Astronomy & Astrophysics manuscript no. 0head_ML_3FGL ©ESO 2019
November 14, 2019

Machine learning methods for 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

Contents

1. Introduction

Machine learning algorithms have been around for some time. Their use in classification, etc. is well studied. However, it is only recently that machine learning has found it's way to astronomical classifications and tasks. Their use, especially with the growth of neural networks has increased exponentially in the past few years. They are now being used in classification of astrophysical sources, as well as in other areas such as reconsttruction of particle tracks (for instance in Icecube etc.).

The fermi large area telescope (LAT) was launched in 2008 for the detection of photons in the gamma-ray regime. The LAT team has since then released 4 catalogs with the 8-year list being released in 2019. These catalogs provide a list of sources, which include AGNs and Pulsars. While a lot of them are associated, many of the sources still remain unassociated.

Attempts to classify these unassociated sources have previously been performed. In 2016, Parkinson et. al. used statistical and machine learning methods like Random Forests and logistic regression to try and classify sources in the 3rd catalog released by the LAT team. They trained on 70 % of the associated sources in the catalog and then tested their results on the rest of the 30 % sources. The methods were used to classify AGNS and Pulsars (and seperately young and millisecond pulsars) and showed accuracy of up to 97%.

In our paper we present our machine learning algorithms and go deeper into their working and data analysis strategies.

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

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

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 multiplpe 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 unasssociated 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.), variablity, 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 seperate 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.

2.1. Details of the analysis

In this subsection we give more details on the analysis:

1. Describe the features that we use for the analysis.
2. Describe the objective function for minimization (accuracy of classification on learning sample). Weighted objective function: give more weight to pulsars, since there are fewer of them in the catalog.
3. Learning curve using all features? *Plot: classification accuracy using the total list of features for learning and test sample as a function of complexity parameter.*
4. Selection of the most important features. *Table: features vs algorithms. Columns: algorithms, rows: features, values: significance.*
5. Selection of meta-parameters. *Plot: classification results for the test sample using a subset of features.*
6. Train the final classifier. *Table: classification accuracy of the final classifiers for different algorithms using the test sample from 3FGL.*

Discuss the general features of the optimal algorithms: which features turn out to be important, what is the depth of the trees, the number of trees in random forests, the depth and number of internal nodes in the neural networks.

2.2. Comparison of the classification algorithms

Plot: classification domains for a pair of features (or different pairs of features, e.g., latitude vs index, index vs curvature, latitude vs variability).

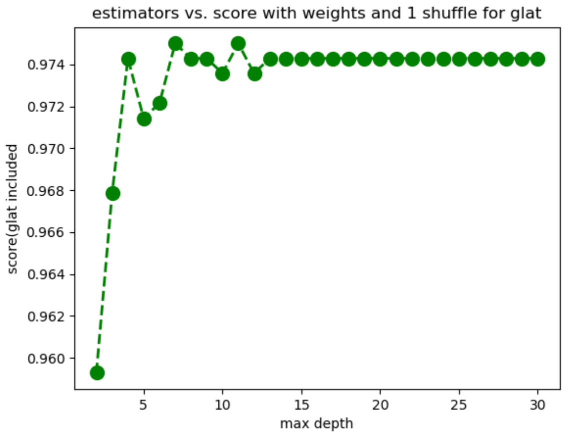

Fig. 1. Example of a figure for one column.

Probabilistic classification? Result: probability for a source to belong to a particular class. Result of classification: table of sources with probabilities for different algorithms. Final probability: the probability for one of the algorithms (for the most precise one?) and uncertainties determined from the other algorithms.

Discuss a few examples where algorithms give different predictions (are these sources at the boundaries of the domains).

Discuss examples where algorithms misclassify sources from the test sample.

3. Prediction for unassociated sources in 3FGL and comparison with 4FGL

Apply the algorithms on unassociated sources in 3FGL.

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. Comparison with 4FGL

5. Conclusions

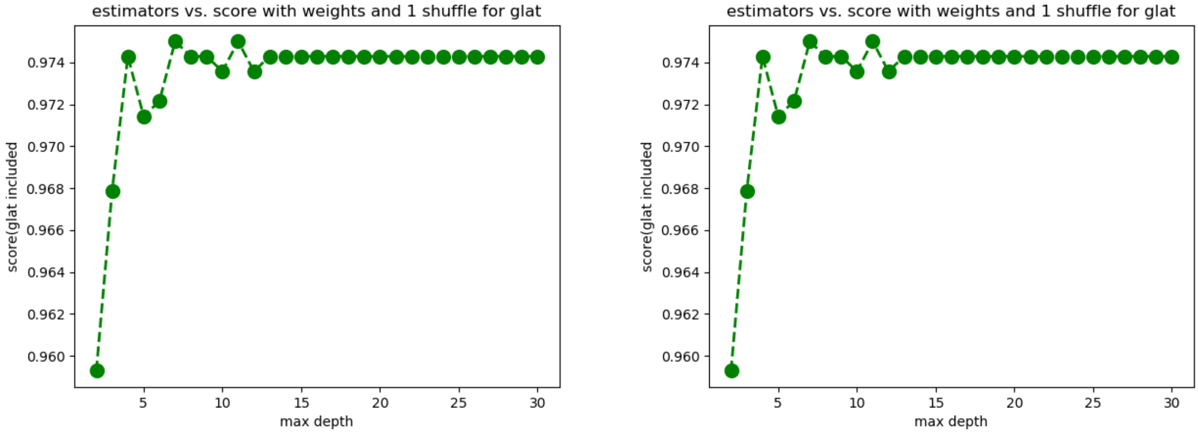

Fig. 2. Example of a figure for both columns.

Appendix A: Appendix

If we need one.