fermitool: a project for the 'Computing Methods for Experimental Physics and Data Analysis' exam

Michele Messina Luana Michela Modafferi

February 26, 2020

Abstract

Develop a tool for a basic statistical exploration of the Fermi-LAT fourth source catalog (4FGL, available at `https://fermi.gsfc.nasa.gov/ssc/data/access/lat/8yr_catalog` in several different formats). The program should be able to display in a graphical fashion the basic characteristics of the sources in the catalog. In addition, use a classification technique of your choice to infer the source class (e.g., AGN or pulsar) for each entry in the catalog, and evaluate the performance of your classification scheme.

GitHub link to our solution: `https://github.com/micmes/SourceClassifier`

1 Libraries and tools

astropy · matplotlib · numpy · pandas · readthedocs · Sphinx · seaborn · scikit-learn · travis · unittest

2 Introduction

The LAT 8-year Source Catalog is currently available as a FITS file, made up of 8 binary table extensions. In this project, we used the LAT Point Source Catalog Extension, which contains all the relevant information about the 5064 sources, and the ExtendedSources extension, which contains detailed information about the extended sources only. We decided to read and perform operations on the data using the *pandas* DataFrame, a very powerful tabular data structure. The DataFrame is processed by a single object: the **Fermi_Dataset** class. We encapsulated all the methods and operations for the data inside this class in order to simplify both the code and the user's interface. We would like to clarify that the goal of this project is to *reproduce* some of the (mostly graphical) results obtained by The Fermi-LAT collaboration (2019). All the scientific statements can be found in the cited article.

3 The Fermi_Dataset class

When making an instance of the Fermi_Dataset class, we pass the DataFrame as the argument (**df**). The 4FGL data is a set of rows and columns where each row is a source and the columns are the variables associated to each source. The methods of the Fermi_Dataset are:

- `clean_column(), remove_nan_rows()` are methods acting on specific columns: the first one normalizes the text of string valued columns and the latter removes the rows that contain NaN values.
- `filtering()` returns a Fermi_Dataset with filtered data, based on a given condition.
- `def_column()` adds a column obtained as a result of a function that acts on two or more columns.
- `galactic_map(), source_hist()` are methods capable of plotting a generic map of the sources and a histogram.
- `dist_models(), plot_spectral_param()` focus on the spectral analysis of the sources.
- `compare_variability()` plots the comparison between the 2 month vs 12 month variability index.
- `classifier()` generates a Decision Tree with the purpose of classifying the unassociated sources.

For more details on each method, we recommend reading the docs.
Each method is tested with python's standard library *unittest*.

4 Discussion and Conclusions

Having opened the original FITS file in a DataFrame, the instance of a Fermi_Dataset object allows the user to generate localization maps, histograms and the spectral and variability characteristics of the sources based on different conditions. In order to make the program more interactive, we also developed a simple parser for command-line options using the *argparse* module. The current parser is a mere prototype of what we had in mind: the potentially infinite number of plots and the lack of time made us settle for a simple tool capable of generating a handful of plots (the ones we considered to be the most interesting).

Some of the outputs are shown on page 3.

The Variability plots are shown in Fig.1. The first subplot shows the distribution of the variability index over 1 year interval. Notice that the trend is that of a power-law overlapped with a χ^2 distribution. The second subplot is the comparison between the 12 month and the 2 month variability indexes. If we compare its trend with the variability thresholds it can be demonstrated that for the majority of sources using longer intervals detects variability better. Both the plots confirm the expected results obtained in the Fermi Large Area Telescope Fourth Source Catalog.

The Spectral plots are shown in Fig.2. The first subplot shows the distribution of the 3 models considered in the catalog: the PowerLaw model is by far the most popular. The second subplot shows the spectral parameters for the significant sources (`'Signif_Avg'` > 30 i.e. $TS > 1000$). As expected, we can see that sources belonging to the same class tend to form clusters in the plot.

The so-called Localization plots are probably the most visually appealing. In Fig. 3 we simply plot all the sources: their totality fill up the whole sky. In Fig. 4 we show only the pulsars (`psr`) and the pulsar wind nebulas (`pwn`), while in Fig. 5 we show their galactic latitude distribution. As expected, they are for the most gathered near the galactic plane. Figure 6 shows the error radius distribution at 95% confidence level. With this figure, we would like the reader to pay attention that our tool allows to manipulate the DataFrame columns before plotting: in this case, for example, it evaluates geometric mean. Fig. 7 shows only the sources of `ExtendedSources` extension; the points are colored accordingly to their semi-major axis value.

As mentioned in Section 3, the Fermi_Dataset features a method capable of inferring the source category thanks to a machine learning decision tree (made with *sklearn*). We chose this particular technique because it is simple to understand and able to handle both numerical and categorical data.

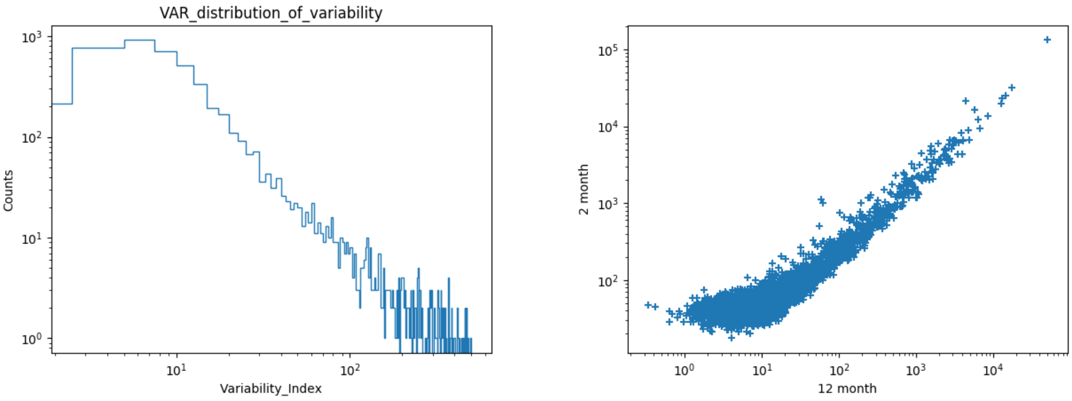
The first operation that was carried out was the cleaning of the data: we lowered all the letters in the CLASS1 column (this means that we don't discriminate the identified sources from the associated ones), we integer-encoded the data, we removed the underpopulated classes and finally filled the NaN values of the DataFrame with the mean value of the column to which the value belonged.

The model selection consisted in evaluating the *accuracy* generated with 5-fold cross-validation. The first attempt resulted in an overfitting model. We tweaked the model's parameters until we were sure that the model wasn't overfitting (see Figure 8). In particular we limited the maximum depth of the tree to 5 and the leaf nodes to 10 (this procedure is called *pruning*). With our new model at hand, we train it on the entire dataset, obtaining an accuracy of 0.95. We can use the classifier to infer the unassociated sources. The predictions are:

class	n. sources predicted
unknown	1333
pulsar	3

References

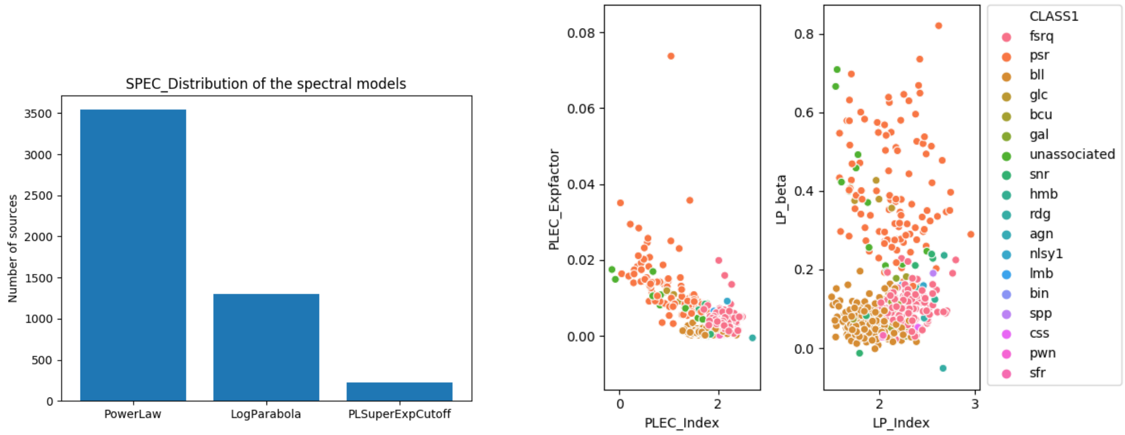
The Fermi-LAT collaboration. (2019, Feb). Fermi Large Area Telescope Fourth Source Catalog. *arXiv e-prints*, arXiv:1902.10045.



(a) Variability Distribution over 1 year intervals

(b) Comparison of the Variability Indexes

Figure 1: Variability Plots



(a) Spectral Model Distribution

(b) Spectral Parameters for most significant sources ('Signif_Avg' > 30)

Figure 2: Spectral Plots

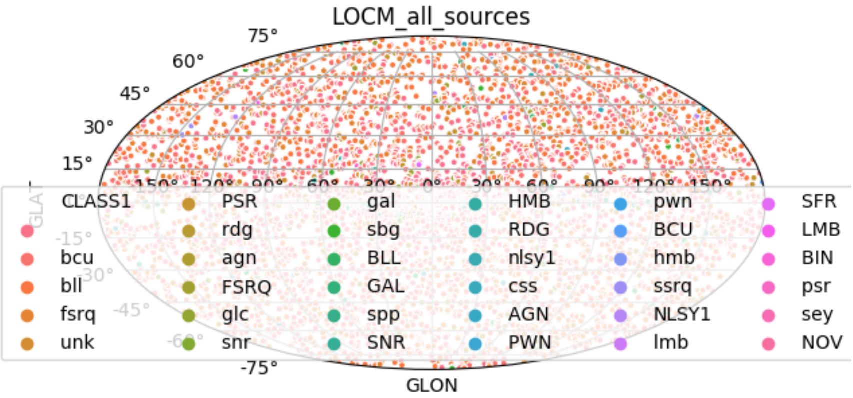
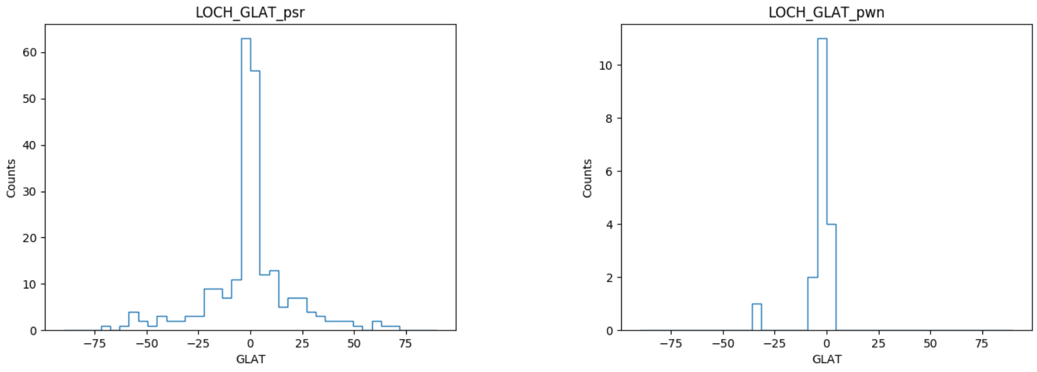


Figure 3: Full Sky Map



(a) Galactic Latitude for pulsars

(b) Galactic Latitude for pulsar wind nebulas

Figure 5: Latitude histograms for `psr` and `pwn`

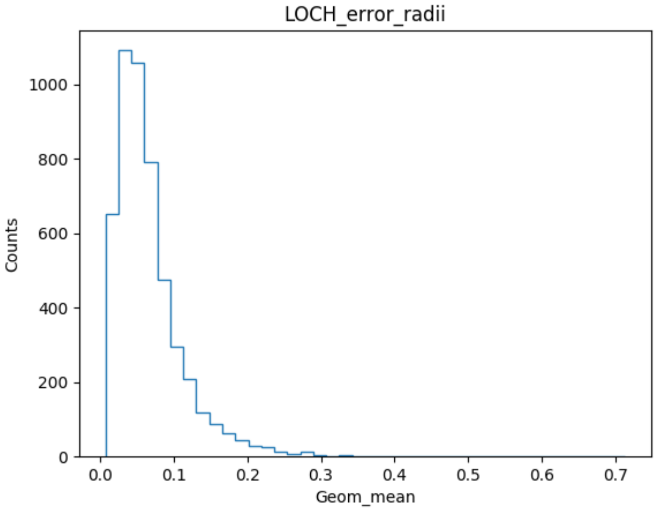


Figure 6: Error radius distribution at 95% confidence level.

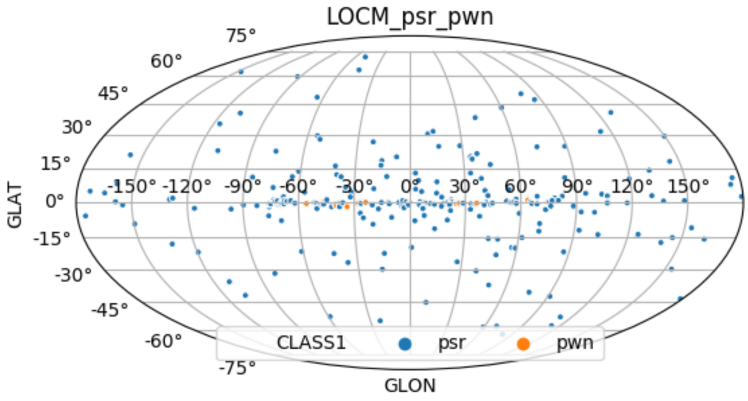


Figure 4: Galactic Map of pulsars and pulsar wind nebulas

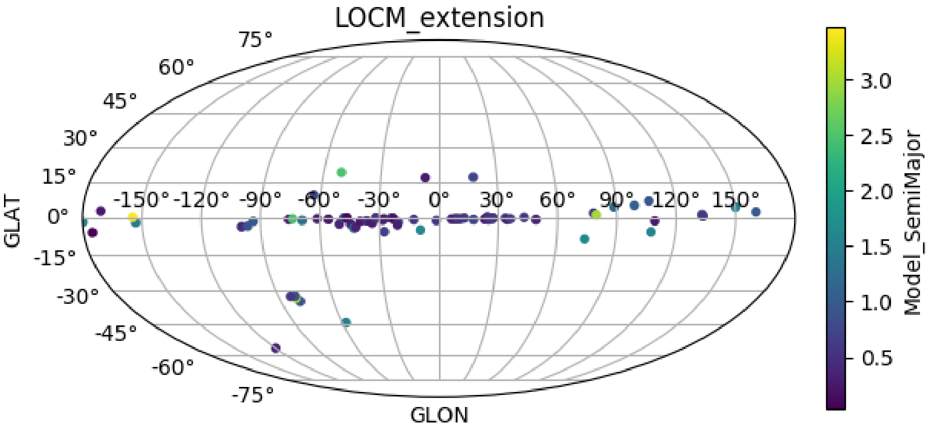


Figure 7: Extended sources colored with their semi-major axis in degrees.

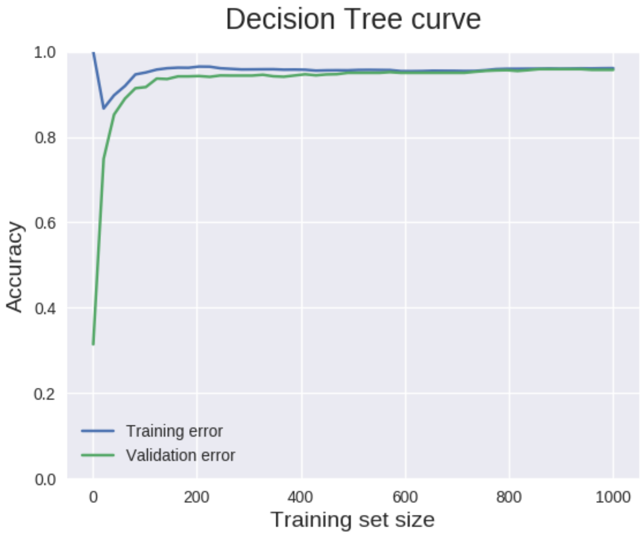


Figure 8: Accuracy of our Decision Tree obtained with cross validation.