Astronomy & Astrophysics manuscript no. 0head_ML_3FGL
March 14, 2020

Machine learning methods for probabilistic catalogs

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

Catalogs of gamma-ray point sources are typically designed to have low false detection rate. False detections may arise, for example, from statistical fluctuations of the background emission, from deficiencies in the diffuse emission model, or from an overlap of faint sources, which results in a detection of a single source. The low false detection rate is achieved by setting a high statistical significance threshold, e.g., 4 or 5 sigma. Although a high detection threshold helps to eliminate most of the false detections due to statistical fluctuations, it is not very effective against deficiencies of the background model or overlapping sources. Moreover, a high threshold removes many objects, which have a high chance to be point-like sources. In other words, the catalogs are typically designed to be clean, but in some cases one may be interested to have a complete catalog. For example, one may want to have a list of all possible pulsar candidates among the unassociated sources in a catalog, which can be derived at the expense of many non-pulsars in the list.

The idea of probabilistic catalogs [Finkbeiner] is to include additional information, which describes a probability that a particular object is a point source or that a particular unassociated PS belongs to a certain class. For

[⋆] e-mail: aakash.bhat@fau.de
[⋆⋆] e-mail: dmitry.mayshev@fau.de

example, about one third of *Fermi*-LAT sources have no firm associations with known Galactic or extragalactic sources. Although the associations are unknown, these sources can still be classified with some probabilities into, e.g., extragalactic or Galactic sources based on their position on the sky, properties of the gamma-ray flux and other features. The classes can be further subdivided into various types of blazars or galaxies for extragalactic sources, or pulsars, pulsar wind nebulae, or supernova remnants for Galactic sources. The classification probability is not unique, it depends on the classification method. The range of probabilities corresponding to different methods can serve as an estimate of the modeling uncertainty of the classification. In case of PS detection, one can derives probabilities that an object is a point source, a statistical fluctuation of the background, a deficiency of the background model, or an overlap of point sources. In this case, the probability will include not only the statistical probability but also the modeling uncertainties.

In this paper we will construct a probabilistic catalog using as an example classification of unassociated sources in the *Fermi*-LAT catalogs. We will start with the Third *Fermi*-LAT catalog (3FGL) and classify the unassociated sources into pulsars and AGNs using the associated sources in 3FGL for training of the classification algorithms. We will use several machine learning algorithms for the classification, e.g., random forest, boosted decision trees, and neural nets (since the number of features and the training sample are small, the neural networks will be rather shallow). We will show applications of the probabilistic catalog for predicting the number of pulsars among the unassociated source and in construction of the source counts as a function of their flux, dN/dS. Since unassociated sources on average have smaller flux than the associated ones, the dN/dS distribution for the probabilistic catalog extends to lower fluxes relative to counting only the associated sources. We will compare the prediction for the number of pulsars and the dN/dS functions with the Forth *Fermi*-LAT catalog (4FGL).

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 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.

When applied on the 3FGL known sources, using 1500 sources to train and the rest to test on, we found (for 10 seeds) the following:

Algorithm Name	Parameters	Accuracy
Random Forest	50 trees and 12 max depth	97.91
Neural Network	200 epochs and 20 neurons in 1 layers	98.2
Gradient Boost	50,15	96.78
Logistic Regression	all solvers	<94

Table 1. Testing Accuracy of 4 algorithms on 3FGL data

2.1. Details of the analysis

2.2. Data and Features

The total number of sources, including unassociated and associated, in the two catalogs is shown below.
[Add Table]

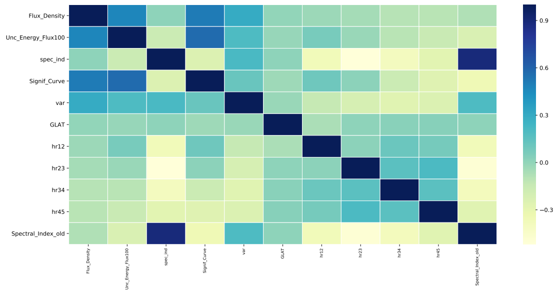

Fig. 1. Correlation matrix for the most important features

The features used for our analysis follow the same idea as the previous studies. The features, along with statistical and methodological details, are given below. A correlation matrix is presented for the most important features as well. The matrix is important for the case where there might be redundant features, in which case using only one of the two features would be a better idea.

[Add Table of features for both catalogs]

Our initial hypothesis was that certain features would be more important for classification than others. For instance, as shown below, one can see a clear distinction between the regimes of AGNs and Pulsars, based on spectral idex and significant curvature. [Add image] While not clearly obvious from the get go, we were also interested in comparing the importance of features based on the algorithms that we were using. Due to the difference in the basic method of Random Forests and Neural Networks, we expected a slight shift in their reliance on certain features. Despite that we hypothesized that features with the most contribution would be among spectral index, variability, and the curvature; as already observed by Parkinson et. al.

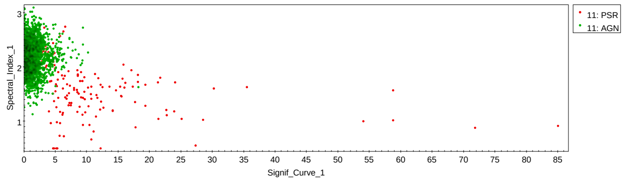

Fig. 2. Differences in AGNs and PSRs from the 3FGL catalog

2.3.

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.

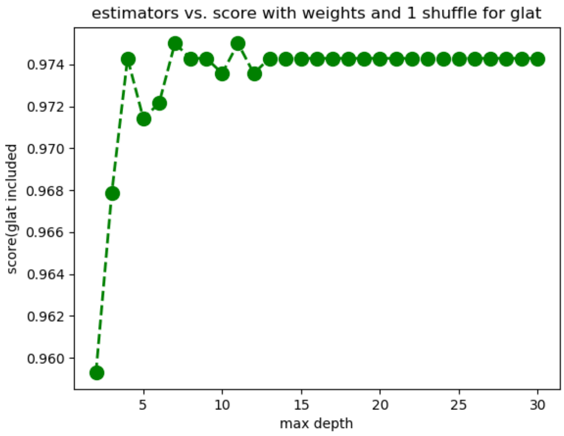

Fig. 3. Example of a figure for one column.

Our hypothesis about feature importances turned out to be correct, as curvature, variability, and spectral index were the most important features. The last hardness ratio was also seen to be quite important, most probably reflecting the end of the spectrum where the AGNs and PSRs shift from each other. These values are given in the table below.

Feature Name	RF (50,15)	GB (50,15)
Flux Density	0	0
Unc Energy Flux100	0	0
Spectral Index	0.16	0.07
Significant curvature	0.28	0.47
var	0.11	0.21
hr12	0.06	0.04
hr23	0.04	0.02
hr34	0.06	0.04
hr45	0.22	0.10
GLAT	0.04	0.01

Table 2. Feature importances for different Algorithm

These importances were found to be consistent for various different algorithm parameters. So while the value might change a bit for different tree architechtures, for instance, the importances of these features were still pronounced.

2.4. 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).

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.

In the case of test data, we worked with three different classification algortithms, namely Random Forests, Ada Boost, and Neural Networks. Here we were mostly concerned with tweaking the parameters of the classification algorithms involved, minimizing the cost of computation and aiming for the most efficient way of classification.

2.4.1. Random Forests

The two main parameters involved in Random Forests are the number of trees and the maximum depth of the trees involved. Figures below shows one instance of the accuracy as a function of maximum depth when the number of trees was kept constant, and as a function of number of trees when the maximum depth was kept constant.

2.4.2. Neural Networks

In the case of neural networks we were concerned with the number of epochs that one would need to tweak, along with a dependence on the number of neurons in the hidden layers. A final improvement involved checking whether multiple hidden layers would actually add to such a classification algorithm or not.
As can be seen in the figure, a complex network with two hidden layers (100 and 5 neurons) reaches the maximum accuracy pretty fast. The results becoming more consistent at higher epochs. A similar result was found for a network having two hidden layers but with only 20 neurons in the first layer. However, such networks could also lead to overtraining, and therefore it is important to check whether such a high accuracy could perhaps be reached by less complicated algorithms, which would drastically reduce the chances of overtraining and allow for a more flexible classificaition methodology.

A consistent and accurate result is found even for networks with only one hidden layer with 20 and 5 neurons in the hidden layer respectively. There seems to be no significant dependence for the number of epochs above 30, and even a simple network with one layer and

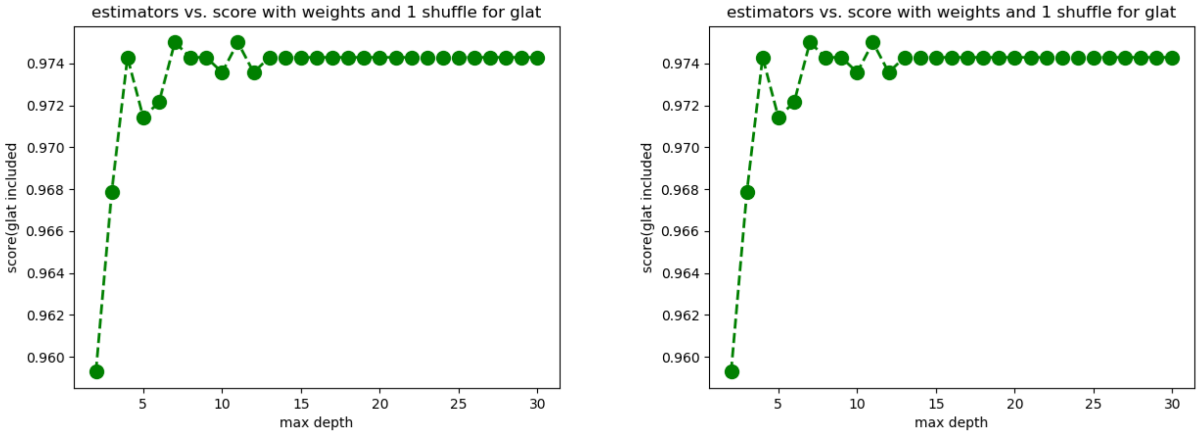

Fig. 4. Example of a figure for both columns.

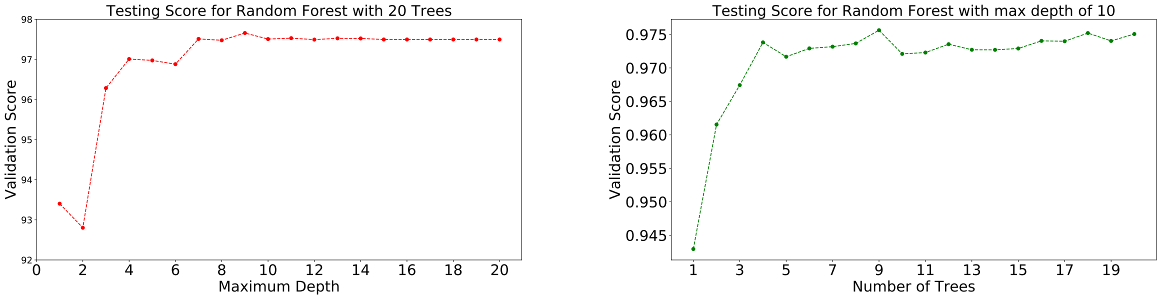

Fig. 5. Random Forests on Testing Data

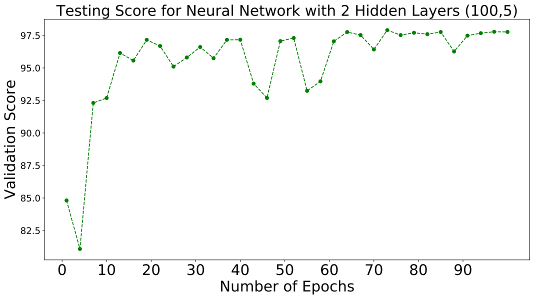

Fig. 6. Example of a figure for one column.

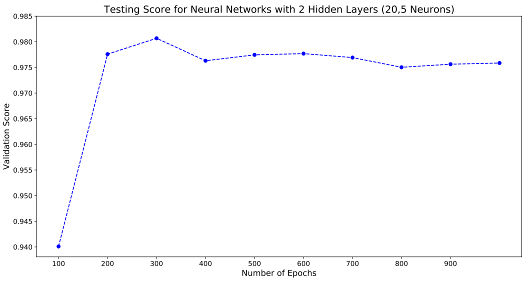

Fig. 7. Neural Network

5 neurons shows a high accuracy for only 30-40 epochs.

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

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

Comment: 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.

In this section we use the best algorithms from the previous section to predict classes for the unassociated sources in the 3FGL. We then use the associations which exist for some of these sources in the 4FGL to check the accuracy of our methods on the unassociated data. In this section we work only with Random Forests, Neural Networks, AdaBoost, and Logistic Regression.

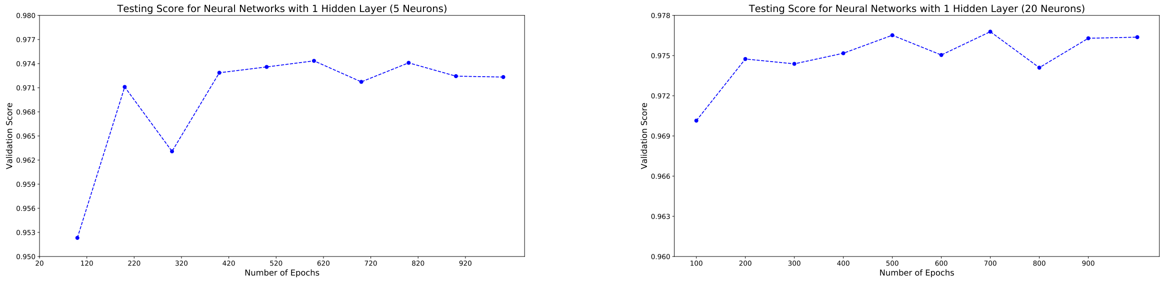

Fig. 8. Neural networks

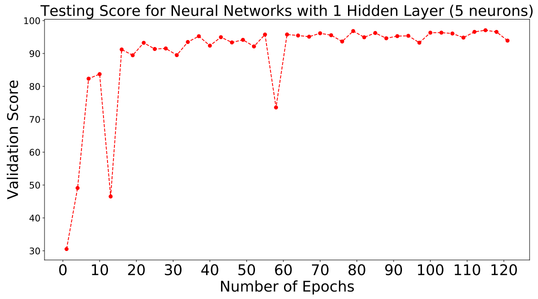

Fig. 9. Neural Network

3.1. 3FGL Unassociated sources with Association in 4FGL

3.2. 3FGL Probabilistic classification

4. Prediction for unassociated source in the 4FGL catalog

5. Conclusions

Appendix A: Appendix

If we need one.