Tata Institute of Fundamental Research

Departmental Project I

A Scheme to Determine Nature of Unassociated Sources in Fourth *Fermi*-LAT Catalog

submitted by
Bihan Banerjee

Supervised by
Dr. Sudip Bhattacharyya

DAA,TIFR
July 23, 2020

1 Background

The Fermi Gamma-ray Space telescope is an international space mission that studies the universe in the energy range 10 keV - 300 GeV. Compared to previous gamma ray telescopes *Fermi* is vastly more capable. The main instrument is called Large Area Telescope(LAT). It is active since June, 2008, and surveying the sky since then. Analyzing all sky data, *Fermi*-LAT collaboration has published 4 point source catalogs. The latest being the fourth(we will refer to it as 4FGL), published in Jan, 2020.(1)

The point source catalog has reported total 5064 sources, in the energy range 50 MeV-300 GeV. Among these, 3636 sources are associated with counterparts in other wavelengths,and their nature has been identified. Among the identified sources, more than 3130 are identified as Active Galactic Nuclei of different kinds. 239 pulsars have been detected. Among other identified sources are Globular cluster, Nova, Supernova remnant, pulsar wind nebulae, binaries etc.(1) For the rest 1428 sources, no counterpart has been found yet, and nature of these sources are undetermined.

To identify these sources, further observations in gamma ray and other wavelengths are obviously needed. However,if we could indicate which source belong to which class beforehand, it would be very useful for further observations. For example, if we somehow find, one particular source is probably a pulsar, then observations can be planned accordingly.

Ackerman(2012)(2) was the first who took a statistical approach to identify the unassociated LAT sources, after publication of 1FGL catalog.(3) It was shown that, in various parameter space, different gamma ray sources cluster up in different regions. Based on these properties, they tried to separate out AGNs and non-AGNs and pulsars and non-pulsars from 1FGL catalog. They also implemented two machine learning methods, decision trees and logistic regression, which was only 57% accurate.

Since then, different machine learning methods to classify the sources has been investigated by different people. However, the most important step was taken by Saz.Parkinson(2016)(4). In this work Random Forest Classifier was first used alongside other methods, for classification of unassociated sources in third *Fermi* catalog (3FGL). It was found that random forest classification can provide 90% accuracy, which is better than the other machine learning methods being used. Parkinson classified the sources into three categories, AGN, pulsar and others. Then using Fermi pulsar catalog, classification of predicted pulsars into two subcategories (Young/millisecond pulsars) was made.

We mainly follow Parkinson's steps to classify the unassociated sources in 4FGL catalog. We have used different set of parameters,(2) and different codes(6), and classified the sources with 95% accuracy. The method is described here(3), and the results are discussed here(4).

Then we took one more step, we have taken 4FGL catalog, and looked for *Chandra* sources within error circles of each *Fermi* source. Then we searched for *Gaia* counterparts within the Chandra error circles. In this way we have made a positional cross match between three catalogs, and obtained *Chandra* and *gaia*

counterparts to a set of Fermi sources. We have been able to obtain counterparts of some unknown sources too, which our codes predict as pulsars.(4).

2 Data

2.1 Instruments, Data products and access

2.1.1 Fermi

Fermi Gamma-ray space telescope is a satellite observatory, launched on June 11,2008. It circles earth every 96 minutes in a 26°at an altitude of 535 km. It carries two main instruments. Large Area Telescope(LAT) and Gamma Ray Burst Monitor(GBM). The observatory can sky the entire gamma-ray sky in every two orbits. LAT detects stable gamma ray sources in the sky, and GBM records the transient events. As mentioned earlier, based on 8 year data(2008-16) collected by LAT, in energy range 50 MeV to 300 GeV, 4FGL was prepared. The catalog is available here: https://fermi.gsfc.nasa.gov/ssc/data/access/ in FITS format.
The Fermi Pulsar catalog is accessed from this link: https://fermi.gsfc.nasa.gov/ssc/data/access/lat/2nd_PSR_catalog/ and the corresponding paper can be found here. (6)

2.1.2 Chandra

Chandra is a spacecraft,carrying X-ray telescopes,launched in July, 1999. Because X-rays are absorbed by Earth's atmosphere, Chandra orbits above it, up to an altitude of 139,000 km in space in an highly eccentric orbit. It can operate in energy range 0.1-10 KeV. Chandra's spatial resolution is $\leq$ 1 arcsec.
We have used Chandra data only to spatially correlate with Fermi sources. For this purpose, we have accessed Chandra catalog through python module astroquery.esasky.

2.1.3 Gaia

Gaia is a space observatory of European Space Agency(ESA), launched in 2013. Gaia can operate in wavelength range: It's main mission is to create a three dimensional map of our galaxy. It's spatial resolution is sub milli-arc second. So it can locate optical sources within our galaxy very precisely. In addition, Gaia also measures parallaxes and using that has obtained distances to multiple sources within the galaxy.[For details see(b)]. We are mainly interested in the distance catalog. Our hope is that, for the high energy sources inside our galaxy, if we can find a counterpart in Gaia distance catalog, we can associate the distance to that high energy source. In this manner, we may find distances to the Gamma ray sources within our galaxy. Second data release of Gaia occured on 25th April, 2018, and we have used that in our work.

We have accessed gaia distance catalog from Vizier, using the python module astroquery.vizier. The catalog id is I/347/gaia dr2.

2.2 Data Preparation for RandomForest Classification

2.2.1 Assigning three classes to all sources:

Not all columns of 4FGL catalog is useful for classification. First, we selected necessary columns from the FITS file and exported it into a .csv file. In 4FGL, there are variety of source classes, but we need only 3 for the first step. Further classification or complexity should be dealt with in next steps. Therefore, we convert all Active Galactic Nuclei subclasses as AGN. The subclasses being: Flat Spectrum Radio Quasar(FSRQ), BL Lactre type Blazar(BLL), Blazar of uncertain type(BCU), Steep Spectrum Radio Quasar(SSRQ), Seyfert Galaxy(SEY), Narrow Line Seyfert Galaxy(NLSY1), Compact Steep Spectrum Radio Source(CSS), other type of AGNs(AGN). For all these sources, we assign a common source class "AGN".
On the other hand, Globular Clusters(glc), binaries(BIN), Low mass binaries(LMB), High mass binaries(HMB),Supernova remnants(SNR), Pulsar wind nebulae(PWN),Nova(NOV), Star Forming Region(SFR), Either of SNR or PWN(SPP), Normal Galaxies(Gal), is assigned with a common source class "others". Lastly, all the Pulsars has the common source class "PSR".
All the Unassociated sources,(in 4FGL "unk" or empty in class field), is assigned with common class "unk".

2.2.2 Feature Selection:

The parameters that we finally use for source classification are called features. For classification, we have not used any positional data,("RA","DEC","GLON" and "GLAT") since that can introduce certain bias and wrong prediction. The features we finally use are described below:

- **Pivot Energy**:The energy, in MeV, at which the error in the differential photon flux is minimal (i.e., the decorrelation energy for the power-law fit). This is derived from the likelihood analysis for 100 MeV - 1 TeV, denoted as E_0 hereafter. It is important for spectral fits, as described below.

- **Flux Densities**: In 4FGL, each source is observed in 7 different bands. Each of the source is then fitted using three models, namely, powerlaw, Log parabola, and power law with an exponential cutoff. The models are given by:

  1. **Powerlaw**

$$\frac{dN}{dE} = k \left(\frac{E}{E_0} \right)^{-\Gamma} \tag{1}$$

2. **Logparabola**

$$\frac{dN}{dE} = k\left(\frac{E}{E_0}\right)^{-\alpha-\beta\log(E/E_0)} \tag{2}$$

3. **Power law with exponential cutoff**

$$\frac{dN}{dE} = k\left(\frac{E}{E_0}\right)^{-\Gamma}\exp\left(a(E_0^b - E^b)\right) \tag{3}$$

This k is known as flux density. For each source, we have used three different flux density from three models, and their uncertainties as two features of the model. However, it turns out, neither k, nor it's uncertainty is an important feature.

- **Spectral Indices**: In eq(1), Γ is the powerlaw index(PL index in catalog). From eq(2), α which is the spectral slope at E_0 is called LP index in the table. β is simply LP beta. Most of the pulsars are fitted with power law with exponential cutoff. Γ is PLEC index, a is PLEC expfactor and b i PLEC exp index, as they appear in table. All of these are used as features to our model.

- **Curvature Significance**: If $\mathcal{L}$ is likelihood function, then curvature significance is defined as
$\mathrm{TS}_{\mathrm{curve}} = 2\log(\mathcal{L}(\text{curved spectrum})/\mathcal{L}(\text{Powerlaw Spectrum}))$.
Where curved spectra can be either of logparabola or power law with exponential cutoff.

- **Variability Index:** Suppose F_{av} is average flux from a source in 1 year interval, and F_{glob} is the total average flux, from 8 year data. Then Variability index (TS_{var}) is defined by:

$$TS_{var} = 2\sum_i \log\left[\frac{\mathcal{L}_i(F_i)}{\mathcal{L}_i(F_{glob})}\right] - max(\chi^2(F_{glob}) - \chi^2(F_{av}))$$

$$\chi^2 = \sum_i \frac{(F_i - F)^2}{\sigma_i^2}$$

where F_i are the individual flux values, $\mathcal{L}_i(F)$ is the likelihood in interval i assuming flux F and σ_i as the error on F_i.
On the other hand, fractional variability is defined as excess variance on top of statistical and systematic fluctuations for each source:

$$V = \frac{1}{N-1}\sum_i (F_i - F_{av})^2$$

$$\delta F = \sqrt{max(V - \frac{\sum_i \sigma_i^2}{N}, 0)}$$

$\delta F/F_{av}$ is defined as the fractional variability.

- **Hardness ratios** : Gamma ray Photons are collected in seven bands: (1)50 MeV to 100 MeV (2)100 to 300 MeV (3) 0.3-1 GeV (4) 1-3 GeV (5) 3-10 GeV (6)10 GeV to 30 GeV (7) 30 GeV to 300 GeV.
  Hardness ratios between two bands are defined by:
  $h_{ij} = \left| \frac{E_i - E_j}{E_i + E_j} \right|$

All these features are inputs of the data. Based on the values of these features, output("class") is predicted. The method is described in next Section.

3 Method

Our work has two separate parts. (1)Prediction of nature of the unknown sources using random forest classification technique.(2)Obtaining counterparts of the predicted sources in other wavelengths by positional cross match. These two methods are described below.

3.1 Random Forest Classification Algorithm

To understand random Forest algorithm, first we need to understand decision trees. Consider a tree. It starts from a root, and then branches separate out. From one branch, more branches appear. The junction point of two branches is called a node. (Denoted by a circle in the figure). Eventually each branch end in leaves.(see figure(1)).

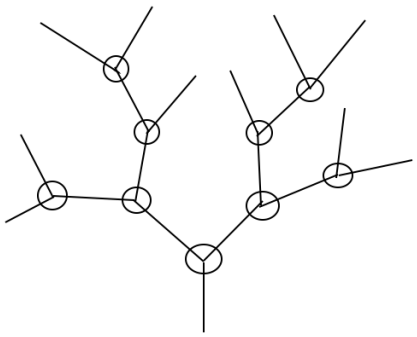

Figure 1: A decision tree

We can break down any classification problem into a series of yes no questions. Now let's again look at the tree. Each node now represents a question(based on the problem). The root, starts with an initial question. And each node asks a different question. If the answer is "yes" we go to the right branch. If the answer is "no", we take the left branch. So depending on answers at each node, we end up at different leaves. These leaves, indeed, represent different

classes.
However, this method is not very accurate, and false classification rate is high. Because of this, instead of a single decision tree we prefer to consider a forest of trees.
In this case, from the data, randomly different subsets of the samples are selected. From each of this subset, a decision tree is made. Given an input, each decision tree predicts a result. The result which gets most voted is shown as the final answer.(See figure 2)

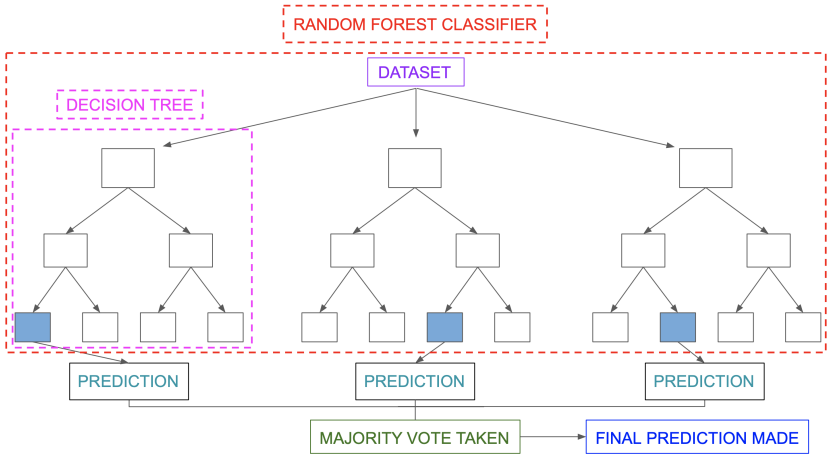

Figure 2: Random Forest of Decision Trees

It can be shown that Randomforest classification has much greater accuracy than simple decision trees. This algorithm is embedded in python library scikit-learn (or sklearn), which we have used.(6)

3.1.1 Implementation in Python

Let's consider our data (2). It has two parts. Input, and output. As input, we have set of parameters described in 2. And as output we have three different classes: AGN, PSR, and others. Now, due to measurement uncertainties, some fields in the input sample are empty or "NULL". First we dropped all such rows which has at least one NULL value. Then, we have checked if the input parameters we are using are highly correlated. For this business, we have performed a t-test, using Scipy.stats.ttest_rel, and omitted the highly correlated features($|\rho| > 0.7$). It turned out, two hardness ratios(h_{67} and h_{23}) were highly correlated, with a t-statistic pvalue $|\rho| > 0.7$, so we dropped h_{23}.
With the rest of the features, we divided all the sources randomly into two parts. 70% we took as training sample, and 30% to test the performance. The whole random forest algorithm is embedded in scikit-learn's RandomForestClassifier. Specifiying n_estimators we can specify the number of trees in the forest. To test

the performance, we ask the system to predict the class of the testing sample. Then we compare the predicted class and the actual class of the testing sample, and find, our result is 95.4% accurate.
When building up decision trees, not all features play equal role. To visualize important and unimportant features we calculate feature_importances_, the Table and Plot is shown in next section4. We have found, The flux densities and their uncertainties are not important at all, we could have as well omitted them in the first place.
Finally with the training model ready in hand, we pass the input data of unknown samples, and the code predicts 1030 AGNs, 106 pulsars and 106 other sources. For each source, probability of being in three categories are computed. For details see here.
Now we come to the next part. **Determining subclasses of the pulsars**. For this work, first we access the Fermi pulsar catalog, and cross match positions of these pulsars with pulsars in 4FGL using CDSXmatch(http://cdsxmatch.u-strasbg.fr/), with maximum angular distance 120 arc sec(Upper limit of CDSXmatch). We find 105 such matches.
In this catalog, each pulsar is divided into three classes, Young Radio Quite, Young Radio Loud, and Milli-second pulsars(MSP), as discussed in earlier section. Since we are only dealing with gamma ray data, we convert both radio Quite and radio loud sources into a single class, Young Pulsars(YNG). Now our question is, **based on only gamma ray observational data(4FGL) can we classify the predicted pulsars into these two categories?(YNG/MSP)**
We keep all the previous feature to train our model, with an addition. We observe that Milli-Second Pulsars and Young pulsars have different distributions over Galactic Longitude. So we use galactic Longitude as an additional feature. We repeat the same process again. This time, due to a smaller sample size, accuracy is smaller, 87% only. The feature importance plot and table is shown in next section.(4). Our code predicts, 48 young pulsars and 59 Milli-second pulsars. For details, see here.

3.2 Positional Cross-matching

With this statistical approach in one hand, it is also important to look for counterparts of unrecognized 4FGL sources in other wavelengths. In this work we are mainly interested in obtaining Gaia Counterparts of LAT pulsars and predicted pulsars. Since distances to sources in Gaia dr2 distance catalog is already known, by correlating, we can obtain distances to 4FGL sources. Although, we must keep in mind, only positional cross match is not enough to associate a gamma ray source to another source in a different wavelength. nevertheless, it is an important first step.
One difficulty to correlate 4FGL catalog with *Gaia DR2* source catalog is that, resolution of LAT is much poor(few arc sec) but *Gaia* instruments have sub milli arc second spatial resolution. So LAT's error circle is much bigger. And within error circle of LAT, in principle, there should be many many *Gaia* sources. So we approach the problem differently. First we search for *Chandra* counterparts

within error circles of each of the sources in 4FGL. *Chandra* has much better resolution than *Fermi* instruments. Hence, Chandra error circles are relatively small. Then we search *Gaia* sources inside these *Chandra* error circles.
We wrote the code in python(here). We have used astroquery module to query *Chandra* and *Gaia* catalogs.
Gaia can measure distances of the sources within our galaxy. So, trying to find *gaia* counterpart of Active Galactic Nuclei is meaningless. In this work, we are mainly interested in obtaing distances to the pulsars in our galaxy. So we keep only the pulsars from 4FGL, and the pulsars predicted by RandomForest Classifier, and look for their counterparts. Distances to some *Fermi* pulsars were already known, we have compared these distances with *Gaia* distances. The results are here

4 Results

4.1 AGN/PSR/Others

4.1.1 Feature Importance

The Features used for Random Forest Classification of unrecognized 4FGL sources, and their importance in the classification is listed below. Corresponding bar chart is given in figure (3).

Feature No	Feature Name	Importance
0	Pivot energy	0.04509
1	PL Flux density	0.00000
2	unc PL Flux Density	0.00000
3	PL index	0.02570
4	Unc PL index	0.04996
5	LP Flux density	0.00000
6	Unc LP Flux density	0.00000
7	LP index	0.02672
8	Unc LP Index	0.03648
9	LP beta	0.05756
10	unc Lp beta	0.03437
11	LP sigCurve	0.10370
12	PLEC flux density	0.00000
13	unc PLEC flux density	0.00000
14	PLEC index	0.03733
15	unc PLEC index	0.03281
16	PLEC expfactor	0.05794
17	unc PLEC expfactor	0.03570
18	PLEC exp index	0.00018
19	PLEC sigcurve	0.09485
20	Npred	0.07071
21	variability index	0.04282
22	frac variability	0.04097
23	unc frac variability	0.01805
24	Variability2	0.04034
25	frac Variability2	0.02623
26	Unc frac Variability2	0.01318
27	h12	0.01932
28	h34	0.02094
29	h45	0.01502
30	h56	0.03402
31	h67	0.02002

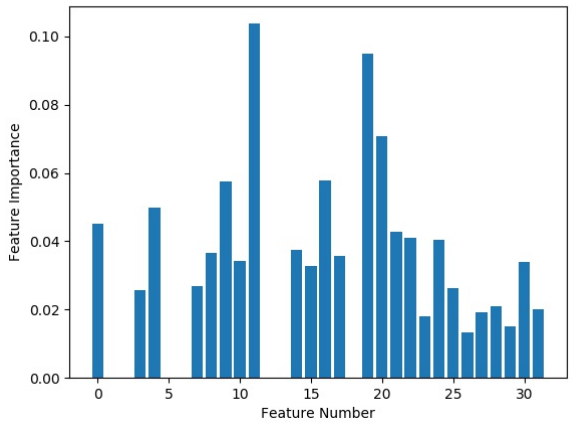

Figure 3: Feature Number vs Feature Importance Graph

4.1.2 Testing Random Classifier Model

First few rows of the results are shown below. The entire file can be found here.(testresult.txt) Prob_class denotes probability of the source to belong to that class.

Actual class(y_t)	Predicted Class(y_p)	Prob_PSR	Prob_AGN	Prob_others
agn	agn	0.0002	0.999	0.0008
agn	agn	0.0018	0.9954	0.0028
agn	agn	0.0	0.999	0.001
PSR	PSR	0.6478	0.173	0.1792
others	others	0.036	0.3698	0.5942

4.1.3 Predicting Classes of unknown Sources

Some of the results are shown below, the entire file can be found here.(newcat.csv)

Source Name	Predicted class	Prob_PSR	Prob_AGN	Prob_others
4FGL J0000.3-7355	agn	0.0002	0.9944	0.0054
4FGL J0002.1+6721	agn	0.0894	0.6962	0.2144
4FGL J0009.2+6847	others	0.129	0.3986	0.4724
4FGL J0039.1+6257	PSR	0.7186	0.0444	0.237

4.2 YNG/MSP?

4.2.1 Feature Importance

Feature no, name and corresponding importance is shown in Table below. The result is plotted as bar chart in figure (4)

Feature No	Feature name	Feature importance
0	GLON	0.10834
1	Pivot energy	0.02842
2	PL Flux Density	0.00000
3	unc PL Flux density	0.00000
4	PL INDEX	0.10950
5	Unc PL Index	0.02941
6	LP Flux density	0.00000
7	Unc LP Flux density	0.00000
8	LP index	0.12908
9	Unc LP index	0.03548
10	LP Beta	0.01694
11	LP sigCurve	0.02482
12	PLEC flux density	0.00000
13	unc PLEC flux density	0.00000
14	PLEC index	0.02822
15	unc PLEC index	0.04928
16	PLEC expfactor	0.02061
17	unc PLEC expfactor	0.03679
18	PLEC exp index	0.00257
19	PLEC sigcurve	0.02401
20	Npred	0.07550
21	variability index	0.02628
22	frac variability	0.01004
23	unc frac variability	0.00762
24	varibility2	0.01943
25	frac variability2	0.01425
26	unc frac variability2	0.01116
27	h12	0.02927
28	h23	0.02760
29	h34	0.04409
30	h45	0.03809
31	h56	0.03309
32	h67	0.02011

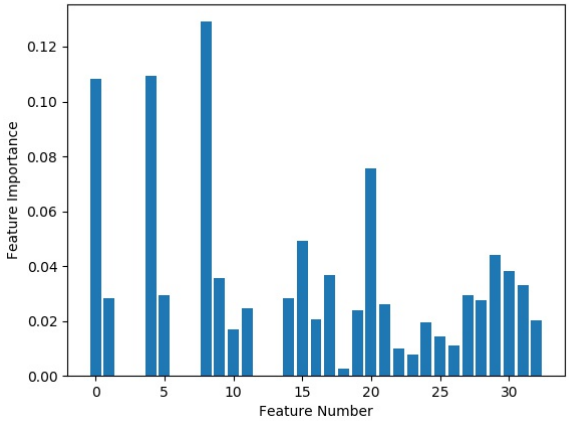

Figure 4: Feature Number vs Feature Importance Graph

4.2.2 Testing Random Classifier Model

Some of the results are shown below, the entire file can be found here.(partest.txt)

Actual class(y_t)	Predicted Class(y_p)	Prob_MSP	Prob_YNG
YNG	YNG	0.2522	0.7478
YNG	YNG	0.0864	0.9136
MSP	MSP	0.7758	0.2242
YNG	YNG	0.0864	0.9136
YNG	MSP	0.5416	0.4584

4.2.3 Predicting classes of Probable Pulsars

Some of the results are shown below. The entire file can be found here.(psrclass.csv)

Source Name	Predicted Class(y_p)	Prob_MSP	Prob_YNG
4FGL J0039.1+6257	MSP	0.61	0.39
4FGL J0212.1+5321	MSP	0.665	0.335
4FGL J0237.8+5238	MSP	0.7384	0.2616
4FGL J0340.4+5302	YNG	0.3356	0.6644
4FGL J0426.5+5434	YNG	0.419	0.581

4.3 Fermi-Chandra-Gaia Positional Crossmatch

Fermi-Chandra-Gaia full Cross-matched table for pulsars can be found here crossmatch.csv, along with comparison to gaia distances.
For two previously unknown sources, one is predicted as millisecond pulsar and the other as young pulsar by our code, we have found Chandra and Gaia

counterparts. Therefore in addition to associate classes to these two sources, we have obtained their distance(shown below:)

S_name	RA	DEC	Pred_class	prob_psr	prob_msp	prob_yng	Dist(kpc)
4FGL J0212.1+5321	33.0445	53.3507	MSP	0.969	0.665	0.335	1.156
4FGL J1748.3-2906	267.0955	-29.1061	YNG	0.4698	0.3454	0.6546	0.277

5 Conclusion

1. Using Random Forest Classifier method, we have classified fermi sources into three categories: AGN, Pulsar and others, with 95 percent accuracy.
2. Taking help from fermi pulsar catalog, we have classified all predicted pulsars into young/Milli Second pulsar classes with 87 percent accuracy.
3. We have correlated fermi-chandra-gaia sources within our galaxy using positional cross match. We have successfully obtained distances to some pulsars, and cross checked with already existing literature.
4. More importantly, using this cross match, we have successfully obtained distances to couple of Milli second pulsar candidtaes,4FGL J0212.1+5321 and 4FGL J1748.3-2906. The first one is a potential pulsar candidate with 96.9% probability, and a milli second pulsar with 66.5% probability. The second one is a pulsar with 46.9% probability and a young pulsar with 65.4% probability. For both the source we have found only one chandra and gaia sources. The distances to these sources are 1.156 kpc and 277 persec.

6 Appendix

6.1 AGN,Pulsar or others?

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.datasets import make_classification
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split
import pandas as pd
import scipy.stats as st

np.random.seed(4)

d=pd.read_csv("trainingf.csv")
d=d.dropna(how='any',axis=0)
X=d.drop(columns=['source name','Spectrum type','RA','DEC',
```

```
'GLON','GLAT','class'])
Y=d['class']
a,b=X.shape

#Testing correlations between parameters.
print("ttest")
l=[]
for i in range(0,b):
        for j in range(0,b):
                if(i>j):
                        t=st.ttest_rel(X.iloc[:,i],X.iloc[:,j])
                        if(t.pvalue>0.7):
                                print(i,j,t.pvalue)
                                l.append(j)
l=sorted(list(set(l)))
N=len(l)
print(l)
for i in range(N):
        X=X.drop(X.columns[l[i]],axis=1)
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.3)

clf=RandomForestClassifier(n_estimators=5000)
clf.fit(X_train,Y_train)
y_pred=clf.predict(X_test)
pred=clf.predict_proba(X_test)
imp=clf.feature_importances_

for i,v in enumerate(imp):
        print("Feature: %0d, Score: %0.5f" % (i,v))
plt.bar([x for x in range(len(imp))],imp)


from sklearn import metrics
print("Accuracy:",metrics.accuracy_score(Y_test,y_pred))
Y_test=Y_test.rename_axis('ID').values

with open("testresult.txt",'w') as zi:
                print("y_test y_pred psr agn others (assc prob)",file=zi)
                for j in range(len(y_pred)):
                        print(Y_test[j],y_pred[j],pred[j][0],
                        pred[j][1],pred[j][2],file=zi)
zi.close()

d1=pd.read_csv("unk2.csv")
d1=d1.dropna(how='any',axis=0)
```

```
d2=d1.drop(columns=['source name','RA','DEC',
'GLON','GLAT','Spectrum type','class'])

for i in range(len(l)):
        d2=d2.drop(d2.columns[l[i]],axis=1)
print(d2.shape)


y1_pred=clf.predict(d2)
pred1=clf.predict_proba(d2)
d1['predicted_class']=y1_pred
d1['prob_psr']=pred1[:,0]
d1['prob_agn']=pred1[:,1]
d1['prob_others']=pred1[:,2]

d1.to_csv("newcat.csv")

plt.xlabel("Feature Number")
plt.ylabel("Feature Importance")
plt.savefig("Feature.jpg")
plt.show()
```

6.2 Classification of Pulsars; Young or Millisecond Pulsar?

```
#Classification of pulsar subclasses.

from sklearn.ensemble import RandomForestClassifier
from sklearn.datasets import make_classification
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split
import pandas as pd
import scipy.stats as st

np.random.seed(0)

df=pd.read_csv("m_psr.csv")
df=df.dropna(how='any',axis=0)
d=df.drop(columns=['angDist','ra','dec','source name','RA','DEC','GLAT',
'Spectrum type','class'])
X=d.drop(columns=['type'])
Y=d['type']

#correlation ttest between parameters:
```

```
print("ttest")
a,b=X.shape

#Testing correlations between parameters.
l=[]
for i in range(0,b):
        for j in range(0,b):
                if(i>j):
                        t=st.ttest_rel(X.iloc[:,i],X.iloc[:,j])
                        if(t.pvalue>0.7):
                                print(i,j,t.pvalue)
                                l.append(j)
l=sorted(list(set(l)))
N=len(l)
print(l)
for i in range(N):
        X=X.drop(X.columns[l[i]],axis=1)
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.3)

clf=RandomForestClassifier(n_estimators=5000)
clf.fit(X_train,Y_train)
y_pred=clf.predict(X_test)
pred=clf.predict_proba(X_test)
imp=clf.feature_importances_

for i,v in enumerate(imp):
        print("Feature: %0d, Score: %0.5f" % (i,v))
plt.bar([x for x in range(len(imp))],imp)


from sklearn import metrics
print("Accuracy:",metrics.accuracy_score(Y_test,y_pred))
Y_test=Y_test.rename_axis('ID').values

with open("psrtest.txt",'w') as zi:
                print("y_test y_pred msp yng (assc prob)",file=zi)
                for j in range(len(y_pred)):
                        print(Y_test[j],y_pred[j],pred[j][0],pred[j][1],file=zi)
zi.close()

d1=pd.read_csv("n_psr.csv")
d1=d1.dropna(how='any',axis=0)

d2=d1.drop(columns=['source name','RA','DEC','GLAT','Spectrum type',
'class','predicted_class','prob_psr','prob_agn','prob_others'])
```

```
for i in range(len(l)):
        d2=d2.drop(d2.columns[l[i]],axis=1)
print(d2.shape)


y1_pred=clf.predict(d2)
pred1=clf.predict_proba(d2)
d1['predicted_class']=y1_pred
d1['prob_MSP']=pred1[:,0]
d1['prob_YNG']=pred1[:,1]

d1.to_csv("psrclass.csv")

plt.xlabel("Feature Number")
plt.ylabel("Feature Importance")
plt.savefig("Feature1.jpg")
plt.show()
```

6.3 Positional Cross Matching

```
from astropy.coordinates import SkyCoord, Angle
from astroquery.esasky import ESASky
from astropy import units as u
from astroquery.vizier import Vizier
import numpy as np

f=np.loadtxt("fermi.txt", dtype=str, delimiter=',')
n=np.loadtxt("listi.txt", dtype=float)
with open("crosstable.txt",'a') as fi:
        #print("FrowNo\tfermiRa(deg)\tfermi dec(deg)\tfermi
        err(deg)\tfermi class
        \tchandra ra(deg)\tchandra dec(deg)\tchandra err(arcsec)
        \tgaia ra(deg)\tgaia dec(deg)\tgaia source dist(persec)",file=fi)
        for j in range(206,len(n)):
                k=int(n[j])
                a=float(f[k,1])
                b=float(f[k,2])
                c=SkyCoord(a,b,unit="deg")
                m=float(f[k,5])
                result = ESASky.query_region_catalogs(c,
                m * u.deg, "chandra-sc2")
                l=len(f[1,:])-10
                y=f[k,l]
                for v in range(0,len(result[0]['ra'])):
                        ra=result[0][v]['ra']
                        dec=result[0][v]['dec']
```

```
err=result[0][v]['err_ellipse_r0']
coord=SkyCoord(ra,dec,unit="deg")
viz=Vizier.query_region(coord,
radius=Angle(err,"arcsec"),
catalog='I/347/gaia2dis')
if viz is not None:
        if(len(viz)>0):
                for s in range(0,len(viz[0]['RA_ICRS'])):
                        gra=viz[0][s]['RA_ICRS']
                        gdec=viz[0][s]['DE_ICRS']
                        grest=viz[0][s]['rest']
                        print("{}\t{}\t{}\t{}
                        \t{}\t{}\t{}\t{}
                        \t{}\t{}\t{}".format(k,a,b,
                        m,y,ra,dec,err,gra,
                        gdec,grest),file=fi)
                        print(j)
```

```
fi.close()
```

7 Acknowledgement

I want to express my gratitude to my supervisor, Dr. Sudip Bhattacharyya for his guidence in this work. I want to thank my friends Subhajit, Tuhin and Akashdeep for their helpful insights. Finally, I want to thank my fiancee Sayanwita Biswas, and my parents for their support in the troubled times.

8 Bibliography

- (1) S.Abdollah, F.Ackero et.al. "Fermi Large Area Telescope Fourth Source Catalog", arXiv:1902.10045v5[astro-ph.HE],17 Jan, 2020
- (2) Ackerman M, "A STATISTICAL APPROACH TO RECOGNIZING SOURCE CLASSES FOR UNASSOCIATED SOURCES IN THE FIRST FERMI-LAT CATALOG", The Astrophysical Journal, 753:83 (22pp), 2012 July 1
- 3Abdo,"Fermi Large Area Telescope First Source Catalog", http://arxiv.org/abs/1002.2280v1
- 4 Saz.Parkinson(2016),"CLASSIFICATION AND RANKING OF FERMI LAT GAMMA-RAY SOURCES FROM THE 3FGL CATALOG USING MACHINE LEARNING TECHNIQUES"http://dx.doi.org/10.3847/0004-637X/820/1/8

- (5) Accero et.al."Fermi Large Area Telescope Third Source Catalog", http://arxiv.org/abs/1501.02003v3
- (6) LAT collaboration, "The Second *Fermi* Large Area Telescope Catalog of Gamma-ray Pulsars" https://arxiv.org/abs/1305.4385
- (7)Gilles Louppe,"Understanding Random Forest, from Theory to Practice" https://arxiv.org/pdf/1407.7502.pdf
- (8) Bailer Jones et.al. "ESTIMATING DISTANCES FROM PARALLAXES IV: DISTANCES TO 1.33 BILLION STARS IN *Gaia* DATA RELEASE 2" http://iopscience.iop.org/article/10.3847/1538-3881/aacb21