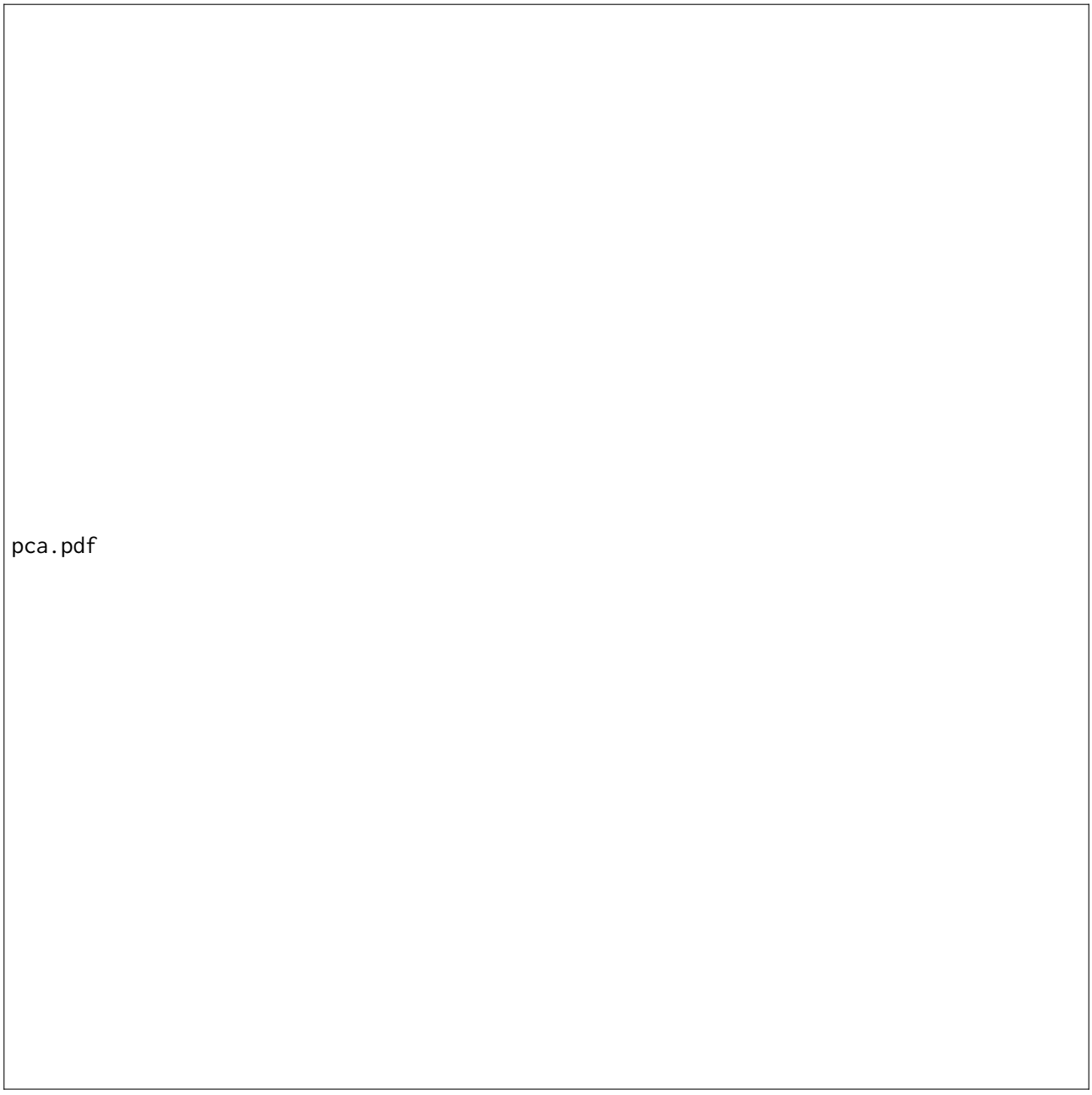

Figure 2: PCA plot of the HTRU2 dataset.