Photometric classification of QSOs from RCS2 using Random Forest

Felipe Barrientos (PUC)

Daniela Carrasco (PUC, U Merlbourne)
Karim Pichara (PUC)

Timo Anguita (UNAB)
Howard Yee (U Toronto)
Mike Gladders (U Chicago)
David Gilbank (SAAO)
Clusters of galaxies are the most massive objects and are sensitive to the cosmology we live in.

We have carried on two large imaging surveys:

- The Red-sequence Cluster Survey 1 (RCS1)
  100 sq deg
  2 bands, R and z
  $1.4 \times 10^7$ objects
  $\sim 10^3$ clusters

- The Red-sequence Cluster Survey 2 (RCS2)
  1000 sq deg
  3+1 bands, grz+i
  $1.1 \times 10^8$ objects
  $N \times 10^4$ clusters
RCS 1 and 2 footprints
Homogeneity of the colors of the early-type (E/S0) population seen at low and intermediate redshifts, and present in all the clusters found to date.

A powerful tool to find clusters of galaxies.
Very good photo-z
Samples of AGNs have a wide range of applications

Galaxy evolution
Intergalactic medium
Local physics around black holes
Large scale structure
Astrometric references for local studies
+ ...

One particular application: multiple lensed quasars

Walsh et al (1979)
Multiple lensed quasars

Many applications:
- Mass determination of the lens
- Environmental effects
- Determination of the Hubble Constant
- The interstellar medium in lens galaxies
- Cosmography with lens statistics (pairs)
- Natural telescopes

Our goal is to find lensed quasars from our quasar candidates and quasars behind galaxy clusters
QbC: The Quasars behind Clusters Survey

QSO

MgII 2803

MgII 2796

$\times$

$d$

☺
Selection of candidates: Random Forest

is a tree-based classification method (Breiman 2001)

Random Forest uses a training set (objects for which the class is known), and the resulting mapping is applied to objects whose class is not available

The process of building a Random Forest for some number of T trees given training data is:
Random Forest

Let \( N \) be the number of objects in the training set. The classifier samples \( N \) objects at random to produce \( S \). This sample \( S \) will be the training set for the growing tree.

- \( M \) is the number of attributes (i.e. colors & mags) for each object, and \( k \) (\(< M\)) variables are selected randomly during the forest growing.

- The predictor variable that provides the best split is used to split the node.

- Each tree is grown to the largest extent possible.
- When a new input is entered into the system, it is run down all of the trees. The result is a voting majority.

- 10-fold cross-validation over the training set to estimate the performance.

- Quantify performance:

<table>
<thead>
<tr>
<th>Known Label</th>
<th>Predicted Label</th>
<th>Positives</th>
<th>Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>True positive</td>
<td>False Negative</td>
</tr>
</tbody>
</table>
| Negative    | False Positive | True Negative | }
## Predicted Label

<table>
<thead>
<tr>
<th>Known Label</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True positive</td>
<td>False Negative</td>
</tr>
<tr>
<td>Negative</td>
<td>False Positive</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

- **Precision:** The percentage of positive predictions that are correct

\[
\frac{TP}{TP + FP}
\]

- **Recall:** The percentage of positive labeled instances that were predicted as positive.

\[
\frac{TP}{TP + FN}
\]
Training Set 1

- Spectroscopically confirmed quasars and stars from SDSS cross-identified in RCS-2.
- 4,916 quasars and 10,595 stars
- Matched with point sources in RCS2 \((r < 0.5'')\)
- Features: \(g, r, i, z, (g-r), (g-i), (g-z), (r-i), (r-z), (i-z)\)

![Graphs showing distributions of stars and QSOs with features](image-url)
Test Set 1 (TS1)

- A total of 1,863,970 objects
- Using 3 features and 70 trees
- 85,085 candidates (<5%)
- Consistent with predictions for LSST

<table>
<thead>
<tr>
<th>TS1</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>QSO</td>
<td>0.875</td>
<td>0.900</td>
</tr>
<tr>
<td>STAR</td>
<td>0.967</td>
<td>0.958</td>
</tr>
</tbody>
</table>
To improve classification add more data: GALEX (Training Set 2)

- NUV (λ 2267 A) photometry
- 1,228 quasars and 815 stars
- Matched with point sources in RCS2 (r < 2.0")
- Features: NUV, g, r, i, z, (g-r), (g-i), (g-z), (r-i), (r-z), (i-z+)
  NUV-g, NUV-r, NUV-i, NUV-z
- About 1.5 mag brighter than Training Set 1

**Images:**
- NUV-g vs g-r
- g-r vs i-z
- r-i vs i-z
Test Set 2 (TS2)

- A total of 16,898 objects
- Using 9 features and 100 trees
- 6556 candidates (38.8%)

<table>
<thead>
<tr>
<th>TS2</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>QSO</td>
<td>0.969</td>
<td>0.976</td>
</tr>
<tr>
<td>STAR</td>
<td>0.963</td>
<td>0.953</td>
</tr>
</tbody>
</table>
... And more data: WISE (Training Set 3)

- W1 (3.4 μm) and W2 (4.6 μm) photometry
- 2,748 quasars and 2,749 stars
- Matched with point sources in RCS2 (r < 2.0")
- Features: W1, W2, g, r, i, z, (g-r), (g-i), (g-z), (r-i), (r-z), (i-z)+
  g-W1, g-W2, r-W1, r-W2, i-W1, i-W2, z-W1, z-W2, W1-W2

Stars

QSOs
Test Set 3 (TS3)

- A total of 242,902 objects
- Using 7 features and 60 trees
- 21,713 candidates (8.94%)

<table>
<thead>
<tr>
<th>TS3</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>QSO</td>
<td>0.993</td>
<td>0.992</td>
</tr>
<tr>
<td>STAR</td>
<td>0.992</td>
<td>0.993</td>
</tr>
</tbody>
</table>
Summary

From a total set of 1,863,970 objects to classify:
- 85,085 quasar candidates from TeS1
- 6,556 quasar candidates from TeS2
- 21,713 quasar candidates from TeS3

91,842 new quasar candidates from RCS-2

3,600 objects classified as quasars from the three test sets
13,691 objects classified as quasars from the two test sets
74,551 objects classified as quasars from only one test set
Spectroscopic confirmation

du Pont Telescope + Boller & Chivens spectrograph
Mag lim r ~ 19
15-16 out of 17 are confirmed QSOs

The selection does work!!

Results in Carrasco et al (submitted)
This process can be applied to other similar datasets

(Galaxy morphology)

Thanks
Morphology using a 2-D galaxy light profile fitting algorithm (L-M chi sq minimization – 5 parameters – own GALFIT)
This accounts for seeing, providing a galaxy size (i.e
• \( r_e \) for an \( r^{1/4} \)-law or \( h \) for an exponential disk)
• + VISUAL CLASSIFICATION
This approach is not longer possible with large datasets

LM minimization is too expensive ( \(~1\ \text{min} / \text{galaxy}\) )

New algorithm

- Populate the parameter space with a discrete number of models
- move convolution out of the fitting process
Model cube
- position angle
- size
- axial ratio

Convolved with the PSF of each CCD

~ 2000 models
Code implemented in Python

First large test.

5 million objects fitted
r-band only
Using 200 cores
In 36 hours

~ 4000 galaxies visually inspected