Quasar Selection from Combined Radio and Optical Surveys using Neural Networks
Ruth Carballo and Antonio Santiago Cofiño
Dpto. de Matemática Aplicada y C. Computación. Universidad de Cantabria, Avda de los Castros s/n, 39005 Santander, Spain
and
José Ignacio González-Serrano
Instituto de Física de Cantabria (CSIC-Univ. Cantabria), Avda de los Castros s/n, 39005 Santander, Spain

Abstract
The application of supervised artificial neural networks (ANNs) for quasar selection is investigated, using the list of candidates and their classification from White et al. (2000). The adopted architectures are 7:1 and 7:2:1, both with seven input parameters - optical and radio data from APM POSS-I (E, O plates) and VLA/FIRST, and a single output interpreted as a quasar probability. Both models were trained on samples of ~ 800 sources and yielded similar performance on independent test samples, with reliability as large as 90 to 80% for completeness from 70 to 90%. For comparison, the quasar fraction from the original list of candidates was 56%. The accuracy found with ANNs is similar to that obtained by White et al. using oblique decision trees and training samples of similar size. In view of the large degree of overlapping between quasars and nonquasars in parameter space, this performance is probably the best that can be achieved with this database. Predictions of the probabilities for the 98 candidates without spectroscopic classification in White et al. are presented, showing a good agreement between the two ANN models and with the values obtained by White et al. Eight of these sources have recent spectroscopic classification from the NASA Extragalactic Database or from the Sloan Digital Sky Survey Data Release 2 and the classes are consistent with their probabilities, reinforcing the ability of ANNs to optimize the selection of quasars. This work presents the first analysis of the performance of ANNs for quasar selection and it shows that ANNs provide a promising technique to single out specific object types in astronomical databases.

An article with the full description of this work has been accepted for publication in Monthly Notices of the Royal Astronomical Society ("Selection of quasar candidates from combined radio and optical survey", Carballo, Cofiño and González-Serrano. © 2004. The Royal Astronomical Society).

1. INTRODUCTION
The full exploitation of the large astronomical databases now available will be only possible with the help of artificial intelligence tools. ANNs have been applied in astronomy mainly for classification of stellar spectra, morphological star/galaxy separation, morphological and spectral galaxy classification, and photometric redshifts of galaxies. A summary of these and other applications can be found in Tagliaferri et al. (2003).

White et al. (2000) present a well-defined list of quasar candidates drawn from the correlation of the VLA/FIRST radio survey with blue starlike sources on APM POSS-I (E and O plates), and the spectroscopic classification of 1130 of the candidates, 636 (56%) being confirmed as quasars. These quasars form the FIRST Bright Quasar Survey of the North Galactic Cap (FBQS-2). Using the sample of candidates with available spectroscopy, the authors trained the oblique decision tree classifier OC1 (Murthy, Kasif and Salzberg 1994), taking as input parameters APM and FIRST data and as the output a value 1 for quasars and 0 for nonquasars, so that the actual output could be interpreted as a quasar probability \( p(Q) \). The performance of any classifier can
be quantified through the efficiency and the completeness of the subsamples selected above a probability threshold $p_c(Q)$. For this case, the efficiency (or reliability) is the fraction of quasars among the candidates with $p(Q) > p_c(Q)$, and the completeness is the fraction of quasars with $p(Q) > p_c(Q)$. White et al. confirmed on test sets that the decision tree classifier OC1 showed a very good performance, allowing to obtain samples with reliability as high as 80% at 90% completeness.

In this work we investigate the performance of ANNs for the selection of quasars using the candidate list in White et al. Our sample includes 1112 of the original 1130 sources, since we rejected those undetected in APM and for which White et al. use APS magnitudes.

2. FITTING AND TESTING TECHNIQUE

The type of ANN we used is the multi-layer perceptron (Bishop 1995), with architectures $7:1$ and $7:2:1$. We assumed that every node is connected to every node in the previous layer and every node in the next layer only. The seven input parameters, similar to those used by White et al., were $E$, $O - E$, $\log_{10} S_p$ (where $S_p$ is the FIRST peak flux density), $S_i/S_p$ (where $S_i$ is the FIRST integrated flux density), the radio-optical separation, and the point spread functions PSF($E$) and PSF($O$). We applied the Levenberg-Marquardt optimization algorithm to minimize the mean of the squared errors $m.se$, the error for each object being the difference between the output (probability of being a quasar) and its target value.

In order to reduce overfitting (i.e. memorization of the outputs rather than modelling) we used training with validation error: the training that is being carried out in the training set is automatically stopped when the error obtained running the trained network in another set, the validation set, does not decrease for a given number of iterations. An additional independent set, the test set, is used to evaluate the ANN performance.

The sample of classified candidates was divided in four sets, each of them with similar fractions of the different object types as the total sample. Setting aside each set, the remaining three were used for the training and validation, and the set itself was used for the test. Repeating the procedure for each of the four sets, we obtained four different classifiers, with the advantage of having used all the objects for the training/validation and all the objects for the test, optimizing the statistics. The size of the test sets, of about 275 objects, insure the inclusion of about a dozen of objects of the classes with fewer members, such as passive galaxies or BL Lac.

The ANN was run $10 \times 10$ times per set, the first factor accounting for different random numbers (for instance for the initial weights) and the second for the use of different splittings to separate the training and validation sets. In order to choose the best ANN we first selected the splitting with better average of $m.se$ for training and validation, in the sense that $m.se$ was both small and in agreement for the training and the validation sets. Then the best ANN of the splitting (with the same criterion) was selected. In the end we had a final ANN for each of the four test sets. Running each ANN for its corresponding test set we obtained $p(Q)$ for the 1112 candidates.

3. PERFORMANCE OF THE NETWORKS

Fig. 1 shows the distribution of $p(Q)$ for the $7:1$ ANN. The majority of the high-$p(Q)$ nonquasars are BL Lac objects. The reliability of quasar selection for $p(Q) > 0.75$ is 91% (353/386) and increases to 98% considering quasar or BL Lac selection (377/386). The corresponding completeness would be 56% (353/627) for quasars and 54% (377/694) for either quasars or BL Lac. The completeness decreases in the latter case since only blue BL Lac are confused with quasars. At the other extreme, there are 36 quasars with $p(Q) < 0.2$, and their most significant differences with respect to the remaining quasars are their redder $O - E$ colours and lower redshifts (twenty-five of them have $z < 0.25$), although they differ also in their wider PSF and larger integrated-to-peak radio flux ratio. The contribution of the host galaxy emission, less perceptible at high redshift, is the most likely explanation for the different input parameters of low-$z$ quasars.
FIGURE 1: Distribution of $p(Q)$ for the 7:1 ANN. The shaded distributions correspond to the objects of the indicated types.

FIGURE 2: Similar to Fig. 1, but for the 7:2:1 architecture.
The distribution of \( p(Q) \) for the 7:2:1 architecture (Fig. 2) is more peaked towards the extreme values, especially for high probabilities, and in this sense is more similar to the quasar probability distribution found by White et al. (2000) with OC1. The reliability of quasar selection for \( p(Q) > 0.85 \) is 89\% and the corresponding completeness 59\%. As occurred for the 7:1 ANN, most of the high-\( p(Q) \) nonquasars are blue BL Lac. Considering either quasars or BL Lac, reliability and completeness change to 96\% and 58\% respectively. Thirty-nine quasars have \( p(Q) < 0.2 \), and again the majority of these quasars have low redshift and redder \( O - E \) colours, wider PSF and larger integrated-to-peak radio flux ratio than the remaining quasars, most likely as a consequence of the contribution of the host galaxy emission.

Fig. 3 shows the efficiency and completeness as a function of the quasar probability threshold for the ANN models and OC1. The three distributions show equally good performances, with reliabilities ranging from 90 to 80\% for completeness from 70 to 90\% respectively. ANNs with more complex architectures were not explored, since the inclusion of the hidden layer - increasing the free parameters of the network from 8 to 19 - did not improve the performance. In view of the large degree of overlapping between quasars and nonquasars in parameter space, this is probably the best accuracy that ANNs or decision trees can achieve with the current database.

4. PROBABILITIES FOR THE UNCLASSIFIED CANDIDATES

The ANN models 7:1 and 7:2:1 were used to predict \( p(Q) \) for the 98 FBQS-2 candidates without spectral classification in White et al. We adopted four classifiers per model, corresponding to the four selected ANNs. Fig. 4a shows the probabilities obtained with the 7:1 architecture - plotted with a different line type for each ANN - and using OC1. There is a good agreement between the probabilities predicted with the four ANNs and between them and the values from OC1. Similar results are found for the 7:2:1 architecture (Fig. 4b). The probabilities obtained for the two ANN models (average of four ANNs per model) and OC1 are listed in Table 4 of Carballo, Cofiño and González-Serrano (2004). The average difference between \( p(Q) \) for the two ANN models (7:2:1 – 7:1) is of only 0.04, with standard deviation 0.06.

Eight of the sources unclassified in White et al. have now an spectroscopic classification either from the NASA Extragalactic Database or from the Sloan Digital Sky Survey Data Release 2. Five quasars (FBQS J083522.7+424258, FBQS J120354.7+371137, FBQS J125142.2+240435, FBQS J153402.2+425249 and FBQS J164733.9+364055) have probabilities ranging from 0.71 to 0.94 (mean 0.82), two Ultraluminous Infrared Galaxies (FBQS J091309.2+413635 and FBQS J125018.1+364914) have \( p \sim 0.1 \) and a star (FBQS J074342.2+321543) has \( p = 0.48 \), reinforcing the ability of ANNs to select quasars.

5. CONCLUSIONS

The performance of neural networks for the selection of quasar candidates from combined radio and optical surveys with photometric and morphological data is analysed. The work is based on the candidate list leading to FBQS-2 (White et al. 2000), and the input parameters used are radio
FIGURE 4: Comparison of the probabilities obtained with OC1 and the ANN models for the 98 candidates unclassified in White et al.
Quasar Selection from Combined Radio and Optical Surveys

flux, integrated-to-peak radio flux ratio, photometry and point spread function in a red and a blue band, and radio-optical position separation.

Two ANN architectures were investigated: a logistic model (7:1) and a model with a hidden layer with two nodes (7:2:1), and both yielded similarly good performances, allowing to obtain subsamples of quasar candidates from FBQS-2 with efficiencies as large as 87% at 80% completeness. For comparison the quasar fraction from the original candidate list was 56%. The efficiencies we find for completeness in the range 70 to 90% are 90–80%, similar to those found by White et al. using the oblique decision tree classifier OC1 and a similar sample size for the training. The lack of a clean separation between quasars and nonquasars in the parameter space certainly limits the accuracy of the classification, and the agreement in the performances obtained favours the interpretation that the three classifiers approach the maximum value achievable with this database. Although none of the two artificial intelligence tools provides a secure quasar classification (say efficiency larger than 95% for a reasonable completeness), they are powerful to prioritize targets for observation.

The probabilities obtained with ANNs for the 98 candidates unclassified in White et al. are found to be in good agreement (average difference 0.04 with standard deviation 0.06) and there is also a good agreement between the results for ANNs and OC1 (average difference 0.02 and standard deviation 0.13). Eight of these sources have recent spectroscopic classification in NED or SDSS DR2: five quasars have probabilities ranging from 0.71 to 0.94 (mean 0.82), two Ultraluminous Infrared Galaxies have $p \sim 0.1$ and a star has $p = 0.48$, reinforcing the ability of ANNs to optimize the selection of quasars.

Acknowledgments

RC and JIGS acknowledge financial support from DGES project PB98-0409-c02-02 and from the Spanish Ministerio de Ciencia y Tecnología under project AYA 2002-03326. This research has made use of the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Funding for the Sloan Digital Sky Survey (SDSS) has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS is managed by the Astrophysical Research Consortium (ARC) for the Participating Institutions. The Participating Institutions are The University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, University of Pittsburgh, Princeton University, the United States Naval Observatory, and the University of Washington.

References

Bishop C.M., 1995. Neural Networks for Pattern Recognition, Oxford University Press

Astronomical Data Analysis III 6
Tagliaferri R. et al., 2003, Neural networks in Astronomy. Neural Networks 16, 297