RAA 2021 Vol. **21** No. **1**, 15(15pp) doi: 10.1088/1674-4527/21/1/15

http://www.raa-journal.org *http://iopscience.iop.org/raa*

***R**esearch in*
***A**stronomy and*
***A**strophysics*

Searching for AGN and pulsar candidates in 4FGL unassociated sources using machine learning

Ke-Rui Zhu[1], Shi-Ju Kang[2] and Yong-Gang Zheng[1]

[1] Department of Physics, Yunnan Normal University, Kunming, Yunnan, 650092, China; *ynzyg@ynu.edu.cn*
[2] School of Physics and Electrical Engineering, Liupanshui Normal University, Liupanshui, Guizhou, 553004, China

Received 2020 May 22; accepted 2020 July 6

Abstract In the fourth Fermi Large Area Telescope source catalog (4FGL), 5064 γ-ray sources are reported, including 3207 active galactic nuclei (AGNs), 239 pulsars, 1336 unassociated sources, 92 sources with weak association with blazars at low Galactic latitudes and 190 other sources. We employ two different supervised machine learning classifiers, combined with the direct observation parameters given by the 4FGL fits table, to search for sources potentially classified as AGNs and pulsars in the 1336 unassociated sources. In order to reduce the error caused by the large difference in the sizes of samples, we divide the classification process into two separate steps in order to identify the AGNs and the pulsars. First, we select the identified AGNs from all of the samples, and then select the identified pulsars from the remaining cases. Using the 4FGL sources associated or identified as AGNs, pulsars and other sources with the features selected through the K-S test and the random forest (RF) feature importance measurement, we trained, optimized and tested our classifier models. Then, the models are applied to classify the 1336 unassociated sources. According to the calculation results of the two classifiers, we report the sensitivity, specificity, accuracy in each step and the class of unassociated sources given by each classifier. The accuracy obtained in the first step is approximately 95%; in the second step, the obtained overall accuracy is approximately 80%. Combining the results of the two classifiers, we predict that there are 583 AGN-type candidates, 115 pulsar-type candidates, 154 other types of γ-ray candidates and 484 of uncertain types.

Key words: gamma rays: galaxies — galaxies: active — methods: statistical

1 INTRODUCTION

Both the Celestial Observation Satellite (COS-B) γ-ray source catalogs (e.g., Hermsen 1981; Pollock et al. 1987) and the Compton Gamma Ray Observatory (CGRO) γ-ray source catalogs (e.g., Fichtel et al. 1994; Thompson et al. 1995; Hartman et al. 1999) contain a small number of sources, most of which are unassociated sources. The identification of MeV-GeV γ-ray sources, over a long period of time, suffers from few detectors and limited angular resolution. In recent years, approximately 20 types of γ-ray sources have been identified (Abdollahi et al. 2020). Most of the identified sources belong to the active galactic nuclei (AGNs) category. It is commonly believed that there is a supermassive black hole (SMBH) in the center of an AGN. Their continuum emission is characterized by high brightness and non-stellar origin. Their broad spectral energy distribution extends from radio to high-energy γ-ray bands (Karas et al. 2019). In the widely accepted unified model paradigm (Urry & Padovani 1995; Ulrich et al. 1997), an AGN is usually associated with a jet that originates from the central SMBH and is filled with relativistic plasmas. Due to their extreme characteristics, the jet of an AGN is an ideal object for studying the acceleration mechanism of high-energy particles. In addition, pulsars are another major observed type; the pulsars' high energy emission mechanism is an open issue. Considering the different locations of the emission region (Harding & Muslimov 1998a), either the polar cap model (Rudak & Dyks 1998; Harding & Muslimov 1998b) or the outer gap model (Cheng et al. 1986; Romani 1996, 2014) is applied to interpret the high-energy emission of pulsars. The latter model is more popular (Saz Parkinson et al. 2016) since a large number of radio-quiet γ-ray pulsars have been identified by Fermi-LAT (Abdo et al. 2009a; Saz Parkinson et al. 2010). However, additional evidence is still required.

In 2008, a new era in the classification of observations began to emerge. High-energy observations have been included in the Fermi catalogs; an abundance of γ-ray sources has been discovered. Over the last decade, the Fermi-Large Area Telescope (LAT) source catalog (FGL) has evolved, including the regular releases of the 0FGL (3 months, Abdo et al. 2009b), 1FGL (11 months, Abdo et al. 2010), 2FGL (2 years, Nolan et al. 2012) and 3FGL (4 years, Acero et al. 2015). Neglecting the incomplete 0FGL, the 1FGL contains 1451 sources including 630 unassociated sources (Abdo et al. 2010). Then the 2FGL reduces the number of these unassociated sources to 576; this catalog contains a total of 1873 sources. The 3FGL contains 3033 sources of which approximately one third are unassociated (Acero et al. 2015). Recently, the Fermi-LAT collaboration has provided a release of the fourth Fermi-LAT source catalog (4FGL)[1]. This catalog features the new γ-ray observation results of an eight-year period from 2008 to 2016 in the 50 MeV to 1 TeV energy range with 4σ confidence level. The 5064 sources contained in the 4FGL are divided into 23 categories (see Abdollahi et al. 2020), in which the number of sources of eight classes of AGNs is 3207, accounting for 63.3% of the total sources. Besides, 239 sources are pulsars, 1336 sources are unassociated, 191 sources are identified in 11 other categories (i.e., pulsar wind nebula, normal galaxy, etc), and 92 sources are labeled as "UNK/unk" in the 4FGL table, which are the sources with weak association with a blazar at low Galactic latitude (marked as UNK in the work). Since the AGNs and pulsars are important for the field of high-energy astrophysics, we evaluate the potential classification of unassociated sources and confirm the AGN and pulsar candidates for the expanded samples.

Machine learning (ML) techniques have become more popular in the field of data mining and data analysis and are receiving attention in a wide variety of domains, including the analysis of astronomical databases (Ball & Brunner 2010; Mirabal et al. 2012; Pesenson et al. 2010; Doert & Errando 2014; Chiaro et al. 2016; Saz Parkinson et al. 2016; Lefaucheur et al. 2017; Salvetti et al. 2017; Baron 2019; Kang et al. 2019a,b; Liodakis & Blinov 2019; Faisst et al. 2019; Fluke & Jacobs 2020). As a cutting-edge cross-disciplinary subject, ML is divided into supervised machine learning (SML) and unsupervised machine learning (USML) algorithms. Based on the clustering algorithm, the USML is utilized to identify the potentially complex relationships among samples. Alternatively, if we focus primarily on the labels of datasets provided artificially, we can employ SML algorithms to realize classification and regression (Baron 2019). The aim of SML classifiers is to establish judgment criteria based on known samples to predict the classification of unknown samples. A wide variety of SML algorithms is available, including logistic regression, decision trees, random forest (RF), support vector machines, neural networks, Bayesian networks, Gaussian finite mixture models, artificial neural networks (ANNs) and many others (e.g., see Feigelson & Babu 2012; Kabacoff 2015).

In recent years, ML algorithms have been widely applied in Fermi data analysis. Many investigators have utilized them to explore the nature of unidentified γ-ray sources, for example, searching for AGNs (Mirabal et al. 2012; Doert & Errando 2014; Saz Parkinson et al. 2016) and pulsars (Mirabal et al. 2012; Saz Parkinson et al. 2016; Luo et al. 2020) in unassociated sources, or evaluating the optical classification of Fermi blazar candidates of uncertain type (BCUs) (Hassan et al. 2013; Chiaro et al. 2016; Lefaucheur et al. 2017; Salvetti et al. 2017; Kang et al. 2019a,b; Liodakis & Blinov 2019).

In the present context, we employ two SML classification methods of both RF and ANN to evaluate the potential classification of the 1336 unassociated sample sources in the 4FGL catalog. The aim is to obtain more potential AGN, pulsar and other γ-ray source (non-AGN and non-pulsar) candidates. The remainder of this paper is organized as follows. In Section 2, we describe the dataset from the 4FGL and select features using the Kolmogorov-Smirnov (K-S) test and RF feature importance measurement. In Section 3, we review SML classification algorithms, dataset partitioning and normalization, and the creation and validation of two individual algorithms (RF and ANN). In Section 4, we test the individual algorithms and composition algorithm, then apply the composition model to the 1336 unassociated sources. Some discussions and the conclusion are given in Section 5.

2 DATASET PREPARATION

In the new release of the 4FGL catalog fits table[2], 5064 γ-ray sources above a 4σ confidence level are reported, and these are divided into 23 categories. Nevertheless, not all samples are available. The nature of UNKs has not been defined, though there is a weak association between UNKs and blazar candidates. Moreover, the bright background at low Galactic latitudes impacts the observation of UNKs, which may lead to deviation in the classification process. So, 92 UNK sources are removed. In the classification, eight classes of AGNs, such as flat spectrum radio quasars,

[1] https://fermi.gsfc.nasa.gov/ssc/data/access/lat/8yr_catalog/

[2] https://fermi.gsfc.nasa.gov/ssc/data/access/lat/8yr_catalog/gll_psc_v21.fit

BL Lac objects and Seyfert galaxies, are labeled as agn. Similarly, we label the pulsars as psr, unassociated sources as unass and the rest of the sources that are identified as other γ-ray sources are labeled as other. The details of the 4972 sources that belong to different categories or labels are shown in Table 1.

As seen in Table 1, the sample is unbalanced. More specifically, the number of AGNs is approximately 15 times the number of pulsars or other types, which can significantly affect the classification results. In order to reduce the influence of the imbalances and improve the prediction accuracy, we divide the classification process into two steps. Firstly, we select the AGN candidates in all of the unassociated samples, and then select the pulsar candidates in the remaining non-AGN samples for the last step. In this way, we expand the non-AGN samples and reduce the error. The classification is done step by step; thus, there are distinct datasets in the two steps (see Table 2).

Each source in the 4FGL catalog contains 333 columns of observed data (Abdollahi et al. 2020). Excluding strings, missing columns, columns without physical significance, errors and historical data, there are 36 usable features: [$F_1 - F_7$]: integral photon flux in the band of 50 to 100 MeV, 100 to 300 MeV, 300 MeV to 1 GeV, 1 to 3 GeV, 3 to 10 GeV, 10 to 30 GeV and 30 to 300 GeV, respectively; [$\nu F_{\nu 1} - \nu F_{\nu 7}$]: spectral energy distribution over the seven bands; [GLON/GLAT]: Galactic longitude/latitude; [E_{100}]: energy flux from 100 MeV to 100 GeV; [F_{1000}]: integral photon flux from 1 to 100 GeV; [Signif_Avg]: source significance in σ units over the 100 MeV to 1 TeV band; [$E_{\rm Pivot}$]: the energy at which error in differential flux is minimal; [$K_{\rm PL}$,PL_Index]: differential flux at pivot energy, photon index in power-law (PL) fit; [$K_{\rm LP}$,LP_Index,LP_beta]: differential flux, photon index at pivot energy, curvature in logarithmic parabola (LP) fit; [$K_{\rm PLEC}$, PLEC_Index,PLEC_Expfactor andPLEC_Exp_Index]: differential flux at pivot energy, low-energy photon index, exponential factor and index in powerlaw with superexponential cutoff (PLEC) fit; [LP_ SigCurv/PLEC_SigCurv]: significance of the fit improvement between PL and LP/PLEC in σ units; [Npred]: predicted number of events in the model; [Variability_Index]: variability index over the full catalog interval; [Variability2_Index]: variability index over two-month intervals; [Frac_Variability/Frac2_Variability]: fractional variability computed from the fluxes in each year/two months.

In order to facilitate normalization and reduce the computational demands of subsequent steps in the process, we calculate the logarithm of the higher scale features (flux, energy, etc).

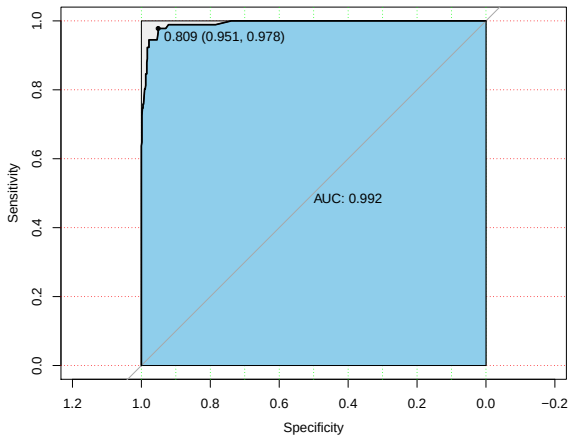

Fig. 1 The ROC curves of the RF classifier with the best hyper-parameter combination for the validation set in the first step. The text in the figure expresses the AUC value, the optimal threshold and the corresponding sensitivity and specificity.

Since different features play different roles in the classifiers, the selection of suitable input features for the SML is necessary. Noticing that, i) More input features do not always result in higher accuracy (Kang et al. 2019b); ii) More features need more computation; iii) Favorable features for the selection of the AGNs are different from those for pulsars, we further select the features for the two steps from the 36 usable features.

The K-S test is a two-sample hypothesis test method, which is often used to evaluate the significance of the distribution difference of the same measurement in two samples (e.g., Xiong & Zhang 2014; Kang et al. 2020). In particular, the K-S test can also be applied for feature selection (e.g., Kang et al. 2019a,b), based on the principle that the greater the distribution difference of the two samples over a feature, the more favorable the feature is in SML classifiers. In addition, feature importance provides a metric on the feature performance evaluation in the RF algorithm. Here, this is measured utilizing the function "*importance*" from the package "*randomForest*" (Liaw & Wiener 2002). In summary, these two test methods are employed to evaluate the 36 usable features. For the purpose of implementing the two-step classification process, we first test the features of AGNs and non-AGNs; then, the same process is applied between pulsars and other γ-ray sources. The pulsars and other γ-ray sources are labeled as non-AGN in the first step.

The test results are displayed in Table 2. In the K-S test, the statistical value D represents the distribution difference level of the feature in the two subclasses, while p signifies the probability that the feature conforms to

Table 1 The Label of 4FGL Samples

Description	Designator	Source count	Label
Non-blazar active galaxy	agn	11	agn
Blazar candidate of uncertain type	bcu	1312	agn
FSRQ type of blazar	fsrq	694	agn
Compact Steep Spectrum radio source	css	5	agn
Narrow line Seyfert 1	nlsy1	9	agn
Radio galaxy	rdg	42	agn
Seyfert galaxy	sey	1	agn
Steep spectrum radio quasar	ssrq	2	agn
BL Lac type of blazar	bll	1131	agn
Binary	bin	1	other
Normal galaxy (or part)	gal	3	other
Globular cluster	glc	30	other
High-mass binary	hmb	8	other
Low-mass binary	lmb	2	other
Nova	nov	1	other
Pulsar wind nebula	pwn	17	other
Starburst galaxy	sbg	7	other
Star-forming region	sfr	3	other
Supernova remnant	snp	40	other
Supernova remnant / Pulsar wind nebula	spp	78	other
Pulsar	psr	239	psr
Unassociated		1336	unass

Column (1): Descriptions of sources for different classes (Abdollahi et al. 2020). Col. (2): Designator of sources for different classes. Col. (3): Source count for different classes. Col. (4): The label of different sources used in this paper.

the same distribution. The RF *Gini* is the mean decrease in accuracy factors given by the measured RF feature importance; these tend to follow the same pattern as the K-S test. According to the selection criterion ($D \geq 0.35$ in the K-S test), 20 better features selected in the first step and eight better features selected in the second step are shown in Table 2. The features above the horizontal line are the features we selected.

3 ESTABLISHING CLASSIFIER MODELS

3.1 Classification Methods

In the field of SML, the dataset contains a certain number of objects. Each object has its own features and a target variable; for classifiers, the target variable is also called a label (Baron 2019). For our work, the dataset contains 5065 sources from the 4FGL catalog, the features are the observation data of these sources and the target variables are the classes of the source.

In a classifier, the model learns the corresponding relationships between input features and target variables. Then, inputting the features of the unknown samples the model outputs the probability P (usually normalized to 0–1) of each sample. Based on the classification threshold (the default value is 0.5 in two-sample classifiers), the unknown samples can be divided into two classes. Therefore, the dataset is further divided according to the role it plays in the classification process. The known samples are put into the training, validation, and test datasets in a certain proportion. The training set is applied to train the classification model. The validation set can help to find the best combination of hyper-parameters (parameters of classifiers, such as the number of trees in RF), classification threshold of different algorithms, or prevent overfitting (see Baron 2019 for more details). The test set is used to evaluate the classifiers' performance in terms of accuracy, sensitivity, etc.

This work employs both RF and ANN algorithms, which contain different origins and characteristics. The RF algorithms are derived from a "decision tree" algorithm, which is a simple classifier algorithm (see e.g., Paul & Utgoff 1989; Duda et al. 2001 for more details). The principle of a decision tree algorithm is to build nodes to make one-to-one judgments, and a large number of nodes constitutes a "tree". However, a limitation of the "decision tree" is an overfitting situation, which leads to a decrease in the accuracy of judgment (Duda et al. 2001). An RF algorithm addresses the overfitting problem by utilizing a combination of a large number of decision trees with weight consideration for each tree (Breiman 2001). Compared with the "decision tree" (Fernández-Delgado et al. 2014), it is a more efficient and accurate classifier. Yet, a traditional RF (Breiman 2001) requires a "clean" dataset, which means that the input of uncertain features or missing values is unfavorable. Recently, the probabilistic random forest (PRF) algorithm has been proposed to deal with uncertain datasets (Reis & Baron 2019; Reis et al. 2019), which also makes the RF algorithms more suitable for astronomy data. As a mature ML classification algorithm, RF is very popular in the field of astronomical data analysis

Table 2 Results of Test

First step				Second step			
Feature	D	p	$RF\ Gini$	Feature	D	p	$RF\ Gini$
logF_4	0.605	0	19.13	PLEC_SigCurv	0.547	0	17.62
log$\nu F_{\nu 4}$	0.603	0	19.86	LP_SigCurv	0.518	0	16.28
LP_SigCurv	0.598	0	17.40	LP_beta	0.434	0	13.17
PLEC_SigCurv	0.591	0	17.69	PLEC_Expfactor	0.399	$< 1 \times 10^{-6}$	12.78
PLEC_Expfactor	0.589	0	18.93	Signif_Avg	0.394	$< 1 \times 10^{-6}$	16.47
log$F_1$000	0.588	0	19.42	log$\nu F_{\nu 7}$	0.379	$< 1 \times 10^{-6}$	14.98
Frac_Variability	0.560	0	17.21	PLEC_Index	0.375	$< 1 \times 10^{-6}$	10.63
LP_beta	0.555	0	19.29	logF_7	0.350	$< 1 \times 10^{-6}$	12.18
log$\nu F_{\nu 3}$	0.545	0	18.23	log$K_{\rm LP}$	0.281	$< 1 \times 10^{-6}$	8.19
logF_5	0.530	0	16.72	log$K_{\rm PLEC}$	0.281	$< 1 \times 10^{-6}$	6.66
logF_3	0.525	0	18.57	log$K_{\rm PL}$	0.267	$< 1 \times 10^{-6}$	7.96
log$\nu F_{\nu 5}$	0.508	0	15.93	$E_{\rm Pivot}$	0.262	$< 1 \times 10^{-6}$	5.93
log$E_1$00	0.503	0	18.02	Npred	0.195	6.48×10^{-4}	10.52
Variability_Index	0.457	0	18.15	logF_5	0.186	1.33×10^{-3}	8.33
Npred	0.446	0	13.72	log$\nu F_{\nu 6}$	0.181	1.97×10^{-3}	8.70
PLEC_Index	0.445	0	18.48	log$F_1$000	0.176	2.88×10^{-3}	9.34
Variability2_Index	0.382	0	19.64	log$\nu F_{\nu 4}$	0.172	3.84×10^{-3}	8.46
Frac2_Variability	0.371	0	19.01	log$\nu F_{\nu 5}$	0.170	4.32×10^{-3}	7.60
log$K_{\rm LP}$	0.360	0	12.75	logF_4	0.167	5.60×10^{-3}	9.16
log$K_{\rm PLEC}$	0.360	0	12.75	logF_6	0.159	9.59×10^{-3}	8.78
log$K_{\rm PL}$	0.335	0	12.67	GLAT	0.153	1.44×10^{-2}	5.04
GLAT	0.329	0	9.76	PL_Index	0.137	3.83×10^{-2}	6.44
Signif_Avg	0.289	0	14.89	Frac_Variability	0.136	4.08×10^{-2}	7.44
log$\nu F_{\nu 2}$	0.288	0	14.15	GLON	0.133	4.68×10^{-2}	4.51
logF_2	0.272	0	12.65	log$\nu F_{\nu 3}$	0.116	1.18×10^{-1}	6.97
PL_Index	0.261	0	12.10	logF_3	0.110	1.57×10^{-1}	7.58
log$\nu F_{\nu 7}$	0.195	$< 1 \times 10^{-6}$	10.51	Frac2_Variability	0.104	2.02×10^{-1}	1.89
LP_Index	0.188	$< 1 \times 10^{-6}$	13.80	LP_Index	0.100	2.37×10^{-1}	5.09
GLON	0.187	$< 1 \times 10^{-6}$	3.72	log$E_1$00	0.095	2.97×10^{-1}	9.07
logF_6	0.169	$< 1 \times 10^{-6}$	14.92	logF_2	0.089	3.75×10^{-1}	3.31
$E_{\rm Pivot}$	0.164	$< 1 \times 10^{-6}$	14.79	log$\nu F_{\nu 2}$	0.089	3.75×10^{-1}	5.54
log$\nu F_{\nu 6}$	0.155	$< 1 \times 10^{-6}$	15.35	Variability2_Index	0.086	4.17×10^{-1}	–2.30
logF_7	0.152	$< 1 \times 10^{-6}$	9.62	Variability_Index	0.085	4.33×10^{-1}	0.46
log$\nu F_{\nu 1}$	0.132	3.46×10^{-6}	10.85	logF_1	0.072	6.35×10^{-1}	4.65
logF_1	0.128	8.41×10^{-6}	10.85	log$\nu F_{\nu 1}$	0.072	6.35×10^{-1}	5.79
PLEC_Exp_Index	0.009	1	0.37	PLEC_Exp_Index	0.013	1	1.00

Columns (1)–(4) display the test results of 36 "usable" features for the first step; Cols. (5)–(8) list the test results of the 36 "usable" features for the second step. Above the horizontal line are the features we selected, 20 for the first step and eight for the second step. Cols. (1) and (5): Tested feature name; Cols. (2)–(3) and (6)–(7): Value of test statistic (D) and p-value (p) for the two-sample K-S test respectively. Cols. (4) and (8): RF mean decrease in accuracy factors given by the function "*importance*" from package "*randomForest*" (Liaw & Wiener 2002).

(e.g., Feigelson & Babu 2003; Calderon & Berlind 2019; Hosenie et al. 2019; Kang et al. 2019a,b).

The ANN algorithms are based on the structure of human brain neurons, and they are implemented in both SML and USML. Owing to their nonlinearity, diversity characteristics and wide applicability in the areas of regression, classification and model prediction, the ANN algorithms are widely utilized in astronomy (e.g., Vanzella et al. 2004; Banerji et al. 2010; Eatough et al. 2010; Brescia et al. 2013, 2014; Ellison et al. 2016; Teimoorinia et al. 2016; Bilicki et al. 2018; Huertas-Company et al. 2018; Naul et al. 2018; Parks et al. 2018; Das & Sanders 2019). The network structure is generally divided into an input layer, one or more hidden neuron layers composed of a large number of nodes, and an output layer. However, the input and output data are generally normalized, which means that normalization and de-normalization conversions are necessary. In addition, there may be extensive computational demands resulting from a large number of neurons (Hussain et al. 2019).

Currently, the R language (R Core Team 2018, version for 3.5.1) is a convenient platform to implement various classifier algorithms. The RF and ANN algorithms are realized using the packages "*randomforest*" (Liaw & Wiener 2002) and "*RSNNS*" (the Stuttgart Neural Network Simulator for *R* language; see, Bergmeir & Benítez 2012), respectively.

For the purpose of evaluating the performance of classifiers, we also employed some other methods. The confusion matrix is a common metric in classifier tests (Baron 2019). The "*class_eval*" (Feigelson & Babu 2003) is a graph function that realizes the visualization of the confusion matrix. More specifically, the horizontal axis indicates the predicted label, the vertical axis represents the true label and the accuracies appear on

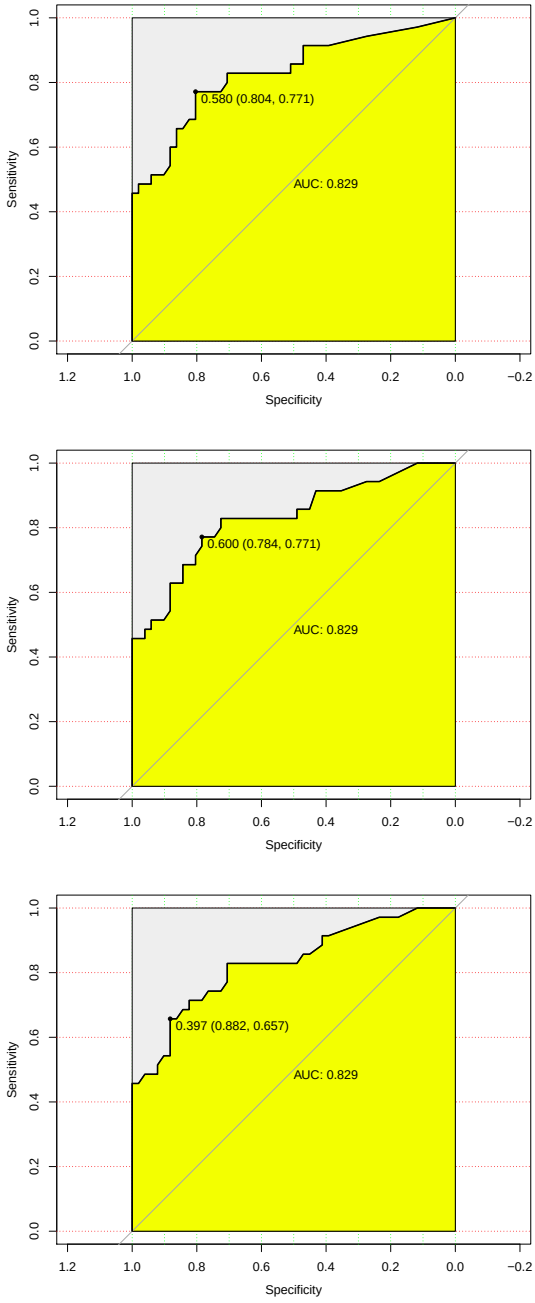

Fig. 2 The ROC curves of the RF classifier with the three best hyper-parameter combinations for the validation set in the second step. The text in the figure expresses the AUC value, the optimal threshold and the corresponding sensitivity and specificity. The different panels correspond to different hyper-parameter combinations

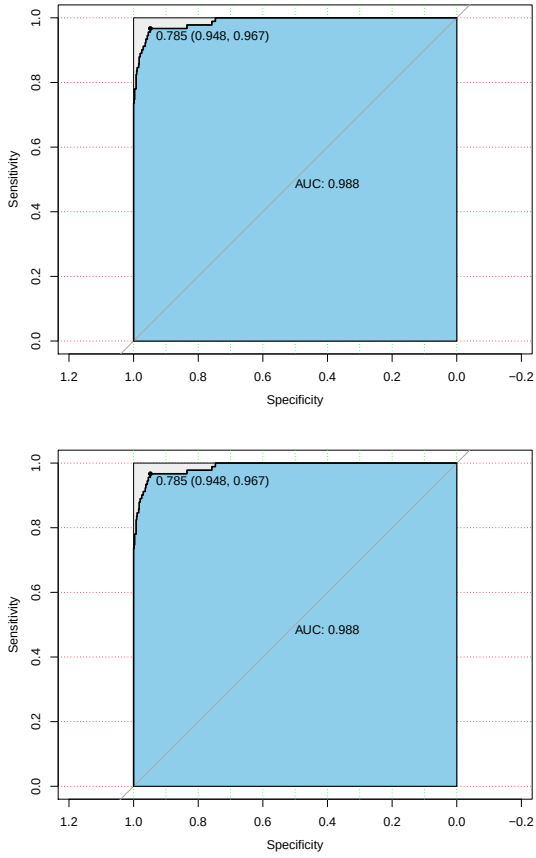

Fig. 3 The ROC curves of the ANN classifier with the two best hyper-parameter combinations for the validation set in the first step. The text in the figure expresses the AUC value, the optimal threshold and the corresponding sensitivity and specificity. The different panels correspond to different hyper-parameter combinations

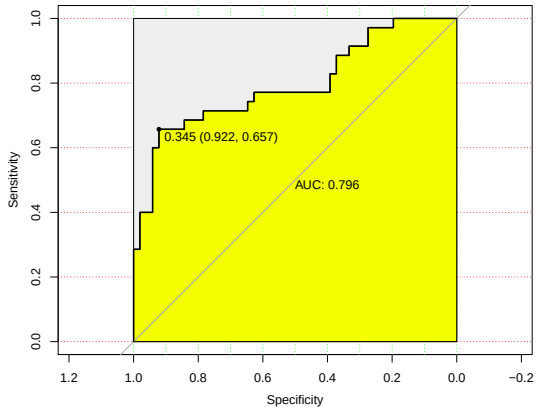

Fig. 4 The ROC curves of the ANN classifier with the best hyper-parameter combination for the validation set in the second step. The text in the figure expresses the AUC value, the optimal threshold and the corresponding sensitivity and specificity.

top of them. In addition, the function "*performance*" (Kabacoff 2015) provides a way to calculate several model performance parameters, including sensitivity (true positive rate), specificity (true negative rate) and overall accuracy based on the confusion matrix. The curves of the receiver operating characteristic (ROC) are another

important classifier performance evaluation metric, which consists of points at which the sensitivity is plotted against the specificity at different classification thresholds (Saz Parkinson et al. 2016) or the true positive rate is plotted against the false positive rate (Baron 2019). The $pROC$ package (Robin et al. 2011) is employed to obtain the ROC curves for sensitivity against specificity of different classifiers and the values of the areas under the ROC curves (AUC) in this work.

In the first step of agn selection, all of the sources in the sample with 20 selected features (see Table 2) are taken into account. A total of 3207 AGNs and 429 non-AGNs, containing 239 pulsars and 190 other sources, are randomly divided into training set, validation set and test set. Considering the impact of randomness on data set partitioning on a single result, we adopt a fixed randomness (random seed of "*12345*") and uniqueness ratio (6:2:2) between training, validation and test set. Again, the 239 pulsars and 190 other gamma-ray sources would be randomly (random seed of "*12345*") divided into training sets, validation sets and test sets in the same ratio (6:2:2) in the second step. In order to obtain uniform results, we also set the random seed as "*12345*" during RF and ANN training.

In addition, the normalization of input features is necessary in the ANN, but not in the RF, and the input target variables (class labels) of the training, validation and test set need to be decoded into a binary matrix. For the purpose of feature normalization, we call the "*normalizeData*" function in the RSNNS package, where there are three modes to choose from, i.e., type"*0_1*" (normalized to the interval from 0 to 1), type "*center*" (the data are centered, i.e., the mean is subtracted) and type "*norm*" (mean zero, variance one) (Bergmeir & Benítez 2012). In the work, our feature normalization type is "*norm*". In addition, the function "*decodeClassLabels*" is adopted for decoding class labels to a binary matrix, while the function "*encodeClassLabels*" is the approach utilized for encoding the binary matrix.

3.2 Model Creation and Validation: RF

In the package "*randomforest*" (Liaw & Wiener 2002), we build the classifier from function "*randomforest*", which contains two hyper-parameters, "*ntree*" and "*mtry*". The "*ntree*" represents the number of trees to grow, and the default value is 500. The "*mtry*" signifies the number of features randomly sampled as candidates at each split ranging from 1 to 8. For classifier, the default value is $\sqrt{n}$, where n is number of features. With all the combinations of the "*ntree*" in the range of 50 to 750 and "*mtry*" in the range of 1 to 8, we train the classifiers using the training set, and calculate the AUCs of the validation set of different hyper-parameter combinations. The hyper-parameter combinations corresponding to the maximum AUC value for the first step are displayed in Table 3 and the corresponding ROC curves are plotted in Figure 1, while those for the second step are shown in Table 3 and Figure 2. In the first step, there is a best hyper-parameter combination, "*ntree=102*" and "*mtry=3*". The best AUC is 0.992, and the thresholds are 0.809 (see Figure 1). In the second step, there are three best hyper-parameter combinations, "*ntree=56*" and "*mtry=4*", "*ntree=65*" and "*mtry=4*", "*ntree=78*" and "*mtry=4*", respectively. The best AUC is 0.829, and the thresholds are 0.580, 0.600 and 0.390, respectively (see Fig. 2).

Accordingly, in several hyper-parameter combinations with the highest AUC values, we adopt the prior one, i.e., we set "*ntree=102*" and "*mtry=3*" for RF in the first step, while the threshold is set to 0.809. "*ntree=56*", "*mtry=4*" and the threshold is set to 0.580 in the second step.

3.3 Model Creation and Validation: ANN

Compared with RF, ANN is more complicated. The package "*RSNNS*" (Bergmeir & Benítez 2012) provides many different types of network structures, including adaptive resonance theory (ART) networks, dynamic learning vector quantization (DLVQ) networks, etc. The most common way to implement an ANN classifier is to construct a multilayer (MLP) network by calling the function "*mlp*". Variable parameters include "*learnFunc*", "*maxit*" and "*size*". The "*learnFunc*" represents the used learning function, which contains different network structures, nonlinear activation functions, whether the back propagation is employed and so on. Since the learning function without back propagation is difficult to be stable in a small number of iterations, which may lead to overfitting, we choose the more common one, "*BackpropBatch*", and the parameters of learning function are set to the default value. The "*maxit*" represents the maximum of iterations to learn, and the default value is 100. The "*size*" is an array that represents the number of hidden layers and the number of neurons per layer. For example, "*c* (2,3,4)" represents three hidden layers, with the number of neurons in each layer being 2, 3 and 4, respectively. Considering the limitation of computation, similar to RF, we evaluated the performance of single and double hidden layer classifiers for all the combinations of neuron number per layer in the range of 1 to 15 and "*maxit*" in the range of 50 to 150. The hyper-parameter combinations corresponding to the maximum AUC value for the first step are shown in Table 3 and the corresponding ROC curves are depicted in

Table 3 The Best Hyper-parameter Combination of Classifiers

	RF				ANN			
	mtry	ntree	Auc	threshold	size	maxit	Auc	threshold
Step 1	3	102	0.992	0.809⋆	9	149	0.988	0.785⋆
					9	150	0.988	0.785
Step 2	4	56	0.829	0.580⋆	c(4,12)	142	0.796	0.345⋆
	4	65	0.829	0.600				
	4	78	0.829	0.390				

Column (1): Step 1 for selection of AGN and step 2 for selection of pulsar; Cols. (2)–(5): The best hyper-parameter combination, the corresponding Auc value and threshold of RF classifier respectively; Cols. (6)–(9): The best hyper-parameter combination, the corresponding Auc value and threshold of ANN classifier respectively. The hyper-parameter combination marked with a symbol ⋆ is the combination obtained in the present context.

Figure 3, while those for the second step are visualized in Table 3 and Figure 4. In the first step, the single hidden layer classifier is better, and there are two best hyper-parameter combinations, "*maxit=149*" and "*size=9*", and "*maxit=150*" and "*size=9*". The best AUC is 0.988, and both the thresholds are 0.785 (see Fig. 3). In the second step, the double hidden layer classifier is better, and there is a best hyper-parameter combination, "*maxit=142*" and "*size=c(4,12)*". The best AUC is 0.796, and the thresholds are 0.345 (see Fig. 4).

In ANN, we employ a single hidden layer classifier with "*maxit=149*", "*size=9*" and threshold of 0.785 in the first step, while a double hidden layer classifier with "*maxit=142*", "*size=c(4,12)*" and threshold 0.345 in the second step.

4 MODEL TESTING

4.1 Individual Algorithm Results

Based on the classifier models created (refer to Section 4), we tested their performance with the test set. In the first step, the test set contains 635 AGNs and 92 non-AGNs, and in the second step it includes 41 pulsars and 45 other γ-ray sources. The test confusion matrixes for the first step are shown in Figure 5, while those for the second step are depicted in Figure 6. The performance of the classifiers are displayed in Table 4.

In the first step, the ANN's accuracy was 0.944, slightly higher than the RF's accuracy of 0.939. For the RF algorithm, 34 out of the total of 635 AGNs were misclassified as non-AGNs, while 10 of a total of 92 non-AGNs were misclassified. The sensitivity for non-AGNs was 0.859, and the specificity for the AGNs was 0.956. For the ANN, 28 out of a total of 635 AGNs were misclassified as non-AGNs, while 13 of the total of 92 non-AGNs were misclassified. The sensitivity for the non-AGNs was 0.891, and the specificity for the AGNs was 0.946.

In the second step, the accuracy was not as good. The overall accuracies of two algorithms were 0.791 and 0.860, respectively. The RF algorithm has a high sensitivity of 0.822 for the other gamma-ray sources and less misclassification (8 out of 45). The specificity for pulsars was 0.791 and 10 of a total of 41 pulsars were misclassified as other sources. In contrast, the ANN has high specificity up to 0.927 for the pulsars and less misclassification (3 out of 41). The sensitivity for the other category was 0.800, and 9 of a total of 45 other sources were misclassified as pulsars.

4.2 Composite Algorithm Results

When combining the two algorithms, we are guided by the principle of common identification, that is, we obtained the classification results only when unassociated sources are classified as the same by both classifiers. When the source classification results of the two classifiers are inconsistent, we consider the sources to be the uncertain type (label as "unc"). For example, the source numbered as 4FGL J0531.7+1241c is obtained as uncertain type, while it is evaluated as an AGN in ANN classifier and evaluated as another γ-ray source in RF classifier. Hence, the accuracy is improved, although the number of candidates is reduced (e.g., Kang et al. 2019b). The combined test results of the two algorithms are provided in Table 5. For the AGNs, there are only nine misclassifications of the 614 candidates obtained, and the overall accuracy is over 98%. In the case of pulsars and other sources, the overall accuracies are 0.886 and 0.914, respectively. There are also some sources of indeterminate type.

Then, we employ the classification model to the 4FGL catalog's dataset of 1336 unassociated sources. The ANN classifier gives 911 AGNs, 166 pulsars and 259 other gamma-ray candidates. The RF classifier gives 585 AGNs, 175 pulsars and 576 other gamma-ray candidates. Combining the results of the two classification algorithms, we obtain 583 AGN candidates, 115 pulsar candidates and 154 other gamma-ray candidates (see Table 6). Figure 7 displays the distribution of the AGN and pulsar candidates over the sky. We find that most pulsar candidates are located near the Galactic plane; 74 pulsar candidates are located at GLAT $|b| \leq 10^\circ$ and 11 candidates are located at GLAT $10^\circ \leq |b| \leq 15^\circ$. The distribution is consistent with the identified pulsars. However, the AGN candidates are

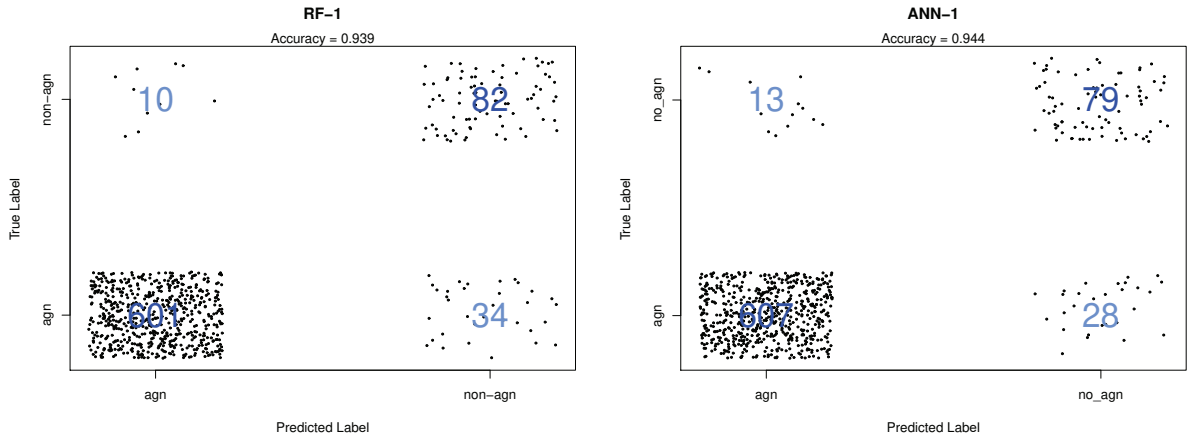

Fig. 5 The test confusion matrix of the two classifiers for the first step. The label agn is the positive sample, while the non-agns are the negative samples.

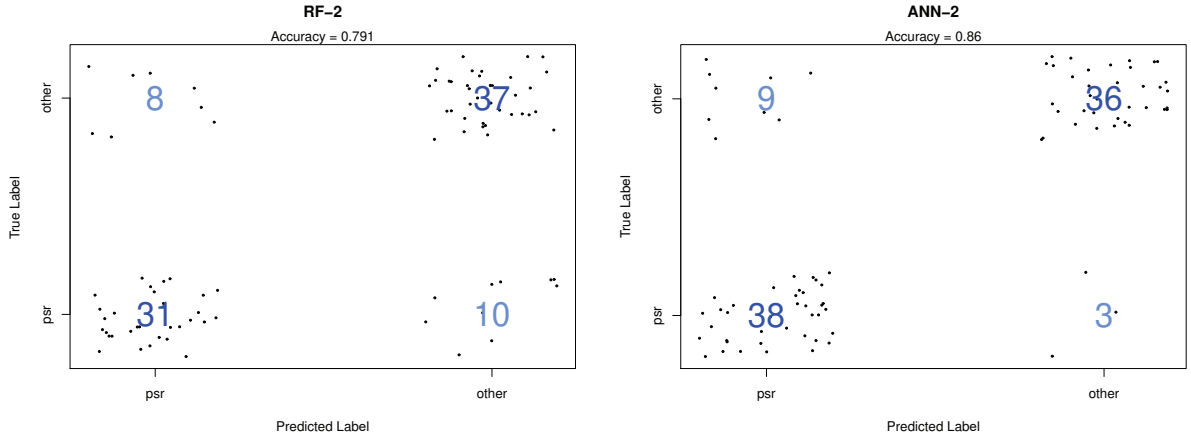

Fig. 6 The test confusion matrix of the two classifiers for the second step. The label psr is the positive sample, while the other is the negative sample.

dispersed all over the sky. Just only 108 AGN candidates are located at GLAT $|b| \leq 10°$. Since the high density distribution of the sources and the bright background are near the Galactic plane, we regard the AGN candidates of low GLAT as difficult to identify.

5 CONCLUSIONS AND DISCUSSION

In this work, we attempt to search for AGN and pulsar candidates in the 4FGL catalog's unassociated samples based on two supervised learning methods. We do not focus on the specific physical mechanism. To accommodate the unbalanced sample, we divide the classification process into two steps. Firstly, we apply the RF and ANN methods with 20 features identified by the K-S test to select AGN candidates in all of the unassociated samples. Then, we utilize the same methods with eight different features to choose pulsar candidates in the remaining non-AGN samples for the second step. By finding the optimal combination of hyper-parameters to optimize the algorithm, we test the performance of our model (accuracy, sensitivity, etc.), and evaluate the labels of the 1336 unassociated samples. The accuracy obtained in the first step is about 95%, and in the second step, the obtained overall accuracy is approximately 80%. Finally, we obtain 583 AGN candidates, 115 pulsar candidates, 154 other type of candidates and 484 uncertain type by combining the results of the two classifiers.

The current context provides the coordinates and the all-sky map of the obtained AGN and pulsar candidates. Meanwhile, the probabilities given by different classifiers for each source are also shown (see Table 5). These could help us to interpret the sky survey, as well as to further examine Fermi unassociated sources by the investigators. We note that AGNs and non-AGNs differ in spectral shape, variability, overall integral flux and flux of partial band (such as from 300 MeV to 10 GeV), which is

Table 4 Test Results for Two Classifiers

Classifier	First step			Second step		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
RF	0.891	0.946	0.939	0.822	0.756	0.791
ANN	0.859	0.956	0.944	0.800	0.927	0.860

Column (1): Classification methods applied in this paper; Cols. (2)–(4) list the test results for the first step: the sensitivity for the non-AGNs and the specificity for AGNs, and overall accuracy, respectively; Cols. (5)–(7) report the test results for the second step: sensitivity for other γ-ray sources, specificity for pulsars and overall accuracy, respectively.

Table 5 Test Results for Classifiers Combined

Class	Label	Count	Errors	Overall accuracy
AGN	agn	605	9	0.985
Pulsar	psr	35	4	0.886
Other γ-ray source	other	35	3	0.914

Columns (1) and (2): Source classes and labels respectively; Cols. (3) and (4): Source count and the number of misclassifications for each label when two classifiers are combined respectively; Col. (5): Overall accuracy for each label when combining the two classifiers.

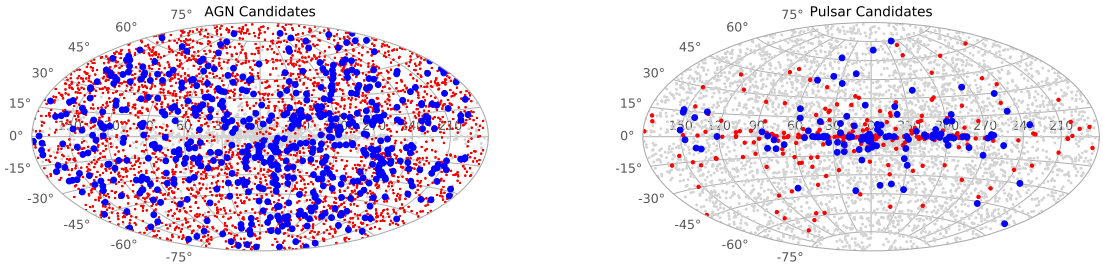

Fig. 7 All sky distribution of AGN (*left*) and pulsar (*right*) candidates in Galactic coordinates. The blue symbols represent candidates we obtained, the red symbols signify the sources of AGN or pulsar identified or associated in 4FGL and gray symbols correspond to all γ-ray sources in 4FGL.

related to the high-energy emission mechanism of AGNs (e.g., Zheng et al. 2016; Zheng & Yang 2016; Zheng et al. 2017). On the other hand, the pulsars and non-pulsars are quite different in terms of spectral shape and higher energy band flux (such as from 30 to 300 GeV), that result from the unique high-energy emission mechanism of γ-ray pulsars (e.g., Cheng et al. 1986; Romani 1996, 2014).

The basis for relying on SML for classification is the training samples and the predicted sample to be classified according to the same distribution in multi-dimensional feature space. When the distributions are different, we encounter the potential problem that the classifier does not perform as well on the unassociated samples as it does on the test samples, which is also known as covariate shift or sample selection bias in astronomy (see Richards et al. 2012; Richards 2012; Luo et al. 2020 for more detail discussions). In the classification of Fermi unassociated sources, the covariate shift exists when comparing the distribution differences of some features between 3FGL and 4FGL (e.g., $Variability\_Index$, see Luo et al. 2020). As observations advance, the features are changing with longer exposures, improvement of observational and statistical methods, and the identification or associate of partial sources. Just as the brighter training samples in variable star classification lead to sample selection bias in the classification of fainter stars (Richards et al. 2012; Richards 2012), there are systematic differences between Fermi associated and unassociated sources. Bright γ-ray sources are more likely to be bright at other wavelengths (radio, optical, X-ray) and therefore more likely to be detected in multiwavelength catalogs that are utilized to associate γ-ray sources. The sources of the associated sample that are used for training and testing the performance of the classification algorithms are generally brighter and detected at higher significance level. On the contrary, the unassociated sources were non-significant. This may lead to the potential problem that the classifier model is not ideal for a predicted target, even though it performs well on the test samples. Similarly, the systematic differences are reflected in the coordinate space. The unassociated samples were biased towards the sources near the Galactic plane, while the associated cases were widely distributed throughout the full-sky, especially in regions with high Galactic latitude. The large number of sources and highlighted backgrounds on the Galactic disk increase the difficulty of source detection. The source distributions in the significance space and Galactic latitude are depicted in Figure 8.

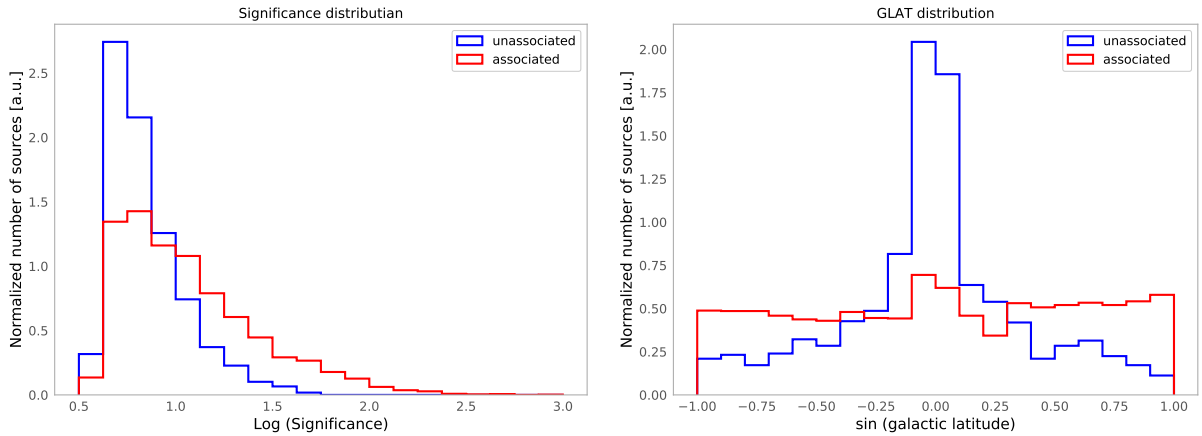

Fig. 8 The normalized source distribution in significance (*left*) and Galactic latitude (*right*).

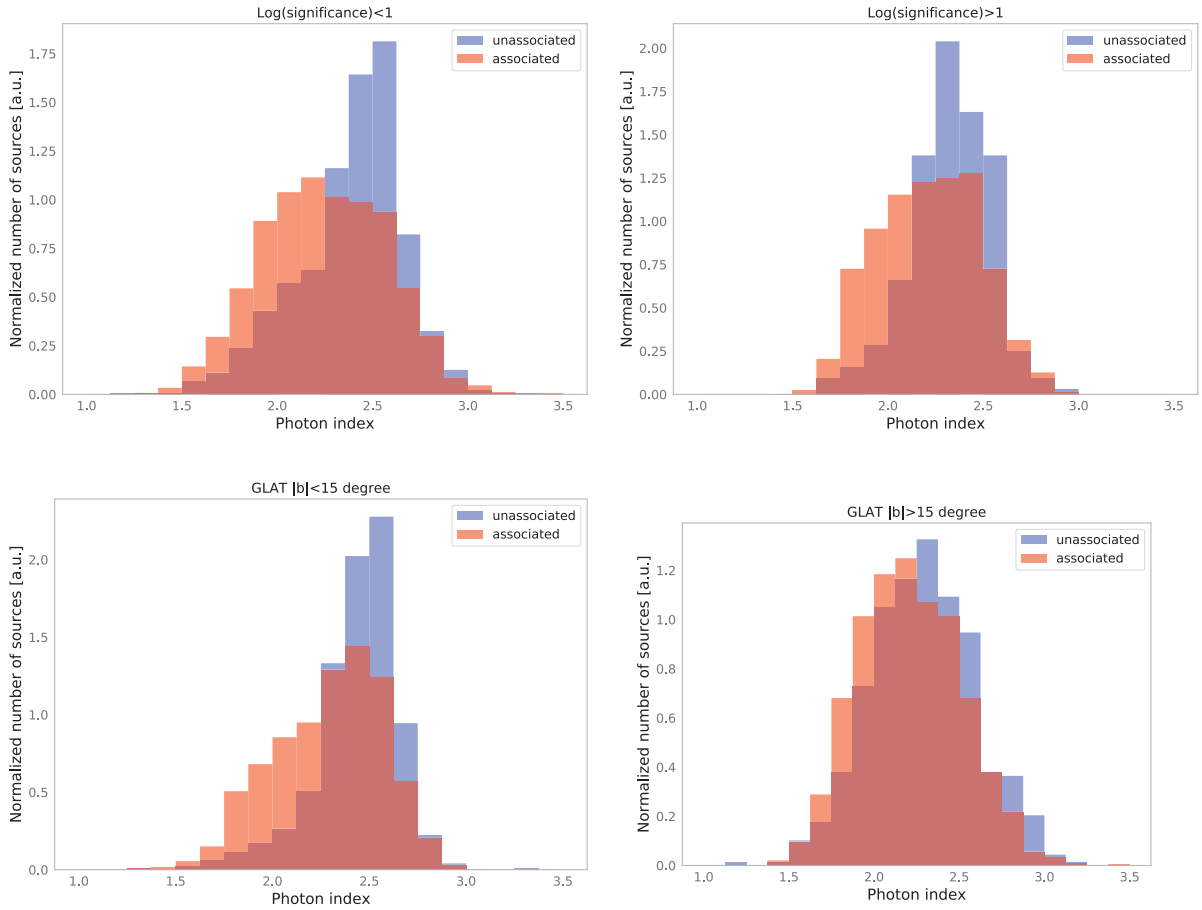

Fig. 9 The normalized distribution of photon spectrum index of darker sources, brighter sources, lower Galactic latitude sources and higher Galactic latitude sources. The darker part is the coincidence region of associated and unassociated samples.

For the purpose of clarifying the influence of systematic differences in distribution of the classification, we divided the Fermi sources into four groups: brighter sources, darker sources, higher Galactic latitude sources and lower Galactic latitude sources. Taking $Photon\_index$ as an example, the normalized distribution diagram is showcased in Figure 9. The spectral indexes of associated and unassociated sources located at higher Galactic latitude are similar in comparison with those of low Galactic latitude and therefore it is a “cleaner” dataset for using feature $Photon\_index$ for classification. For significance, the distribution differences of associated and

Table 6 The Classification of Unassociated Sources

4FGL_name	R.A.	Dec	$P_{\rm RF1}$	$P_{\rm RF2}$	$C_{\rm RF}$	$P_{\rm ANN1}$	$P_{\rm ANN2}$	$C_{\rm ANN}$	$C_{\rm com}$	$A_{\rm name}$	$C_{\rm LR-P}$	$C_{\rm RF-P}$
...	...	...	...	...	...	...	...	...	...			
4FGL J0530.0-6900e	82.50	-69.00	0.09	0.02	other	0.59	0.10	other	other	3FGL J0524.5-6937	AGN	AGN
4FGL J0531.7+1241c	82.94	12.69	0.63	0.32	other	0.81		agn	unc			
4FGL J0531.8-6639e	82.97	–66.65	0.83		agn	0.98		agn	agn	3FGL J0525.2-6614	AGN	AGN
4FGL J0532.6+3358	83.17	33.98	0.82		agn	0.99		agn	agn			
4FGL J0533.6+5945	83.42	59.76	0.28	0.98	psr	0.41	0.94	psr	psr	3FGL J0533.2+5944	PSR	AGN
4FGL J0533.9+2838	83.48	28.64	0.83		agn	1.00		agn	agn			
4FGL J0534.0+3746c	83.51	37.77	0.71	0.39	other	1.00		agn	unc			
4FGL J0534.2+2751	83.57	27.86	0.99		agn	0.97		agn	agn			
4FGL J0535.1-5422	83.78	–54.38	0.93		agn	0.97		agn	agn			
4FGL J0535.3+0934	83.84	9.58	0.98		agn	1.00		agn	agn			
4FGL J0536.1-1205	84.03	–12.09	0.90		agn	0.92		agn	agn			
4FGL J0537.5+0959	84.38	9.99	0.98		agn	1.00		agn	agn	3FGL J0537.0+0957	AGN	AGN
4FGL J0538.9+3549c	84.74	35.83	0.33	0.09	other	0.53	0.20	other	other			
4FGL J0539.2-6333	84.82	–63.55	1.00		agn	1.00		agn	agn			
4FGL J0540.0-7552	85.01	–75.88	0.90		agn	1.00		agn	agn	3FGL J0539.9-7553	AGN	AGN
4FGL J0540.2+0655	85.05	6.92	1.00		agn	0.99		agn	agn			
4FGL J0540.6+5540	85.17	55.67	1.00		agn	1.00		agn	agn			
4FGL J0540.7+3611	85.18	36.20	0.53	0.41	other	0.40	0.66	psr	unc			
4FGL J0543.5-8741	85.89	–87.69	1.00		agn	1.00		agn	agn	3FGL J0542.2-8737	AGN	AGN
4FGL J0543.9-0418	85.98	–4.31	0.81		agn	0.98		agn	agn			
4FGL J0544.4+2238	86.11	22.64	0.71	0.21	other	0.98		agn	unc	3FGL J0544.7+2239	AGN	AGN
4FGL J0544.8+5209	86.22	52.16	1.00		agn	1.00		agn	agn			
4FGL J0545.7+6016	86.44	60.27	0.16	0.93	psr	0.58	0.94	psr	psr	3FGL J0545.6+6019	PSR	PSR
...	...	...	...	...	...	...	...	...	...			

Column (1) shows the 4FGL names. The right ascension and declination of sources are listed in Cols. (2)–(3), respectively. Cols. (4)–(5) report the score given by ANN classifier for the first ($P_{\rm ANN1}$) and second ($P_{\rm ANN2}$) step. Sources with a step 1 score below the threshold 0.789 are classified as non-AGNs and brought into the step 2 classification. The classification ($C_{\rm ANN}$) given in the ANN is listed in Col. (6). Cols. (7)–(8) report the scores given by the RF classifier for the first ($P_{\rm RF1}$) and second ($P_{\rm RF2}$) steps. Sources with a step 1 score below the threshold 0.739 are classified as non-AGNs and brought into the step 2 classification. The classification ($C_{\rm RF}$) given in the RF is listed in Col. (9). Col. (10) lists the classification results of the two algorithms combined ("unc" means uncertain source). Col. (11) expresses the associated name ($A_{\rm name}$) in the other FGL. The cross-matching results for the 3FGL catalog's unassociated source classification results Saz Parkinson et al. (2016) obtained using logistic regression ($C_{\rm LR-P}$) and RF ($C_{\rm RF-P}$) are listed in Cols. (12) and (13), respectively. Table 6 is published in its entirety in machine-readable format (e.g., Table6_result.xlsx). A portion is shown here for guidance regarding its form and content.

Table 7 Comparison of 3FGL and 4FGL Results

Method	Label	Count[b]	Obtained predictions			
			agn	psr	other	unc
	Count[a]	334	96	73	24	141
LR^P	agn	146	84	6	4	52
	psr	188	12	67	20	89
RF^P	agn	163	87	9	4	63
	psr	171	9	64	20	78

Column (1): The classifiers obtained in Saz Parkinson et al. (2016); Cols. (2) and (3): The classification results for 356 common sources from Saz Parkinson et al. (2016), where the row count[a] expresses our classification results for 356 common sources; Cols. (4)–(7): Cross comparison results.

Table 8 Results from Two Supervised Classifiers for Simultaneous Classification of Three Different Types

Classifier	Features	Test				Prediction		
		$\rm Acc_{agn}$	$\rm Acc_{psr}$	$\rm Acc_{other}$	$\rm Acc_{overall}$	agn	psr	other
RF	20 features (See Table 2, left)	0.992	0.653	0.488	0.939	959	133	244
ANN	20 features (See Table 2, left)	0.998	0.673	0	0.917	1216	120	0
RF	8 features (See Table 22, right)	0.995	0.633	0.349	0.933	1112	106	118
ANN	8 features (See Table 2, right)	1	0.286	0	0.893	1221	115	0

Column (1): Classification methods; Col. (2): The used features; Cols. (3)–(6): Test results; Cols. (7)–(9): Prediction results.

unassociated samples in both brighter and darker sources are large, while the difference proportion in the brighter source is slightly smaller.

Due to the limitation of astronomical observation, sample selection bias is almost inevitable. A simple classifier with few features reduces the possibility of covariant shift (Luo et al. 2020). Using hardness ratios instead of direct observations like individual fluxes and variability index to keep information about the spectral shape, or modifying the observations to obtain more intrinsic features might solve the problem. Grouping the prediction samples and searching for suitable training samples to refine the classification process are less likely to encounter the problem of sample selection bias.

Table 9 The Influence of Sample Proportion on Learning Results

Sample method	Sample ratio ($n_{\rm non-AGN}/n_{\rm AGN}$)	RF			ANN		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
No	0.127	0.891	0.946	0.939	0.859	0.956	0.944
Undersampling	0.200	0.891	0.935	0.930	0.913	0.899	0.901
Undersampling	0.500	0.934	0.869	0.878	0.946	0.639	0.678
Undersampling	1.000	0.946	0.795	0.814	0.967	0.398	0.470
Oversampling	0.200	0.891	0.948	0.941	0.913	0.917	0.916
Oversampling	0.500	0.913	0.937	0.934	0.946	0.655	0.692
Oversampling	1.000	0.913	0.332	0.930	0.957	0.356	0.432
SMOTE	0.254	0.891	0.935	0.930	0.913	0.874	0.879
SMOTE	0.508	0.913	0.912	0.912	0.946	0.641	0.680
SMOTE	1.017	0.913	0.883	0.887	0.978	0.339	0.420

Column (1): Sample methods; Col. (2): The sampling ratio, that is, the ratio of the number of non-AGNs to the number of AGNs; Cols. (3)–(5): The performance of ANN classifier in sampling test; Cols. (6)–(8): The performance of RF classifier in sampling test.

Recently, Saz Parkinson et al. (2016) divided all of the 3FGL catalog's sources into AGNs and pulsars based on the logistic regression (LR) and RF algorithms. Cross-matching the 4FGL predictions (1336 unassociated sources) from the present work and 3FGL predictions (3033 sources, see Saz Parkinson et al. 2016[3]), we obtained 334 common sources (see Table 6). In the 3FGL predictions of common sources, 146 sources were classified as AGNs and 188 sources were classified as pulsars in LR^{P},[4]; 163 sources were classified as AGNs and 171 sources were classified as pulsars in their RF^{P} based classifier. In our predictions for classifier combination, 96 sources belong to the AGN type, 73 belong to the pulsar type and the number of sources classified as other or uncertain type are 24 and 141, respectively. Cross-matching the results (see Table 7), the majority of sources (approximately 89%) had the same classification for sources of AGN and pulsar types in our predictions. In Luo et al. (2020), they searched 20 millisecond pulsar candidates from the 4FGL unidentified sources employing a two-layer cascade method prompted by investigating the factors affecting ML classifications. Cross-matching the 20 millisecond pulsar candidates given by Luo et al. (2020), nine sources are evaluated as pulsars, two sources are classified as AGNs and nine sources are uncertain type in our predictions. In addition, we note that the nine pulsar candidates have higher significance in their results, while two AGNs exhibit lower significance. Most of our predictions are consistent with other previous results. However, the predictions of a fraction of sources are inconsistent, and the evaluation of their true classification needs further study in the future.

We have tried to put all of the unassociated samples (i.e., 1336) into the algorithms at the same time and classify them into three types directly. Although an overall accuracy of over 0.9 can be obtained (see Table 8), the approach has several limitations. Firstly, the result is unstable, especially for the other type, and the results produced by various classifiers are quite different. Secondly, the imbalance of the samples reduces the accuracy. Almost no predicting sample is classified as the other type, mainly resulting from fewer other type samples with insignificant characteristics. The presence of more AGN type of training samples leads to more unassociated samples to be evaluated as AGN type. For unbalanced samples, this can result in higher accuracy but it does not mean the classifier is good. Thirdly, there is a large difference in the selection of suitable features for different samples. For example, based on the results of the K-S test, "logF4" is the best feature in evaluation of the AGNs and non-AGNs, but it is not a good feature in the discrimination of pulsars and non-pulsars. In order to obtain a higher confidence level, we employ a step-by-step feature selection and classification approach at the expense of higher computational demands.

We have adopted a step-by-step classification strategy to reduce the large gap in the sample size. However, there is still a class imbalance issue even in the classification process, especially for the first stage of the AGN selection. Undersampling and oversampling are important statistical methods to solve the imbalance of samples. The SMOTE algorithm (Siriseriwan 2019) is a method to improve the oversampling by constructing new samples; this can reduce the overfitting consequences of oversampling to some extent. We have studied the effect of different sample proportions on the results (see Table 9) by different sampling methods (oversampling, undersampling and SMOTE). There are several important observations for these results. Firstly, in the optimal classifier model utilized in this paper, the application of a sampling method reduces the accuracy of the classifier. Secondly, in comparison with the ANN algorithm, the RF algorithm has better performance against sample change. Thirdly, in the oversampling, the sensitivity of the non-AGN samples does not increase after the increase of non-AGN samples,

[3] Also see `https://www.physics.hku.hk/~pablo/pulsarness/Step_08_Results.html`

[4] The logistic regression and RF model used in Saz Parkinson et al. (2016)

which may be due to the overfitting. However, overfitting has been avoided in the SMOTE method, which we plan to consider in future work.

There are some differences between these classifiers' results presented in the preceding section. These results should be treated with caution. Similarly, the accuracy of the second step is not as satisfactory as that of the first one, mainly because the uniqueness of the various sources is not significant enough due to the limited sample size. This also applies to the first step; although we try to expand the non-AGN sample, the gap is still too large. However, as the number of observations in catalogs progress, the situation can be gradually improved.

A potential drawback of this work is that the results are only obtained from the data in the 4FGL catalog. Due to the limitations of astronomical observations, the limited sample used to diagnose the classification of the 4FGL catalog's unassociated samples cannot be completely accurate, and the same applies to our results. In addition, the threshold value of feature selection and the details in the algorithm (random seed, etc.) can directly influence our results, which need to be further addressed in the future and are beyond the scope of this work. In addition, the impact of sample selection bias is only discussed but not resolved, which needs to be further addressed in future work.

ACKNOWLEDGEMENTS

We thank the anonymous referee for very constructive and helpful comments and suggestions, which greatly helped us to improve our paper. This work is partially supported by the National Natural Science Foundation of China (Grant Nos. 11763005 and 11873043), the Science and Technology Foundation of Guizhou Province (QKHJC[2019]1290), and the Research Foundation for Scientific Elitists of the Department of Education of Guizhou Province (QJHKYZ[2018]068).

References

Abdo, A. A., Ackermann, M., Ajello, M., et al. 2009a, Science, 325, 840

Abdo, A. A., Ackermann, M., Ajello, M., et al. 2009b, ApJS, 183, 46

Abdo, A. A., Ackermann, M., Ajello, M., et al. 2010, ApJS, 188, 405

Abdollahi, S., Acero, F., Ackermann, M., et al. 2020, ApJS, 247, 33

Acero, F., Ackermann, M., Ajello, M., et al. 2015, ApJS, 218, 23

Ball, N. M., & Brunner, R. J. 2010, International Journal of Modern Physics D, 19, 1001716

Banerji, M., Lahav, O., Lintott, C. J., et al. 2010, MNRAS, 406, 342

Baron, D. 2019, arXiv e-prints, arXiv:1904.07248

Bergmeir, C., & Benítez, J. M. 2012, Journal of Statistical Software, 46, 1

Bilicki, M., Hoekstra, H., Brown, M. J. I., et al. 2018, A&A, 616, A69

Breiman, L. 2001, Machine Learning, 45, 5

Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., & Mercurio, A. 2013, ApJ, 772, 140

Brescia, M., Cavuoti, S., Longo, G., & De Stefano, V. 2014, A&A, 568, A126

Calderon, V. F., & Berlind, A. A. 2019, MNRAS, 490, 2367

Cheng, K. S., Ho, C., & Ruderman, M. 1986, ApJ, 300, 500

Chiaro, G., Salvetti, D., La Mura, G., et al. 2016, MNRAS, 462, 3180

Das, P., & Sanders, J. L. 2019, MNRAS, 484, 294

Doert, M., & Errando, M. 2014, ApJ, 782, 41

Duda, R. O., Hart, P. E., & Stork, D. G. 2001, Pattern Classification (Wiley,)

Eatough, R. P., Molkenthin, N., Kramer, M., et al. 2010, MNRAS, 407, 2443

Ellison, S. L., Teimoorinia, H., Rosario, D. J., & Mendel, J. T. 2016, MNRAS, 458, L34

Faisst, A. L., Prakash, A., Capak, P. L., & Lee, B. 2019, ApJL, 881, L9

Feigelson, E. D., & Babu, G. J., eds. 2003, Random Forests: Finding Quasars, In: Statistical Challenges in Astronomy. Third Statistical Challenges in Modern Astronomy (SCMA III) Conference, University Park, PA, USA, July 18 - 21 2001. Eric D. Feigelson, G. Jogesh Babu (eds.). New York: Springer, ISBN 0-387-95546-1, 2003, 243

Feigelson, E. D., & Babu, G. J. 2012, Modern Statistical Methods for Astronomy

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. 2014, Journal of Machine Learning Research, 15, 3133

Fichtel, C. E., Bertsch, D. L., Chiang, J., et al. 1994, ApJS, 94, 551

Fluke, C. J., & Jacobs, C. 2020, WIREs Data Mining and Knowledge Discovery, 10, e1349 3

Harding, A. K., & Muslimov, A. G. 1998a, ApJ, 508, 328

Harding, A. K., & Muslimov, A. G. 1998b, in Neutron Stars and Pulsars: Thirty Years after the Discovery, ed. N. Shibazaki, 311

Hartman, R. C., Bertsch, D. L., Bloom, S. D., et al. 1999, ApJS, 123, 79

Hassan, T., Mirabal, N., Contreras, J. L., & Oya, I. 2013, MNRAS, 428, 220

Hermsen, W. 1981, Philosophical Transactions of the Royal Society of London Series A, 301, 519

Hosenie, Z., Lyon, R. J., Stappers, B. W., & Mootoovaloo, A. 2019, MNRAS, 488, 4858

Huertas-Company, M., Primack, J. R., Dekel, A., et al. 2018, ApJ, 858, 114

Hussain, K. F., Yousef Bassyouni, M., & Gelenbe, E. 2019, arXiv:1906.08864

Kabacoff, R. 2015, R in Action, Second Edition (Manning Publications Co.)
Kang, S.-J., Fan, J.-H., Mao, W., et al. 2019a, ApJ, 872, 189
Kang, S.-J., Li, E., Ou, W., et al. 2019b, ApJ, 887, 134
Kang, S.-J., Zhu, K., Feng, J., et al. 2020, ApJ, 891, 87
Karas, V., Svoboda, J., & Zajacek, M. 2019, arXiv:1901.06507
Lefaucheur, J., Boisson, C., Goldoni, P., & Pita, S. 2017, International Cosmic Ray Conference, 35, 600
Liaw, A., & Wiener, M. 2002, R News, 2, 18
Liodakis, I., & Blinov, D. 2019, MNRAS, 486, 3415
Luo, S., Leung, A. P., Hui, C. Y., & Li, K. L. 2020, MNRAS, 163
Mirabal, N., Frías-Martinez, V., Hassan, T., & Frías-Martinez, E. 2012, MNRAS, 424, L64
Naul, B., Bloom, J. S., Pérez, F., & van der Walt, S. 2018, Nature Astronomy, 2, 151
Nolan, P. L., Abdo, A. A., Ackermann, M., et al. 2012, ApJS, 199, 31
Parks, D., Prochaska, J. X., Dong, S., & Cai, Z. 2018, MNRAS, 476, 1151
Paul, E., & Utgoff. 1989, Machine Learning, 4, 161 7
Pesenson, M., Pesenson, I., McCollum, B., & Byalsky, M. 2010, in Proc. SPIE, Vol. 7740, Software and Cyberinfrastructure for Astronomy, 77400L
Pollock, A. M. T., Bloemen, J. B. G. M., Grenier, I. A., et al. 1987, International Cosmic Ray Conference, 1, 88
R Core Team. 2018, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria
Reis, I., & Baron, D. 2019, PRF: Probabilistic Random Forest
Reis, I., Baron, D., & Shahaf, S. 2019, AJ, 157, 16
Richards, J. W. 2012, Overcoming Sample Selection Bias in Variable Star Classification, eds. L. M. Sarro, L. Eyer, W. O'Mullane, & J. De Ridder, 2, Astrostatistics and Data Mining, eds. L. M. Sarro, L. Eyer, W. O'Mullane, & J. De Ridder, 2, 213
Richards, J. W., Starr, D. L., Brink, H., et al. 2012, ApJ, 744, 192
Robin, X., Turck, N., Hainard, A., et al. 2011, BMC Bioinformatics, 12, 77
Romani, R. W. 1996, ApJ, 470, 469
Romani, R. W. 2014, Science, 344, 159
Rudak, B., & Dyks, J. 1998, in Astronomical Society of the Pacific Conference Series, 138, 1997 Pacific Rim Conference on Stellar Astrophysics, eds. K. L. Chan, K. S. Cheng, & H. P. Singh, 281
Salvetti, D., Chiaro, G., La Mura, G., & Thompson, D. J. 2017, MNRAS, 470, 1291
Saz Parkinson, P. M., Xu, H., Yu, P. L. H., et al. 2016, ApJ, 820, 8
Saz Parkinson, P. M., Dormody, M., Ziegler, M., et al. 2010, ApJ, 725, 571
Siriseriwan, W. 2019, smotefamily: A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE, r package version 1.3.1
Teimoorinia, H., Bluck, A. F. L., & Ellison, S. L. 2016, MNRAS, 457, 2086
Thompson, D. J., Bertsch, D. L., Dingus, B. L., et al. 1995, ApJS, 101, 259
Ulrich, M.-H., Maraschi, L., & Urry, C. M. 1997, ARA&A, 35, 445
Urry, C. M., & Padovani, P. 1995, PASP, 107, 803
Vanzella, E., Cristiani, S., Fontana, A., et al. 2004, A&A, 423, 761
Xiong, D. R., & Zhang, X. 2014, MNRAS, 441, 3375
Zheng, Y. G., & Yang, C. Y. 2016, MNRAS, 457, 3535
Zheng, Y. G., Yang, C. Y., & Kang, S. J. 2016, A&A, 585, A8
Zheng, Y. G., Yang, C. Y., Zhang, L., & Wang, J. C. 2017, ApJS, 228, 1