Classification of gamma radiation source types using machine learning

Suzdalov Gleb

Abstract

3FGL (LAT 4-Year Point Source Catalog) is a catalog of gamma-ray sources detected by the Fermi Large Area Telescope (LAT) space gamma-ray observatory. These sources include both extragalactic and Galactic sources. Extragalactic sources include active galactic nuclei and their subcategory, blazars. Galaxies, on the other hand, include pulsars, which are notable for their gamma-ray emission. Some gamma-ray pulsars emit outside the gamma-ray range, making it difficult to detect their pulsations using ground-based telescopes alone. Searching for these pulses is a computationally intensive task [1], as the pulsation frequencies vary over time and the frequency of gamma-ray photons from the source are millions of times lower than the expected pulsation frequency. In order to optimize computing resources, we conducted a preliminary classification of sources using machine learning algorithms. This classification allowed us to identify potential targets for further investigation.

Introduction

The paper addresses the challenge of object classification, which is made difficult by the small size of the data set and its skewed distribution towards blazars. This can be seen in the figure 1. To solve this problem, three classification methods were employed: a forest-based classifier, adaboost, and neural networks. Additionally, when using neural networks for classification, data augmentation techniques were applied.

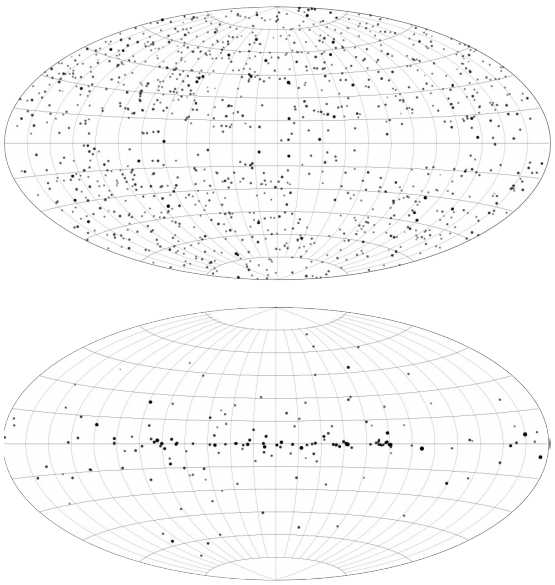

Figure 1: *Visualization of the dataset. Blazars (above) and pulsars (below) in the sky*

Classification

Data description

The dataset [2] contains information on 3034 objects, 1010 of which are unidentified. The first column contains the name of the object. Columns 2 and 3 contain the equatorial coordinates of the object in degrees. Columns 4-6 provide information on the parameters of the measurement error ellipse for the source. Columns 7-28 contain information on the spectrum parameters, while column 29 provides the variability index. Column 30 indicates the source code for identified sources or contains a value of "NULL" for unidentified objects.

Tree classifier

In order to categorize objects, a classifier based on a forest was employed. Two classes were created to implement this approach: a tree class, which builds a tree using a greedy algorithm, and a forest class, which contains multiple trees and provides an answer based on these trees. The results of this classification for the test sample are illustrated in the figure 2.

As can be observed from the graph, the algorithm performs well in classifying blazars, but it performs poorly in classifying pulsars. This is likely due to a skew in the dataset, as there are significantly more blazars compared to pulsars.

Boosting

In order to enhance the classification of pulsars, the AdaBoost machine learning algorithm was employed [3]. The results for the test dataset are shown in the figure 3.

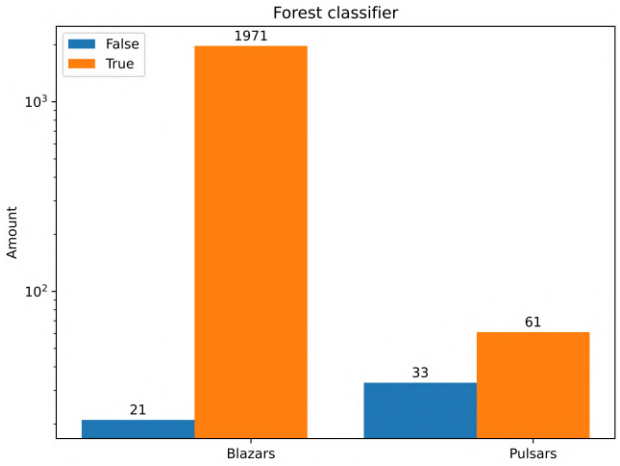

Figure 2: *Test sample classification with forest classifier*

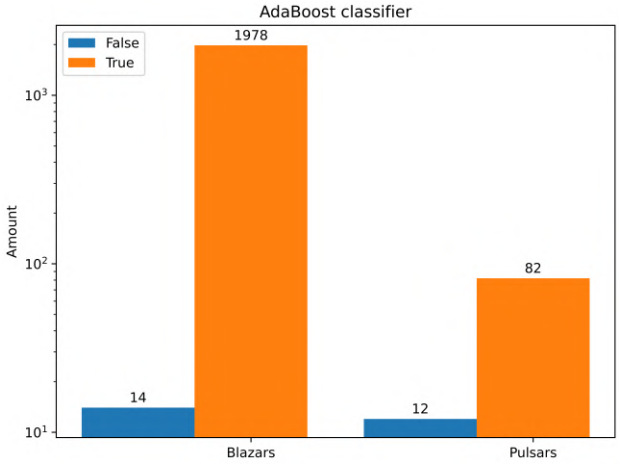

Figure 3: *Test sample classification with AdaBoost classifier*

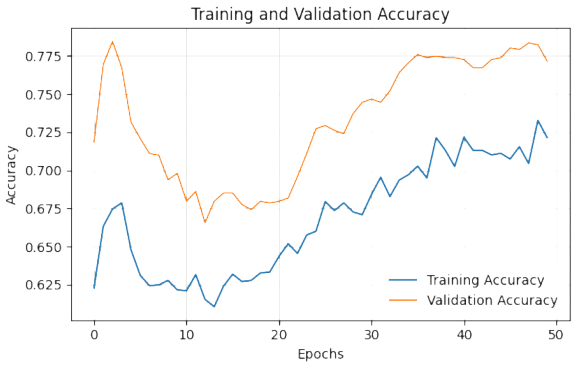

Figure 4: *Training and Validation Accuracy without data augmentation*

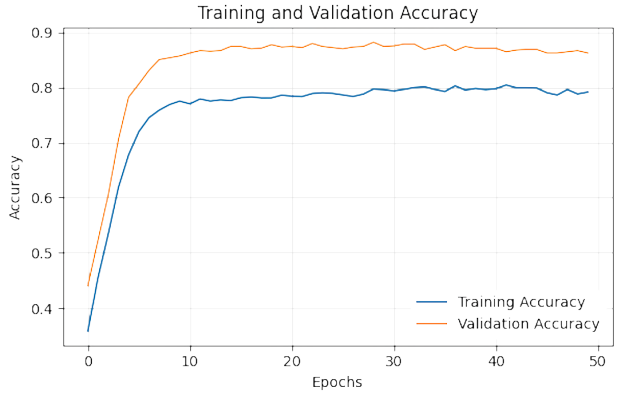

Figure 5: *Training and Validation Accuracy with data augmentation*

As can be seen from the charts, boosting has greatly improved the outcome, from which we can conclude that boosting may be beneficial for addressing skewness in the dataset. This is logical, as it increases the influence of objects that were previously misclassified. However, the accuracy of pulsars remains poor, with more than 10% of objects being misclassified.

Neural network classifier and data augmentation

In order to enhance the input data, the following approach was employed: given that physical measurements are subject to error, it was deemed feasible to additionally generate simulated data from pulsars, which would differ from the actual data within the bounds of the standard error of measurement.

The neural network is implemented using the curses library and has three layers with 64, 32, and 8 neurons, respectively. Categorical cross-entropy is used as the loss function, and accuracy is employed as the metric.

Figure 4 represents the results obtained without the use of data augmentation, while figure 5 represents the results achieved using data augmentation.

Results

During the course of the project, the issue of object classification was addressed in three different ways: utilizing a forest-based classifier, boosting, and neural networks. Based on the results obtained, it was found that boosting and a neural network combined with data augmentation achieved the best performance. Nevertheless, it should be noted that AdaBoost provides a finite level of accuracy comparable to that achieved through data augmentation. Therefore, this algorithm appears to be the most suitable approach for addressing this issue, as data augmentation may have a negative impact on the physical interpretation

of the data, and some parameters may be given excessive weight relative to their actual contribution to the overall picture.

References

[1] H. J. Pletsch et al. "Discovery of nine gamma-ray pulsars infermilarge area telescope data using a new blind search method". In: *The Astrophysical Journal* 744.2 (Dec. 2011), p. 105. ISSN: 1538-4357. DOI: 10.1088/0004-637x/744/2/105. URL: http://dx.doi.org/10.1088/0004-637X/744/2/105.

[2] Fermi Gamma-ray Space Telescope. "LAT 4-year Source Catalog". In: (2019-2023). URL: https://fermi.gsfc.nasa.gov/ssc/data/access/lat/4yr_catalog/.

[3] Yoav Freund and Robert E. Schapire. "A Short Introduction to Boosting". In: (1999). URL: https://api.semanticscholar.org/CorpusID:9621074.