Use of neural networks for the identification of new $z \geq 3.6$ QSOs from FIRST-SDSS DR5

R. Carballo, J.I. González-Serrano, C.R. Benn, F. Jiménez-Luján

Abstract We aim to obtain a complete sample of $z \geq 3.6$ radio QSOs from FIRST sources having star-like counterparts in the SDSS DR5 photometric survey ($r_{\rm AB} \leq 20.2$). The starting sample of FIRST-DR5 pairs includes 4250 objects with DR5 spectra, 52 of these being $z \geq 3.6$ QSOs. Simple supervised neural networks, trained on these sources, using optical photometry and radio data, are very effective for identifying high-z QSOs, yielding 96 per cent completeness and 62 per cent efficiency. Applying these networks to the 4415 FIRST-DR5 sources *without* DR5 spectra we found 58 $z \geq 3.6$ QSO candidates. We obtained spectra of 27 of them, confirming 17 as high-z QSOs. Spectra of 13 additional candidates from the literature and SDSS DR6 revealed seven more $z \geq 3.6$ QSOs, giving an overall efficiency of 60 per cent (24/40). None of the non-candidates with spectra from NED or DR6 is a $z \geq 3.6$ QSO, consistently with a high completeness. The initial sample of high-z QSOs is increased from 52 to 76 sources (a factor 1.46). From the new identifications and candidates we estimate an incompleteness of SDSS for the spectroscopic classification of FIRST $3.6 \leq z \leq 4.6$ QSOs of 15 per cent for $r \leq 20.2$.

R. Carballo
Dpto. de Matemática Aplicada y Ciencias de la Computación, Univ. de Cantabria. ETSI Caminos, Canales y Puertos, Avda de los Castros s/n, E-39005 Santander, Spain, e-mail: carballor@unican.es

J.I. González-Serrano
Instituto de Física de Cantabria (CSIC-Universidad de Cantabria), Avda de los Castros s/n, E-39005 Santander, Spain, e-mail: gserrano@ifca.unican.es

C.R. Benn
Isaac Newton Group, Apartado 321, E-38700 Santa Cruz de La Palma, Spain, e-mail: crb@ing.iac.es

F. Jiménez-Luján
Instituto de Física de Cantabria (CSIC-UC) and Dpto. Física Moderna, Avda de los Castros s/n, E-39005 Santander, Spain, e-mail: jimenezf@ifca.unican.es

1 Introduction

Homogeneous statistical samples of high-redshift quasi-stellar objects (QSOs) allow not only investigation of the QSO phenomenon itself, but also provide important information for a wide variety of studies. In particular, the luminosity function of high-z QSOs provides strong constraints on the theory of the accretion of matter onto supermassive black holes in the nuclei of galaxies. The increasing evidence for a relation between the formation of galaxy bulges and supermassive black holes [1] [2] emphasises the importance of understanding the role of QSO activity in the formation and evolution of galaxies.

Although radio-loud (RL) QSOs are a small subset of the QSO population, samples of high-z RL QSOs benefit from higher completeness, due to the drastically reduced contamination by stars in samples of radio selected QSO candidates, compared to optically (colour) selected QSO candidates [3]. Moreover, the connection between radio and optical activity, which still needs to be understood, requires a comparison between radio-loud and radio-quiet QSO populations. Many studies suggest that RL QSOs reside in more massive galaxies and harbour more massive central black holes than radio-quiet QSOs, but the point is still controversial (references for and against at [4]).

We aim to obtain a homogeneous sample of high-z RL QSOs drawn from correlation of the Faint Images of the Radio Sky at Twenty cm Survey (FIRST, [5]) with unresolved objects in the Sloan Digital Sky Survey (SDSS, [6]) Data Release 5 (DR5, [7]). The area of overlap between FIRST and the DR5 imaging survey is ~ 7391 deg^2 and the number of selected FIRST-SDSS matches is 8665. SDSS provides: (i) *ugriz* photometry, which is a powerful tool for separating high-z QSOs from stars, QSOs with z below ~ 3.6 or unresolved low-z galaxies; (ii) morphological classification, essential for distinguishing between high-z QSOs and galaxies or resolved low-z active galactic nuclei (hereafter AGN); and (iii) spectroscopy of many of our candidates (4250), selected as spectroscopic targets by SDSS DR5. Since SDSS spectroscopy lags the imaging, the DR5 spectroscopic area is lower, with ~ 5553 deg^2 in the overlap with FIRST. Most of the candidates with available spectrum were classified by SDSS as QSOs, i.e. a high-excitation emission line with FWHM ≥ 1000 km sec^{-1} was detected. The rest are galaxies, stars and objects of 'unknown' class. 52 DR5 sources were classified as $z \geq 3.6$ QSOs.

Our approach to obtain a high-z QSO sample was to extend the set of 52 high-z QSOs by applying automated learning techniques, specifically neural networks (NNs), to the 8665 FIRST-SDSS DR5 photometric matches. NNs have been shown to be powerful tools for classification and regression in many fields of astronomy, allowing to predict object classes and/or estimate astrophysical parameters.

QSO selection and estimation of QSO photometric redshifts are of prime importance for the SDSS project. Various studies addressed the problem using different machine learning approaches. [8] applied a probability density analysis based on kernel density estimation of the colour distribution of stars and spectroscopically confirmed QSOs in SDSS DR1, to classify, as stars or QSOs, over 10^5 unresolved UV-excess QSO candidates (z up to $2.4-3.0$) with 95 per cent efficiency and com-

pleteness. Moreover they provide photometric redshifts, based on the procedure described in [9], with deviations $|\Delta z| < 0.3$ and $|\Delta z| < 0.1$ for 86 per cent and 65 per cent of the QSOs with available spectroscopic redshift, respectively.

[10] applied the oblique decision tree classifier ClassX to separate SDSS-DR2 photometric objects into 25 classes (stars, red stars, 10 redshift bins for galaxies, and 13 for AGN) using colour information and morphology. For each of the 12 redshift bins for AGN with $\Delta z = 0.2$ and covering $0 \leq z \leq 2.4$, the completeness obtained lies in the range $43-81$ per cent, with average 63 per cent, and the efficiency in the range $48-77$ per cent, with average 62 per cent. For the high-redshift bin, with $z = 2.4-6.0$, the efficiency is good, ~ 75 per cent, although the completeness drops to 14 per cent.

[11] applied decision trees, trained on the SDSS-DR3 objects with available spectroscopy, to classify all photometric objects ($> 10^8$) in SDSS-DR3 in one of the three categories of star, galaxy or nsng (neither star nor galaxy), the latter including QSOs and 'unknown'. A blind test on the 2dF QSO Redshift Survey (2QZ, [12]), using the 8739 2QZ-SDSS matches, yielded 95 per cent completeness and 87 per cent efficiency. The authors did not discuss how the performance depends on redshift. [13] applied the nearest neighbour method to calculate photometric redshifts of DR5 QSOs reaching an encouraging result, with an average dispersion between photometric and spectroscopic redshift of $\sigma = 0.343 \pm 0.005$ for 11000 QSOs with redshifts up to $z = 4$.

2 Method, discussion and conclusions

2.1 Selection of the sample

The initial sample of 8665 sources is made of *all* FIRST sources with $S_{1.4\ \mathrm{GHz}} > 1.0$ mJy and with an unresolved counterpart at SDSS-DR5 within 1.5 arcsec of the radio position, with clean photometry and PSF magnitude $15 \leq r_{\mathrm{AB}} < 20.2$. 4250 of the sources (49 per cent) have DR5 spectra, belonging to the classes of QSOs (3808), stars (230), galaxies (59) and 'unknown' (153). 52 of the QSOs have $z \geq 3.6$.

Since the sample was obtained using a one-to-one match between radio and optical positions, within a 1.5 arcsec radius, the class of double-lobe QSOs without detected radio cores is missed. The incompleteness due to this effect is expected to be around 3.7 per cent [14].

2.2 Separability of high-z QSOs in the labelled sample

Our goal is to train a classifier to recognize high-z QSOs among the 4415 objects *without* SDSS spectra, i.e. the 'unlabelled' sources, after learning the class prop-

erties from the 4250 objects *with* spectra, i.e. the 'labelled' sources. The adopted procedure was to consider a two-class problem, with high-z QSOs as one class and the remaining types as the other.

We trained a feed-forward NN using the logistic linear discriminant, with a layer for the input parameters, i.e. the data, and an output layer with a single variable y, set during training to 1 for high-z QSOs and 0 for the remaining classes. The adopted error function was the mean of the squared errors of the outputs and optimization was obtained using the Levenberg-Marquardt algorithm. The output values, $0 \leq y \leq 1$, give the degree of similarity with the class of high-z QSOs. Objects with y exceeding a given threshold y_c are classified as high-z QSO candidates.

The classifier has to be tested using objects not used for the learning. We separated training and test objects adopting the partition 'leave-one-out', repeatedly dividing the data set of m instances into a training set of size $m-1$ and a test set of size 1, in all possible ways. This procedure yielded m classifiers. Since the m objects (4250) are used for testing, a good estimate of the performance can be obtained.

As input data we tried various combinations of SDSS photometry (r magnitude and $u-g$, $g-r$, $r-i$ and $i-z$ colours), optical-radio separation and total and peak radio flux, finding for the set of the first six parameters the best results, yielding 96 ± 14 per cent completeness (50/52) and 62 ± 9 per cent efficiency (50/81) for the threshold $y_c = 0.1$. The 31 contaminants include 28 QSOs, 19 of them with $3.2 \leq z < 3.6$, very close to the redshift cut.

12 of the 52 high-z QSOs show broad absorption lines (BALs) or self-absorption at Lyα, and the classifier recovers 11 of them, therefore being effective in the selection of QSOs with this type of absorption features.

2.3 Identification of new high-z QSOs in the unlabelled sample

The unlabelled sample includes the sources in the DR5 spectroscopic area not selected as spectroscopic targets by SDSS (2059 objects, compared to 4250 labelled in the same region), and the sources in the DR5 photometric area but outside the DR5 spectroscopic area (2356 objects). We expect a reasonable overlap between labelled sources and the sources located outside the spectroscopic area, since a large fraction of the latter would have been SDSS spectroscopic targets if included in the spectroscopic area. A poorer overlap is expected for the remaining sources, and in this sense the NNs would be performing an extrapolation. The input variable showing the largest difference between labelled and unlabelled sources is the r-band magnitude, about 0.5 magnitudes fainter for the latter.

The application of the NNs to the 4415 unlabelled sources yielded 58 high-z QSO candidates. The quality of the selection of high-z QSOs was tested with available spectroscopy from the NASA Extragalactic Database (NED), from a dedicated observing programme for this work, and from spectroscopic classifications from SDSS DR6.

Regarding the NED, four of the candidates were spectroscopically classified, all being QSOs with $z > 3.3$ (three with redshifts 4.33, 4.17 and 3.694 [15] and one with $z = 3.305$[16]), and no high-z QSO was found among the non-candidates.

Spectra of other 27 candidates were obtained with the ISIS spectropgraph at the WHT (La Palma) on 2007 April and July. 26 were classified as high-z QSOs (17), QSOs with $3.17 \leq z < 3.6$ (7), QSOs with $1.07 \leq z < 1.34$ (4); the other one remaining unclassified.

SDSS DR6 provided the spectra of nine additional candidates, including four QSOs with $3.6 \leq z \leq 3.8$, one at $z = 3.40$, two galaxies with $z = 0.45$ and $z = 0.58$, and two sources labelled as unknown. Amongst the 4357 non-candidates, 898 have spectra from DR6, and two are classified as high-z QSOs although with low confidence (SDSS 075559+113211, SDSS 161836+153313). We found that these SDSS redshifts are incorrect, due the interpretation of the Mg II emission line as Lyα.

2.4 Performance of the selection of new high-z QSOs

From a total of 40 candidates with available spectra, 24 are high-z QSOs, yielding an overall efficiency in the selection of new high-z QSOs of 60 ± 12 per cent. This value shows a good agreement with the efficiency obtained for the training sample, 62 ± 9. We note that the efficiency increases with *NN* output, going from 22 per cent for outputs $0.1 \leq y < 0.55$ to 91 per cent for the range $0.55 \leq y \leq 1$. The remaining 16 contaminants include seven QSOs with $3.15 \leq z \leq 3.6$, confirming the high rate of QSOs with z near the threshold $z = 3.6$ previously found in the training sample.

None of the non-candidates with spectra available from NED or DR6 is a $z \geq 3.6$ QSO, therefore we have no evidence of incompleteness regarding high-z QSOs with matches in these catalogues. Since NED spectroscopic identifications are assigned from a variety of surveys different than SDSS, the database provides a blind test of the good completeness of the classifier for DR5 unlabelled sources.

The number of RL high-z QSOs was increased from 52 in the initial sample to 76, i.e. a factor 1.46.

Adopting for the 18 candidates which still lack spectroscopy a weighted efficiency of 37 per cent (four candidates with $y \geq 0.55$ and 14 with $y < 0.55$), we calculate ~ 7 additional $z \geq 3.6$ QSOs. The FIRST-DR5 sample of high-z QSOs is thus expected to contain ~ 83 QSOs (52+24+7). Adopting as a lower limit for completeness the nominal value of 96 per cent found for the labelled sample, we calculate for the set of 31 high-z QSOs obtained by the classifier (24 discovered and 7 predicted) a minimum ~ 1 missed high-z QSO.

2.5 Completeness of SDSS for the selection of FIRST high-z QSOs

Our results allow us to obtain an estimate of the incompleteness of SDSS for the spectroscopic classification of FIRST high-z QSOs. 47 of the high-z QSO candidates are located in the spectroscopic area covered by DR6, and 17 of them are $z \geq 3.6$ QSOs, ten included in the DR6 spectroscopic catalogue and seven not included. 15 candidates in this area still lack spectroscopy, and assuming for them a weighted efficiency of 31 per cent (two candidates with $y \geq 0.55$ and 13 with $y < 0.55$), we expect another four high-z QSOs. From this calculation we estimate 11 FIRST high-z QSOs missed by SDSS (7 QSOs and 4 candidates), which when compared to 62 (52+10) identifications yields an incompleteness of SDSS for the spectroscopic classification of FIRST $3.6 \leq z \leq 4.6$ QSOs of ~ 15 per cent (11/73) for $r \leq 20.2$.

2.6 Future work

We plan to analyze the optical luminosity function of FIRST-SDSS QSOs at $3.6 \leq z \leq 4.6$ on the basis of this sample. Concurrently we expect to carry out spectroscopic observations of the 18 candidates without spectra. Given the efficacy of our approach, we intend to extend the sample using updated SDSS data releases, increasing the number of training sources and the number of high-z QSO candidates, for which subsequent spectroscopy will be planned. We envisage using additional infrared data via UKIDSS (UKIRT Infrared Deep Sky Survey).

References

1. Kormendy J., Richstone D., 1995, ARA&A, 33, 581
2. Magorrian J. et al., 1998, AJ, 115, 2285
3. Richards G.T. et al., 2006, AJ, 131, 2766
4. Cirasuolo M., Magliocchetti M., Gentile G., Celotti A., Cristiani S., Danese L., 2006, MNRAS, 371, 695
5. Becker R.H., White R.L., Helfand D.J., 1995, ApJ, 450, 559
6. York D.G. et al., 2000, AJ, 120, 1579
7. Adelman-McCarthy J. et al., 2007, ApJS 172, 634
8. Richards G.T. et al., 2004, ApJSS, 155, 257
9. Weinstein M.A. et al., 2004, ApJSS, 155, 243
10. Suchkov A. A., Hanisch R.J., Margon B., 2005, AJ 130, 2439
11. Ball N.M., Brunner R.J., Myers A.D., Tcheng D., 2006, ApJ, 650, 497
12. Crom S.M., Smith R.J., Boyle B.J., Shanks T., Miller L., Outram P.J. Loaring N.S., 2004, MNRAS, 349, 1397
13. Ball N.M., Brunner R.J., Myers A.D., Strand N.E., Alberts S.L., Tcheng D., 2008, ApJ, 683, 12
14. de Vries W.H., Becker R.H., White R.L., 2006, AJ 131, 666
15. Benn C.R., Vigotti M., Pedani M., Holt J., Mack K.-H., Curran R., Sánchez S.F., 2002, MNRAS, 329, 221
16. Mason K.O. et al., 2000, MNRAS, 311, 456