LOFAR SOURCES IDENTIFICATION WITH MACHINE LEARNING

Credit: Cyril Tasse and the LOFAR surveys team.

LARA ALEGRE (SHE) - PHD STUDENT
SUPERVISORS: PROF. PHILIP BEST & DR. JOSE SABATER

www.lofar-surveys.org
ELAIS-N1
20.0 μJy/bm

e.g.
LoTSS

LOFAR Ttwo-Metre Sky Survey

LoTSS-DR1
- HETDEX
- 424 deg² (2% LoTSS)
- 58 pointings
- Radio sources: 318542
- Optical counterparts: 71% of the radio sources (PanSTARSS, WISE), Williams et al., 2019

LoTSS-DR2
- 13h and 0h fields
- 5700 deg² (27% LoTSS)
- Radio sources: 4.3M
- in prep
- Status of DR2 (observations + LGZ)

RGZ status at 2020-12-24 12:58:52
LoTSS-DR1
CROSS-IDENTIFICATION

WILLIAMS ET AL., 2019 “FLOWCHART”
ONE BIG DECISION TREE

LIKELIHOOD RATIO TECHNIQUE
& VISUAL ANALYSIS

LOFAR GALAXY ZOO

PyBDSF gaussians
LOFAR radio (150 MHz)
FIRST radio (1.4-GHz)

Source name: ILTJ13142.18+503610.6 (RA 202.936 DEC 50.603)

- WISE (IR.W1 band)
- PANSTRARRS (optical, r band)

RADIO PYBDSF SOURCES
- MULTIPLE RADIO COMPONENTS
- EXTENDED EMISSION
- BLENDED
**LoTSS-DR I**

**Cross-identification**

### PYBDSF Sources

<table>
<thead>
<tr>
<th>Decision tree outcome</th>
<th>Suitable for LR</th>
<th>Need visual inspection</th>
<th>Artefacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>295364</td>
<td>294129</td>
<td>1096</td>
</tr>
<tr>
<td>LGZ</td>
<td>8432</td>
<td>3143</td>
<td>5051</td>
</tr>
<tr>
<td>Prefilter</td>
<td>21099</td>
<td>10079</td>
<td>9604</td>
</tr>
<tr>
<td>Artefacts</td>
<td>799</td>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td>Total</td>
<td>325694</td>
<td>307352</td>
<td>15751</td>
</tr>
</tbody>
</table>

- **VISUAL INSPECTED 29531**
- **NEEDED VISUAL INSPECTION 15751**
- KEEP VISUAL INSPECTION LOW
- KEEP NUMBER OF SOURCES WRONGLY ACCEPTED BY LR LOW
- USE THE CHARACTERISTICS OF THE SOURCES AS INPUT FEATURES

- **> 99%** CORRECT DECISION TREE OUTCOME
- **~ 60%** DECISION TREE OUTCOME
- **~ 45%**
BINARY CLASSIFIER

CLASS 1
LR
- Pybdsf sources that were not associated with other PyBDSF sources
- were not deblended
- sources for which LR gave correct optical ID (or correctly lack of ID)

CLASS 0
LGZ
- PyBDSF sources that were associated with other sources in LGZ
- deblended into more than one source
- LR obtained incorrect ID

• Number of sources in class 0: 15751
• Number of sources in Class 1: 307352
• Exclude the artefacts: 2591

CREATE A BALANCED DATASET

DOWNsampling THE MAJORITY CLASS
75% TRAIN 25% TEST
Machine learning features

Baseline (BL)
- Maj
- Min
- Total_Flux
- Peak_Flux
- log_n_gauss

Likelihood Ratio (LR)
- l_r
- l_r_dist

Gaussian properties (GAUS)
- gauss_maj
- gauss_min
- gauss_flux_ratio
- log_gauss_l_r_tlv
- gauss_l_r_dist
- log_highest_l_r_tlv

Nearest Neighbours (NN)
- NN_45
- NN_dist
- NN_flux_ratio
- log_NN_l_r_tlv
- NN_l_r_dist

Cyclic 10x10 SOM (SOM)
- 10x10_closest_prototype_x1
- 10x10_closest_prototype_x2
- 10x10_closest_prototype_y1
- 10x10_closest_prototype_y2
Method - Supervised ML
Ensembles of Decision Trees

- Minimization of total loss
- More weight to models with better performance

Majority of the votes
Random Forest

Hyperparameters | Search values | Best GBC
--- | --- | ---
learning_rate | 0.001, 0.01, 0.05, 0.1, 0.5, 1 | 0.01
n_estimators | 100, 250, 500, 1000 | 500
max_depth | range (1, 11, steps = 1) | 8
subsample | range (0.05, 1.01, steps = 0.05) | 0.15
min_samples_split | range (2, 21, steps = 1) | 12
min_samples_leaf | range (1, 21, steps = 1) | 5
max_features | range (0.05, 1.01, steps = 0.05) | 0.6
RESULTS

MODEL PERFORMANCE

<table>
<thead>
<tr>
<th>Table</th>
<th>test</th>
<th>train</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.9460</td>
<td>0.9590</td>
</tr>
<tr>
<td>F1-score 1</td>
<td>0.9452</td>
<td>0.9582</td>
</tr>
<tr>
<td>F1-score 0</td>
<td>0.9468</td>
<td>0.9597</td>
</tr>
</tbody>
</table>

- Train vs test performance: avoid overfitting
- F1-score: performance on the different classes
- 96.4% of the sources that need visual inspection are sent to LGZ (but a different component of the same source may be sent to visual inspection)

Apply corrections
RESULTS
Fails and Corrections

False Positives

Multi Component Source
ILTJ105709.24+484041.0

Blended Source
ILTJ145409.19+503619.4
RESULTS

THRESHOLD VALUE AND CONFUSION MATRIX

- Corrected and not corrected
- Threshold value of 18%

Visual inspections - 10%
False positives - 30%

Confusion matrix, tlv=0.18

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGZ</td>
<td>801</td>
<td>14950</td>
</tr>
<tr>
<td>(1096)</td>
<td>(14655)</td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>295537</td>
<td>11815</td>
</tr>
<tr>
<td>(294129)</td>
<td>(13230)</td>
<td></td>
</tr>
</tbody>
</table>
**SUMMARY & WORK IN PROGRESS**

**ML**
- This project
  - Perform the selection of the sources for which the statistical matching is not reliable
  - Reduces the nr of sources visually inspected
  - Improves the flowchart process

**DL**
- In progress
  - Select the multiple components and blended sources
  - Allows to identify single components were the LR fails

**LOFAR TEAM**
- Cross-ids for multiple component sources
- Automate the cross-matching for the single component extended sources