Probabilistic classification of Fermi-LAT sources

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Dmitry Malyshev & Aakash Bhat

dmitry.malyshev@fau.de, aakashbhat7@gmail.com

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

Based on A&A 660, A87 (2022), arXiv:2102.07642

Abstract

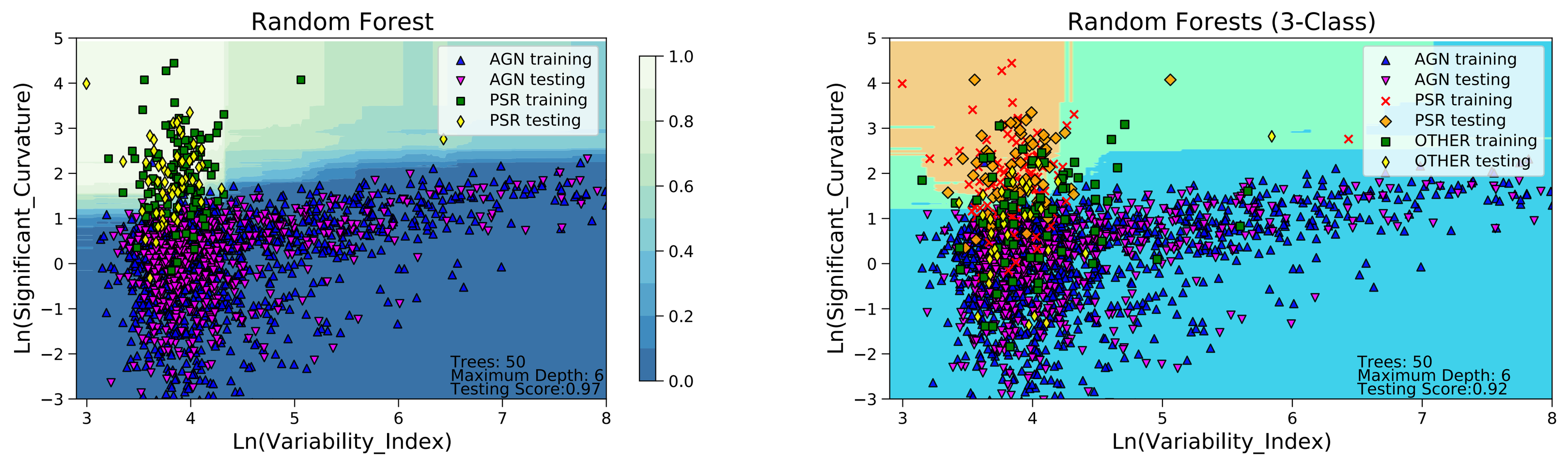
Classification of sources is one of the most important tasks in astronomy and astrophysics. About one third of sources in Fermi LAT catalogs are unclassified due to absence of plausible associations. We determine probabilistic classification of unassociated sources in the 3FGL [1] and 4FGL-DR2 [2] catalogs using machine learning (ML) methods into two and three classes. We argue that probabilistic classification can be used not only to determine the most likely classes of sources, but also to perform population studies taking into account all unassociated sources. For example, the expected density of active galactic nuclei (AGNs) including unassociated sources weighted by probabilities is approximately isotropic, while the density of associated AGNs has a dip in the Galactic plane. We perform several tests of classification, including comparison of predictions in the 3FGL catalog with associations in 4FGL-DR2.

Methods

Algorithms: boosted decision trees (BDT), random forests (RF), logistic regression (LR), and neural networks (NN).
Data: for training we use associated sources in Fermi-LAT catalogs [1, 2]. We consider 2-class classification: AGNs and pulsars, and 3-class classification: AGNs, pulsars, and other Galactic sources. We split the associated sources into 70% training and 30% testing samples.
Features: we use parameters of sources, such as spectral index, position on the sky, variability index etc. which have less then 0.75 correlation among themselves. For 3FGL catalog we select eleven features, for 4FGL-DR2 – sixteen features.
Meta-parameters: we optimize meta-parameters (e.g., depth of trees, number of neurons) in order to obtain the best accuracy of prediction without overfitting the data.

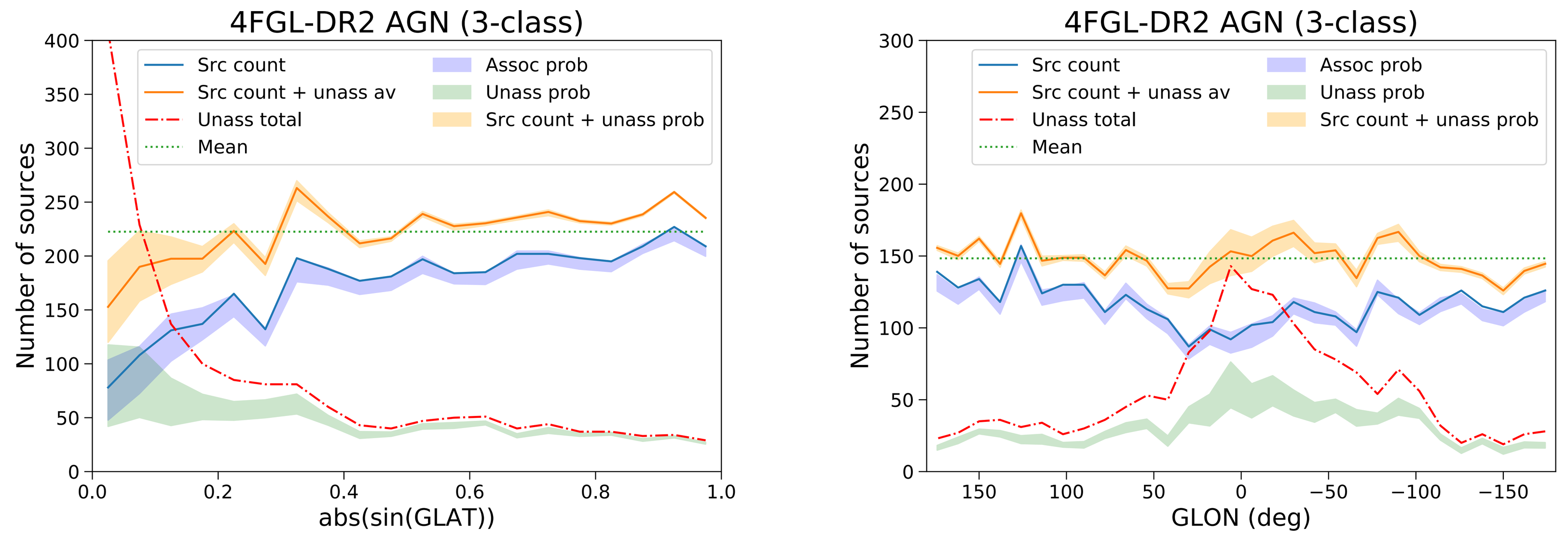
Classification

We classify Fermi-LAT sources using the four ML algorithms trained without and with oversampling. An example of probability domains for RF algorithm using two features (for visualization purposes) in the 2- and 3-class classification of 3FGL sources is presented below. The full probabilistic catalogs as well as lists of most likely pulsar and other galactic source candidates among unassociated sources are available online [3].



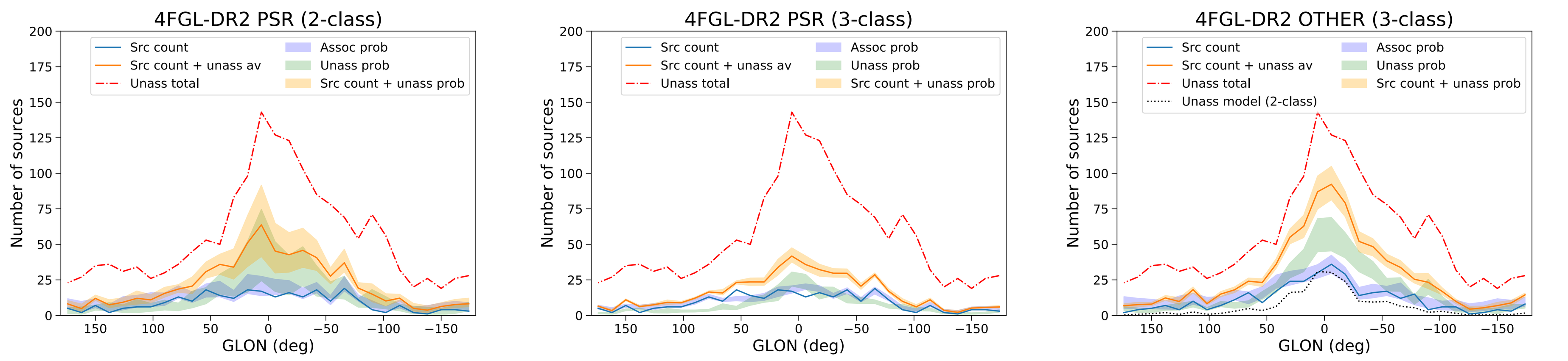
Population studies

As an example of an application for populations studies, we show the distribution of AGNs as a function of Galactic latitude and longitude (Glat and Glon) in case of 3-class classification of 4FGL-DR2 sources. The distribution of associated AGNs is not isotropic, while including the unassociated sources weighted by the AGN-like probability makes the total expected distribution consistent with isotropy, as expected for extragalactic sources.



Two- vs three-class classification

There has been a debate whether the 3-class classification of Fermi LAT sources is stable [4]. We find that 3-class classification provides stable results with accuracies comparable to the 2-class case. As an example, we show the distribution of expected numbers of pulsars and other Galactic sources as a function of Glat. In the 2-class case the distribution has large uncertainties near the Galactic center (GC) and potentially large number of pulsars among unassociated sources. In the 3-class case most of unassociated sources are attributed to other sources near the GC.



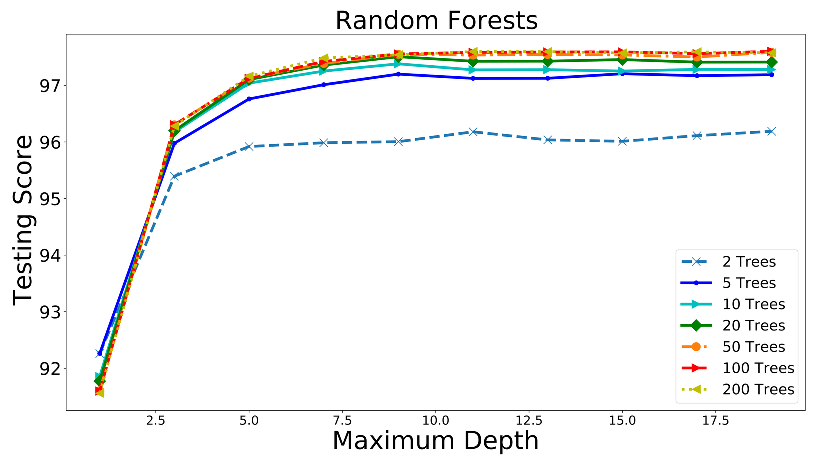
Conclusions

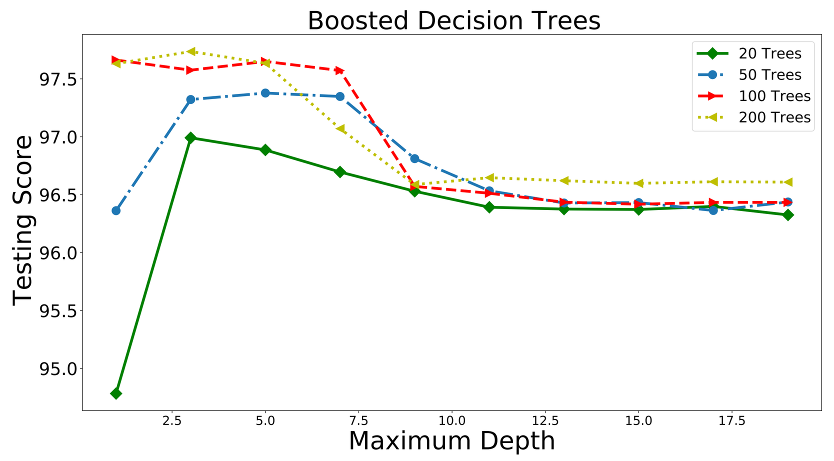
- We determine probabilistic classification of sources in the 3FGL and 4FGL-DR2 catalogs using four ML methods.
- The full probabilistic catalogs (including class probabilities for all ML methods and all sources) in the 2- and 3-class cases as well as lists of most likely pulsar and other Galactic source candidates among unassociated sources are available online [3].
- We find that the three-class classification provides stable results with accuracy of classification similar or better than the 2-class classification (if we assume that non-pulsar Galactic sources are "misclassified" in the 2-class case). At the same time, the 3-class case gives classification probabilities for non-pulsar Galactic sources, which are attributed to pulsar or AGN classes in the 2-class case.

Details of the derivation and tests


Optimisation of meta-parameters

Test of accuracy for RF and BDT algorithms as a function of tree depth and the number of trees in the estimator. It is interesting to note that RF algorithm is stable against overfitting as the depth of the trees increases (there is no decrease in the accuracy for large depths), while BDT has the best performance with depth of trees between about 3 and 8. For LR we have tested the number of iterations and different optimizers, while for NN we have tested the numbers of hidden layers, neurons, and training epochs, as well as different activation functions (Relu, tanh) and optimization algorithms (Adam, LBFGS).





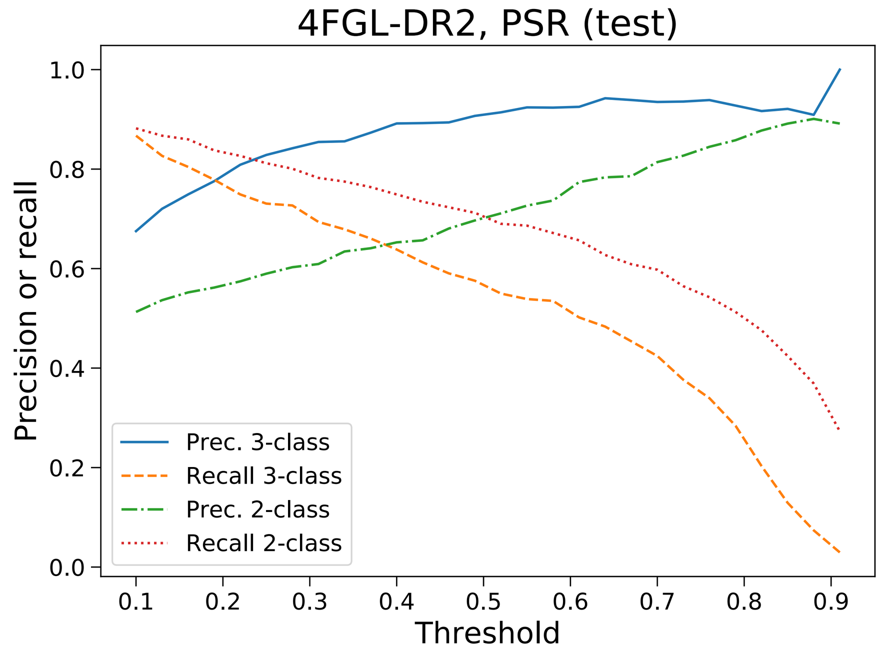
Test of 3FGL classifications with 4FGL

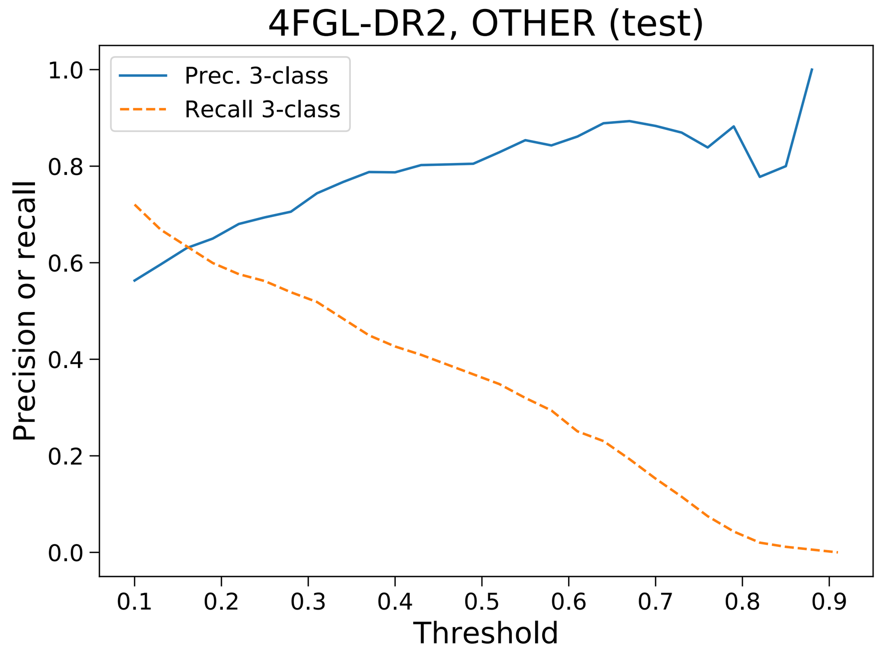
We compare probabilistic classifications of unassociated 3FGL sources with associations in the 4FGL-DR2 catalog (when available). We find that the accuracy based on this comparison is worse than the accuracy based on the testing samples, which can be due to systematic effects not accounted for in training, such as differences in distributions of associated and unassociated sources in feature space and larger uncertainties for unassociated sources, which have smaller fluxes.

Algorithm	Parameters	Testing Accuracy	Std. Dev.	Comparison with 4FGL-DR2 Accuracy
RF	50 trees, max depth 6	97.37	0.60	91.09
RF_O		97.90	0.50	89.44
BDT	100 trees, max depth 2	97.65	0.54	90.43
BDT_O		97.79	0.51	91.75
NN	300 epochs, 11 neurons, LBFGS	97.29	0.97	90.10
NN_O		94.31	5.13	87.13
LR	200 iterations, LBFGS solver	97.63	0.54	90.43
LR_O		93.68	0.99	85.15

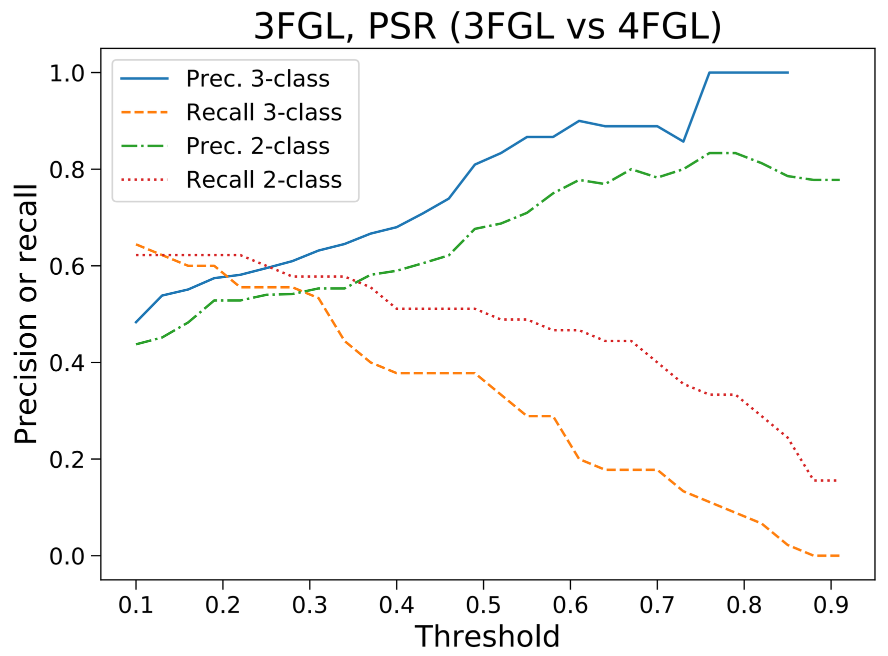
Precision and recall

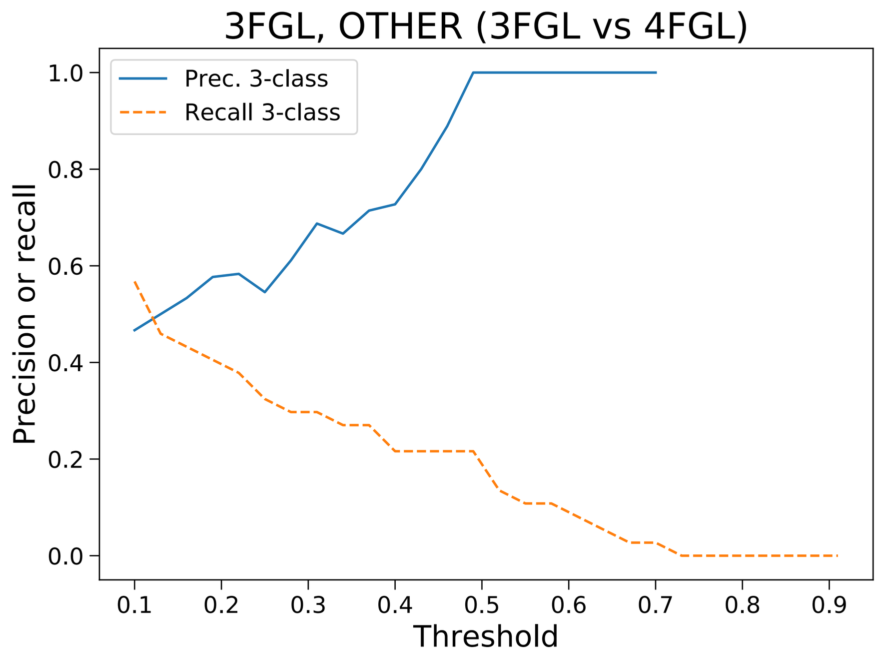
In order to determine the number of candidate sources, one usually selects a probability threshold, above which the unassociated sources are attributed to a particular class. Below we show the precision (the fraction of true sources among predictions) and recall (the fraction of true sources relative to the number of associated sources) for the test samples of sources. For the classification we use "all algorithms agree" condition, i.e., the probabilities for all algorithms are above the threshold. We note that the expected precision is larger in the 3-class case compared to the 2-class case, while the recall is smaller.





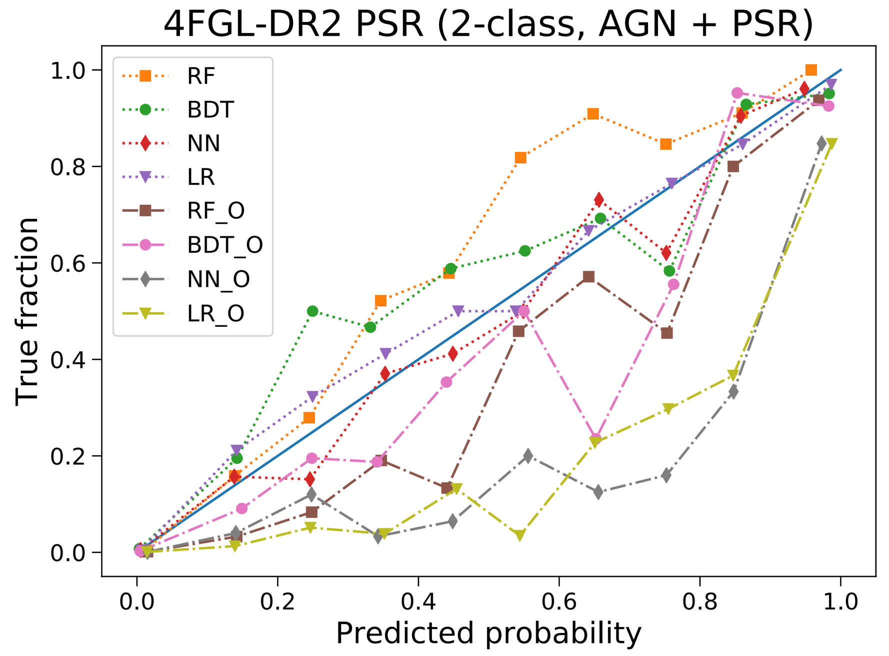
We also calculate precision and recall for unassociated 3FGL sources using the available associations in the 4FGL-DR2 catalog in order to determine the "true" classes. The performance for the unassociated sources is slightly worse than the performance for the test samples.

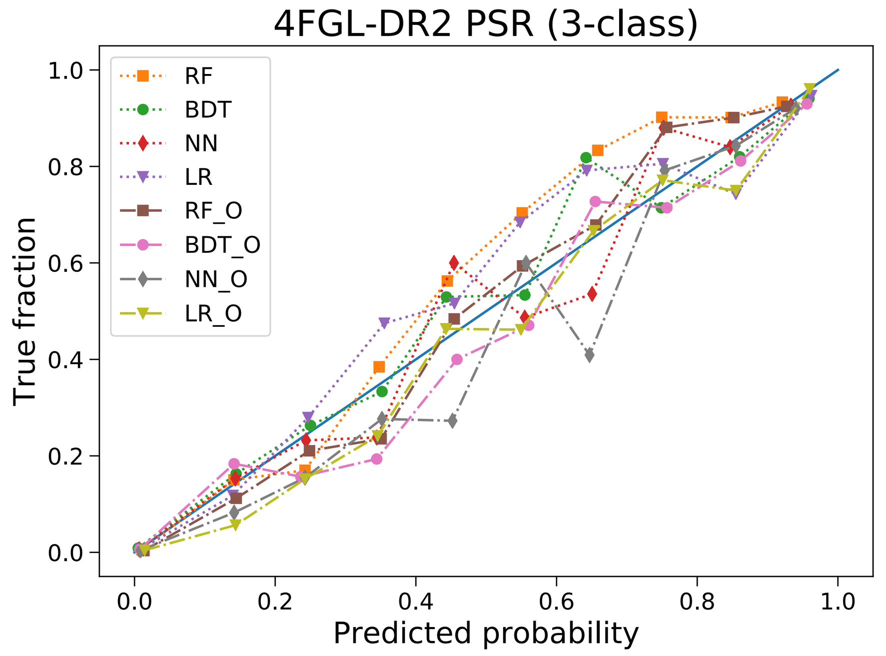




Reliability diagrams

In order to test the interpretation of classification scores provided by the algorithms as probabilities, we compute reliability diagrams. Reliability diagrams compare predicted probability (x-axis) to the fraction of true sources in bins of predicted probabilities (y-axis). The reliability diagrams for different ML methods show a wider spread in the 2-class case (left panel) compared to the 3-class case (right panel) around the optimal prediction (y = x line), even if we take only AGNs and pulsars into account in the 2-class case.





References

[1] F. Acero et al., "Fermi Large Area Telescope Third Source Catalog", ApJS, 218, 23 (2015), arXiv:1501.02003
[2] J. Ballet et al., "Fermi Large Area Telescope Fourth Source Catalog Data Release 2", arXiv:2005.11208 (2020)
[3] A. Bhat, D. Malyshev, "Machine learning methods for constructing probabilistic Fermi-LAT catalogs", A&A 660, A87 (2022), arXiv:2102.07642
[4] K. Zhu et al, "Searching for AGN and pulsar candidates in 4FGL unassociated sources using machine learning", RAA 21, 015, (2021), arXiv:2001.06010