

Machine learning classification of unassociated *Fermi* LAT sources

A. Bhat ^{*1} and D. Malyshev ^{**1}

Erlangen Centre for Astroparticle Physics, Erwin-Rommel-Str. 1, Erlangen, Germany

Received September 15, 1996; accepted March 16, 1997

ABSTRACT

Context. Classification of sources is one of the most important tasks in astronomy. Sources detected in one wavelength band, e.g., in gamma rays, may have several possible associations in other wavebands or there may be no plausible association candidates.

Aims. In this work, we take unassociated sources in the third *Fermi*-LAT point source catalog (3FGL) and suggest associations to known classes of gamma-ray sources using machine learning methods trained on associated sources in the 3FGL.

Methods. We use several machine learning methods to separate *Fermi*-LAT sources into two major classes: pulsars and active galactic nuclei (AGNs). We evaluate the dependence of results on meta-parameters of the ML methods, such as the depth of the tree in tree-based classification methods and the number of layers in neural networks. We test the performance of the methods with a test sample drawn from the associated sources in 3FGL. We compare the predictions with the preliminary forth *Fermi*-LAT catalog (4FGL).

Results. Summary of results

Key words. Methods: statistical – Catalogs

1 Contents

2 1 Introduction

3 2 Methods

4	2.1	Details of the analysis	3
5	2.2	Data and Features	3
6	2.3	3
7	2.4	Comparison of the classification algorithms	3

8 3 Prediction for unassociated sources in 9 3FGL and comparison with 4FGL

10 4 Prediction for unassociated source in the 11 4FGL catalog

12 5 Conclusions

13 A Appendix

14 1. Introduction

15 Catalogs of gamma-ray sources such as 3FGL or 4FGL
16 contain many sources without associations, e.g., X for
17 3FGL and Y for 4FGL. In order for a source to be asso-
18 ciated with a known source, such as a blazar or a pulsar,
19 it has to pass relatively strict selection criteria, which
20 ensure that the rate of errors in associations is low. How-
21 ever, in some problems the desired property is complete-
22 ness rather than purity. In other words, one may need to
23 select as many members of a particular class as it is pos-
24 sible, which usually comes at the expense of higher rate
25 of false associations. For example, this situation can arise
26 if we look for something new or unusual, in this case one
27 would like to pay attention to all possible members of
28 the class, even though there may be many non-members
29 added to this list. In this case it is useful to have a softer
30 selection criterion, such as a probability to belong to a
31 certain class, rather than a discrete classification. Then
32 one has an option to select a pure subsample with very
33 high probability to belong to the class, or a larger sub-
34 sample where the probability to miss a source is low but
35 for some of the sources the probability to belong to the
36 class can be also not high, e.g., even less than 50%.

* e-mail: aakash.bhat@fau.de

** e-mail: dmitry.malyshev@fau.de

Discrete or probabilistic classification of unassociated gamma-ray sources can be achieved with machine learning (ML) algorithms. The basic idea is that the algorithms are trained to classify sources based on their characteristics, such as the spectral index and position on the sky, for a set of sources with known classifications. Then the application of the classification algorithm for a sources with unknown classification allows one to make a prediction to which class it belongs. The main goal of this paper is to apply several ML algorithms for the classification of sources in 3FGL and 4FGL catalogs. In particular we will use logistic regression, decision trees, random forest, neural networks algorithms. We will

1. Train the classifiers on associated sources in the 3FGL catalog in Section 2;
2. Make prediction for unassociated sources in the 3FGL catalog and compare with the associations in the newer 4FGL catalog in Section 3;
3. Retrain the classifiers for the 4FGL catalog and make predictions for the unassociated sources in 4FGL in Section 4.

Discussion of literature and what is new in this paper.

2. Methods

Our methodology for classification was dependent on two things: The data that we had, which needed to be cleaned and the algorithms that we needed to apply. For this we decided on using the 3rd catalog of F-LAT (3FGL from hereon) for initial training and testing, the 4th catalog (FL8Y from hereon) for further testing and predictions, and machine learning algorithms like Random Forests, Logistic Regression, Decision Trees, and Neural Networks. All of the machine learning algorithms were taken from the python module sklearn, including Neural Networks. A neural network using Keras was also attempted; however, due to the classification being on only two classes, we discarded it in favour of the sklearn algorithm which was much faster.

Our data was similar to that used by Parkinson et. al. We cleaned the 3FGL catalog to have sources which were both associated and unassociated but with no missing values. We then used the associated sources which were classified as either AGNs (with multiple labels) or Pulsars, to get a list of 1905 sources. The rest of the sources without problematic values were then used as unassociated sources, which we used later on for testing and prediction. The FL8Y presented us with another way of testing the accuracy of our methods. We predicted the classifications of unassociated sources in the 3FGL and used the FL8Y to check how

many of these unassociated sources, which now had associations in the FL8Y, actually had the right prediction.

The raw data of the catalog had a lot of different features that could be used for classification. However, going by the previous studies, we decided on using the most important features, which included Flux density and the error on it, spectral index, the curvature, hardness ratios (as defined by Parkinson et. al.), variability, and also the galactic latitude, the last of which was used even in the classification of AGN and Pulsars (as opposed to Parkinson, who used it only for the young and milli-second pulsar distinction). In features where the values were high, we used the logarithmic scale to better separate the sources. The complete list of sources, along with some statistics, is given in the table below. The influence of the features on the classification, especially the differences in the various methodologies is discussed in much more detail in the next section.

One of the main aims of our project was to understand and optimize the machine learning methods which we were using. So apart from the features which were in the data itself, we also theorized and experimented with the parameters of the algorithms themselves. We wanted to find the fastest and cost-effective way of using certain methods, without going into regimes of under and over-fitting the data. Parameters which we studied range from Depth and Number of trees in Forest based methods to the number of hidden layers and epochs in neural networks. The details are given in the next section, where we discuss our expectations and the resulting behaviour of our algorithms.

In our general the Methodology was as follows.

1. Split the PS with known classification into learning and test samples.
2. Use the learning sample for training and for selection of features. In particular, continuous parameters, such as the thresholds in the decision trees or mixing matrices in neural networks, are determined from the learning sample.
3. Meta-parameters, which encode the complexity of the methods, such as the depth of the decision trees, are determined from the best performance on the test sample.

After the above had been completed we were ready both with our final data and our optimum algorithms. We then applied and sought the results using both the catalogs in our possession. This is discussed in detail in section 4 and 5.

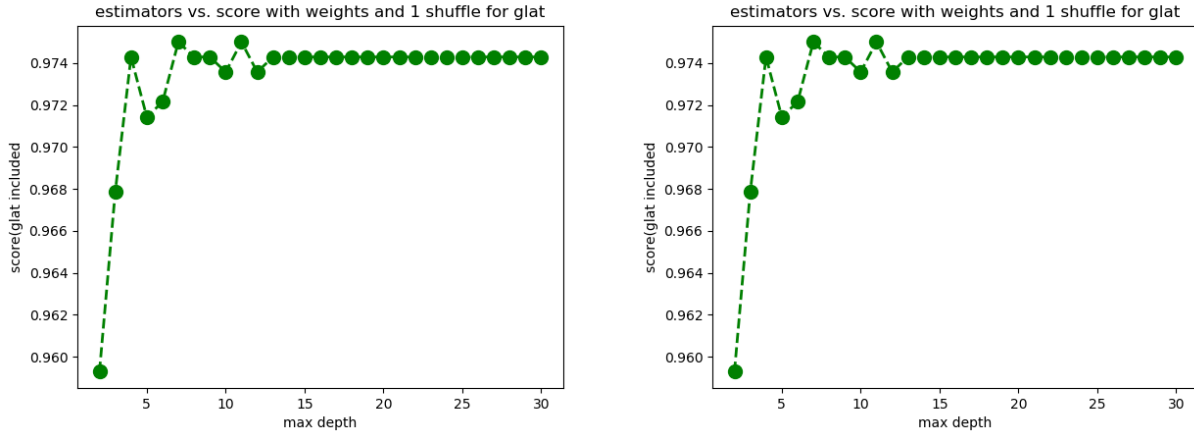


Fig. 3. Example of a figure for both columns.

Plot: add unassociated sources on the plots with domains for the best algorithm.

Create a table with sources which are more likely to be pulsars (select about 20 the most likely candidates). Compare the accuracy of the algorithm for the sources which have an associate now in the 4FGL catalog.

4. Prediction for unassociated source in the 4FGL catalog

5. Conclusions

220 **Appendix A: Appendix**

221 If we need one.