PROCEEDINGS OF SCIENCE

ONLINE ICRC 2021
THE ASTROPARTICLE PHYSICS CONFERENCE
Berlin | Germany
37th International Cosmic Ray Conference
12–23 July 2021

Detection of new Misaligned Active Galactic Nuclei in the Fermi-LAT Fourth Source Catalog using machine learning techniques

Luca Deval,[a,*] **Fiorenza Donato**[b,c] **and Mattia Di Mauro**[b]

[a] *Institute for Astroparticle Physics, Karlsruhe Institute of Technology (KIT) Karlsruhe, Germany*
[b] *Istituto Nazionale di Fisica Nucleare Sezione di Torino,Turin,Italy*
[c] *Dipartimento di Fisica, Università di Torino Turin,Italy*

E-mail: luca.deval@kit.edu, donato@to.infn.it, dimauro.mattia@gmail.com

Active galactic nuclei (AGN) are the most luminous and abundant objects in the γ-ray sky. AGN with jets misaligned along the line-of-sight (MAGN) appear fainter than the brighter blazars, but are expected more numerous. Fermi Large Area Telescope (LAT) detected 40 MAGNs compared to 1943 blazars. The aim of this study is to identify new MAGN candidates in the blazars of uncertain type (BCUs) listed in the Fermi-LAT 10-year Source Catalog using an artificial neural network (ANN). The statistical tests applied to the trained ANN reveals that a classification with machine learning techniques is feasible with high accuracy and precision. The trained ANN has been applied to the 1120 BCUs which have been classified into 655 BL Lacs and 314 flat spectrum radio quasars (FSRQs). Among the re-classified BCUs, the possible MAGN candidates have been determined by applying thresholds on the spectral index, variability index and gamma-ray luminosity. Our results led to 36 possible MAGN candidates, which respect the main physical properties of the 40 MAGN already listed in the Fourth Fermi Catalog.

37th International Cosmic Ray Conference (ICRC 2021)
July 12th – 23rd, 2021
Online – Berlin, Germany

*Presenter

PoS(ICRC2021)667

Contents

1. Introduction

The *Fermi*-LAT had detected, over the years of observation, more than five thousand Galactic and extragalactic pointlike (and extended) gamma-ray sources. The increased number of observed sources has enlarged the population of known gamma-ray objects and broadened our comprehension of the Galactic diffuse emissions and extragalactic diffuse backgrounds. The observed sources have been collected in public catalogs which last released is the *Fermi*-LAT catalog Data Release 2 (4FGL-DR2) [2] based on 10 years of data acquisition. The applied analysis methods are the same as the 4FGL which led to the classification of 5788 sources. According to the unified scheme [1], Active Galactic Nuclei (hereafter AGN) are categorized as a function of their jet orientation: a jet which is misaligned of 14° with respect to the line-of-sight indicates a separation between blazars and non-blazars objects. AGN which axes are misaligned with respect to the line-of-sight (hereafter MAGN) have weaker luminosities but are expected to be more numerous although only 40 have been detected.

Given the complexity and size of astronomical dataset, machine learning algorithms have shown to be useful tools in particular when they have been applied to objects classification problem. Machine learning algorithm are used with the intention of learning a function (*target function*) $f(x) = y$ which associates to a multidimensional input x an output y [4]. The algorithm is based on models and techniques that provide the necessary information to determine an approximation of the target function without knowing it explicitly. A machine learning model consists in a parametric function's space $H = h_\theta$ where $\theta \in \mathbb{R}^N$ called *hypothesis set* and a optimization algorithm called *optimizer*. During the training phase, the optimizer tests the hypothesis space on the data searching for a configuration of the parameters $\bar{\theta}$ such that $h_{\bar{\theta}} \approx f(x)$. When the training phase is completed the model provides an algorithm able to use the approximation of the target function for estimating numerically y on previously unseen data. This can be applied to the statistical interference with the aim of approximating probability distributions. A common approach to the problem is to infer the conditional probability $P(y|x)$ with a discriminant model. In order to train such model data and target need to be provided: this machine learning's framework is called *supervised learning* [5].

The key factor regarding machine learning model is data. Data-sets need to be preprocessed and they need to be dimensionally suitable to extrapolate relevant statistical information. The preprocessing acquires a crucial role and it requires algorithms capable of mapping data in a new space which is easier to analyze. This process is named *features extraction* [6]. In the *Fermi*-LAT catalog [3] are listed 84 features related to every source of which only a few proved to be relevant regarding the classification of the blazars. The extracted features have been used to train the neural network and then they have been used to propose a classification of the BCUs. BCUs are blazars candidates of uncertain type therefore the neural network labeled the unclassified objects as BL Lacs and FSRQs. In order to find new possible MAGN among the re-classified BCUs, we imposed thresholds based on the physical properties of the 40 listed MAGN in the 4FGL [2] namely, the spectral index (Γ), variability index and γ-ray luminosity (L_γ). This selection led to the detection of 36 possible MAGN candidates.

2. Machine learning algorithm

Fully connected neural network can be visualized as graph-like structure in which nodes are referred as neurons which are located in layers and they are connected to all the neurons of the previous and following layers. Each connection between neurons has a weight and a bias term which indicates the strength of the connection. The input pattern is transmitted from the input layer to the next which is called hidden layer(s). After that, the output signal of each layer is used as the input signal to the next one, until the output signal of the last layer is computed. The output layer has as many neurons as the number of classes which have been provided. The neuron with the highest value, which is associated with a greater model confidence, determines the classification of the element in analysis. Therefore, the final output of the neural network is the likelihood of belonging to a given class. Finding the correct combination of weights is referred to as *training* which is accomplished with a gradient descent technique [8] and the employment of a loss function. The loss function shows the error between the provided labeled targets and the network's output. By using gradient descent the network computes the derivative of the loss function with respect to the weights and updates following the direction of the loss function gradient [6]. The value of the loss function is computed with a back-propagation algorithm in which the output values are compared with the correct answers and the error is then fed back through the network. By using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the loss function by small steps. This process is repeated for a sufficiently large number of training cycles which lead the network to a state where the error on the calculations is rather small. Once the network has been trained, it is compared with the test set which is a sample that is not used during the training process [7]. The agreement of the network's prediction with the labeled elements contained in the test sample is then evaluated considering the *confusion matrix*.

The model used in this work consists in a Feed-forward fully connected neural network with 4 layers: the input layer which is composed by 14 neurons, 2 hidden layers consisting of 30 neurons each and a final layer constituted of 2 neurons. The number of neurons in the input layer matches the selected features related to the analyzed sources. Regarding the number of neurons in the hidden layers, it was found that the total amount of neurons should be higher than the input features although an excessively high number of neurons would have led to overfitting. Moreover, in order to avoid overfitting the number of epoch has been selected with the early-stopping technique [7]. The data employed in the training and test set have been selected randomly, although the fluctuation in performance depending on which sources are considered might be important. In order to minimize the fluctuations, the model has been trained 100 times. Therefore, the values of the metrics used for the network's evaluation are mean values. The training set consists in the 70% (702 BL Lacs, 386 FSRQs) of the available data while the remaining 30% (375 BL Lacs, 208 FSRQs) have been used as test set. The desired outcome for training samples is set to 0 and 1 respectively for FSRQs and BL Lacs. Therefore, the output neurons return the likelihood (L) of a source belonging to either class such as $L_{BLLac} + L_{FSRQ} = 1$ for each source. The network is evaluated by computing the accuracy, recall, precision and AUC-score. A source is labeled as BL Lac if $0.8 < L_{BLLac} < 1$ while it is classified as FSRQ if $0 < L_{FSRQ} < 0.2$. The sources with a likelihood with values ranging between 0.2 and 0.8 are considered as unclassified. By applying the inputs to the network, we obtained 357 sources labeled as BL Lacs, 174 sources labeled as FSRQs and 52 sources were left unclassified. In Figure 1a is presented the histogram of the likelihood for BL Lacs and FSRQs. BL Lacs are concentrated towards a $L_{BLLac} \to 1$ while FSRQs have a $L_{FSRQ} \to 0$. The unclassified sources are concentrated in the region where the likelihood is ranging between 0.2 and 0.8. The accuracy related to the trained network is 0.91 which indicates that the 9% of the sources is left unclassified. The main discrepancy in the comparison with the test set is related to the number of FSRQs. The disagreement is potentially due to the difference in the number of available FSRQs in the training set compared to the number of available BL Lacs. The total of false positive and false negative is rather low in fact the precision has a value of 0.92 while the recall is 0.94. The AUC score is 0.97 which indicates not only a low number of false negative and false positive but also a rather good proficiency in characterize positive and negative classes.

The classification of labeled sources shown a rather good agreement with the labels presented in the 4FGL. Therefore we applied the trained network to unknown sources with the intent of presenting a classification into BL Lac and FSRQ. The proposed classification is achieved considering a 0.9 precision which implies that BCUs with a likelihood $L_{BCU} \in [0.8, 1]$ are labeled as a BL Lac while objects with a likelihood $L_{BCU} \in [0, 0.2]$ are labeled as FSRQ. The BCUs with a likelihood ranging from 0.2 and 0.8 are left unclassified. We used as input for the BCUs the same features used for the known sources and we classified the 1120 objects in 694 BL Lacs, 333 FSRQs while 91 sources have been left unclassified as presented in Figure 1b. The proposed classification of the BCUs shows that 62% of the sources has been labeled as BL Lacs while only 30% has been labeled as FSRQs. These results are, however, comparable with the ratio of classified elements listed in the *Fermi* catalog. In order to infer sources which could be consider as MAGN we applied on the re-classified BCUs physical constraints, which are based on the physical properties of the 40 MAGN listed in the 4FGL. By imposing the selection criteria to the re-classified BCU sample we obtained 36 MAGN candidates which meet the physical properties of the 40 MAGN listed in the 4FGL. In Figure 2a the 36 candidates are shown alongside with the sources employed in the analysis while in Figure 2b the candidates are shown only alongside the 40 MAGN listed in the catalog. It is important to notice that the proposed candidates characterized by an high value of L_γ entered a region densely populated by BL Lacs, therefore a further analysis will be conducted in order to investigate the robustness of their nature [9].

3. Conclusion

In this work we focused our attention on the study of MAGN with the aim of increasing their number. Increasing the number of available MAGN may lead to a better understanding of this γ-ray sources and moreover it could result in a better comprehension of the contribution of MAGN to the Isotropic Gamma-Ray Background(IGRB). With the intention of identifying possible new MAGN-candidate, we analyzed the BCUs that, due to their properties, may contain MAGN. In order to assess this hypothesis we employed a Feed-forward fully connected neural network which has been trained on the labeled BL Lacs and FSRQs listed in the *Fermi*-LAT catalog. The statistical tests applied on the trained network provide evidence that a classification with machine learning for the sources in the 4FGL is feasible. Therefore, we applied the trained network to unclassified objects and performed a classification. The categorization of the BCUs showed that by applying thresholds on the main features linked to the physics of MAGN it is possible to obtain valuable MAGN-candidates. As a result of this analysis 36 possible MAGN-candidates have been identified.

Bibliography

[1] Urry, C. M., & Padovani, P., *Unified schemes for radio-loud active galactic nuclei*, *Publications of the Astronomical Society of the Pacific* **107(715)**,803 (1995)

[2] Ballet, J., Burnett, T. H., Digel, S. W., & Lott, B., *Fermi Large Area Telescope Fourth Source Catalog Data Release 2*. **arXiv** :2005.11208 (2020)

[3] Abdollahi, S., Acero, F., Ackermann, M., Ajello, M., Atwood, W. B., Axelsson, M., & Siskind, E. J., *Fermi large area telescope fourth source catalog*. *The Astrophysical Journal Supplement Series* **247(1)**, 33. (2020)

[4] Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. T., *Learning from data*. New York, NY, USA: AMLBook. (2012)

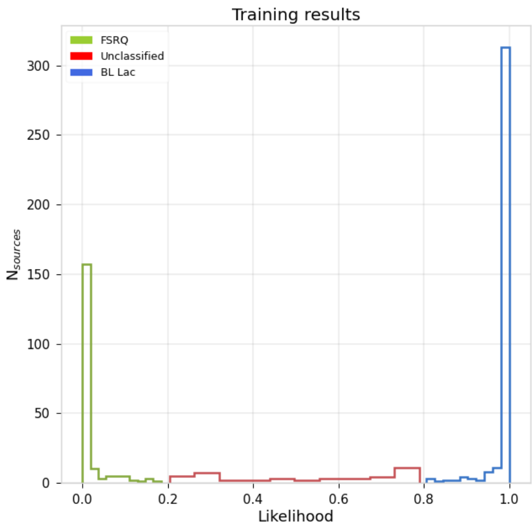
[5] Bishop, C. M. *Pattern recognition and machine learning*. Springer. (2006)

[6] LeCun, Y., Bengio, Y., & Hinton, G., *Deep learning*. *Nature* **521(7553)**. (2015)

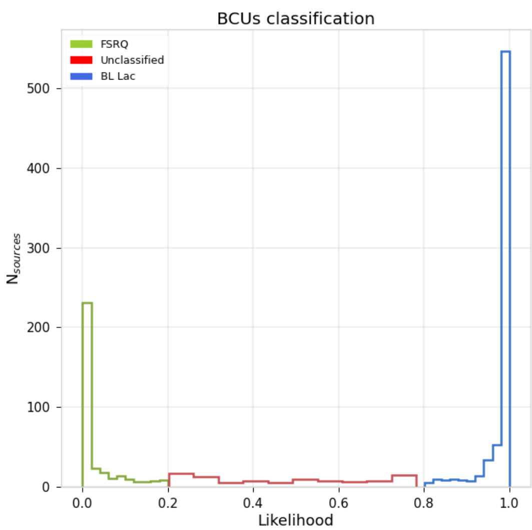
[7] Mehta, P., Bukov, M., Wang, C. H., Day, A. G., Richardson, C., Fisher, C. K., & Schwab, D. J., *A high-bias, low-variance introduction to machine learning for physicists*. *Physics reports*, **810**. (2019).

[8] Ruder, S., *An overview of gradient descent optimization algorithms*. **arXiv**:1609.04747. (2016).

[9] Deval, L., Donato, F., Di Mauro, M., *New misaligned AGN inferred by machine learning techniques*, To appear.

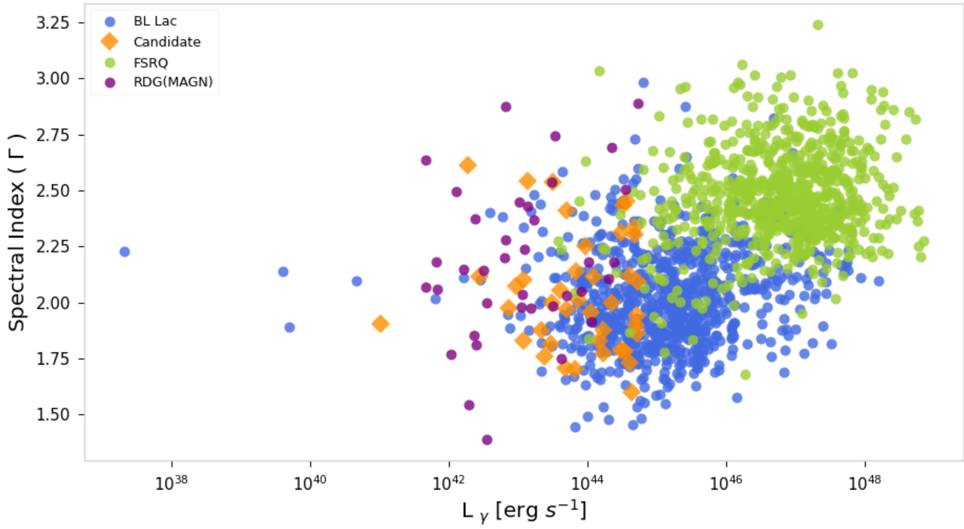


(a) Results for the trained neural network with known BL Lacs and FSRQs. The network has been tested with a sample composed of 381 BL Lacs and 202 FSRQs. The blue bars indicate the 357 sources classified by the ANN as BL Lacs, the green bars indicate the 174 sources labeled as FSRQs, while the red bars indicates the 52 sources left unclassified by the neural network.

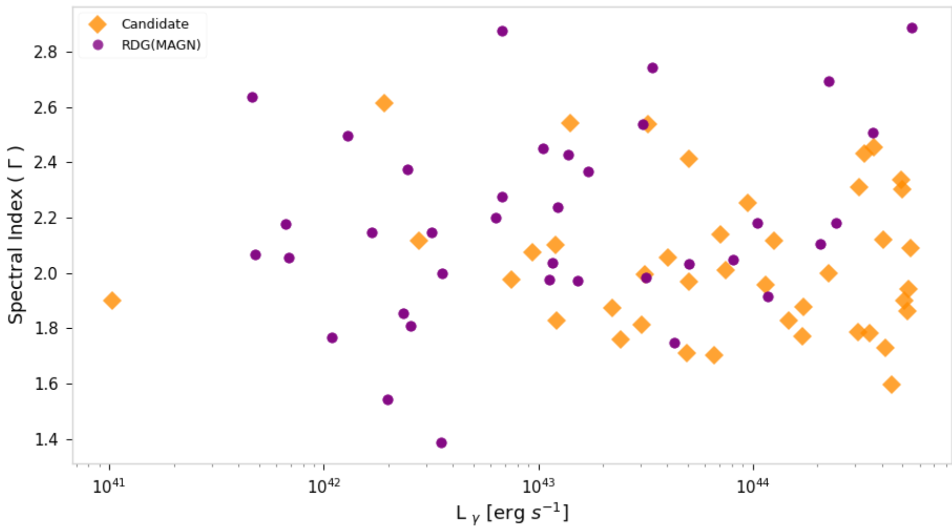


(b) Results for the classification of the 1120 BCUs in the Fourth *Fermi* catalog. The blue bars represent the 694 BL Lacs while the green bars represent the 333 FSRQs. The red bars indicate the 91 sources which are not classified by the neural network.

Figure 1



(a) Plot of the spectral index Γ as function of L_γ for BL Lacs (blue dots) FSRQs (green dots) and MAGN (purple dots) used through the analysis. Orange diamonds represent the MAGN-candidates obtained with the machine learning analysis



(b) Plot of the spectral index Γ as function of L_γ for MAGN (purple dots) used through the analysis and the MAGN-candidates obtained with the machine learning analysis.

Figure 2