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 16, 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]. For more information about pulsars, see the Appendix.

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 statistics 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 **??**.

Looking at the positive (≈ 10%) *vs* negative (≈ 90%) ratio and the pairplot in **??**, the raw data is imbalanced. It can be confirmed simply by eye from the Principal Component Analysis (PCA) plotted in **??**. Then, the data needs to be balanced beforehand, so that the trained model is not biased across the majority class, reducing the possibility of producing false positives.

3 Preprocessing of the data & training of the SVM

In order to compare the performance of the model trained with imbalanced and balanced data, the raw data is splitted into training (80%) and test (20%) sets. Therefore, the training set data is balanced using three methods:

Table 1: Metrics for the different classifiers.

Classifier	Accuracy	Precision	Recall
Imbalanced	0.9740	0.9321	0.7671
Oversampled	0.9595	0.7196	0.9006
SMOTE	0.9578	0.7090	0.9006
Undersampled	0.9534	0.6841	0.8944
NearMiss	0.7578	0.2560	0.8882
Class-weight	0.9609	0.7321	0.8913

- Oversampling (plus the SMOTE method):
- Undersampling (plus the NearMiss method):
- Class-weight:

The SVM is trained, for each one of the training sets, and tested with the test set. Some simple metrics are calculated [1] in order to compare the performance of the different methods, and are listed in table 1.

Then, the program chooses the best classifier (as the one whose mean value of the three metrics is the largest) and stores it for the problem data. In the case listed in table 1, the chosen classifier is class-weight, with a high accuracy of $\approx 96\%$.

4 Problem data & Results

The chosen trained SVM classifier can be used to classify possible pulsar candidates. As an example, 3 candidates from the LOFAR Tied-Array All-sky Survey (LOTAAS1) [7] dataset, listed in the `problem.csv` file, are classified with this model. The results are shown in table 2.

5 Conclusions

Support Vector Machines are powerful methods for classification and regression problems. One of its main advantages is its ability to handle complex and high-dimensional data, achieve high accuracy, and be less sensitive to outliers than other algorithms. Also, its capacity control and ease of changing the implemented decision surface makes it an extremely powerful and universal learning machine [1].

With this report, the very difficult and tedious task of labeling pulsar star candidates is solved and automatized with a very simple code and very few computation time[2] , resulting in a very robust and easily extensible model.

Table 2: Results for the problem data.

[1] The results vary slightly at every run of the program, due to the randomization algorithms used in the balancing methods.

[2] The dataset used is not very large, so that the training time is not too long. Eventhough, the final accuracy of the model is very good and, in my opinion, more than good enough for the task.

Appendix

A Pulsars

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, 9]. Here is when machine learning algorithms come in hand to automatically label pulsar candidates, which is the purpose of this report.