NEURAL NETWORKS FOR THE SELECTION OF QUASARS FROM RADIO-OPTICAL SURVEYS

R. Carballo¹, J. I. González-Serrano² and A. S. Cofiño¹

¹ Departamento de Matemática Aplicada y C. Computación, Universidad de Cantabria. Avda. de los Castros s/n. E-39005 Santander, Spain

² Instituto de Física de Cantabria. Avda. de los Castros s/n. 39005 Santander. Spain

Received: 2004 December 1

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 and 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 in the original list of candidates was 56%. The accuracy found with ANNs is similar to that obtained by White et al. using decision trees and training samples of similar size. 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. This work presents the first analysis of the performance of ANNs for quasar selection, showing that ANNs provide a promising technique to single out specific object types in astronomical databases.

Key words: methods: data analysis – methods: statistical – quasars: general

1. INTRODUCTION

Artificial neural networks (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) publish a list of quasar candidates drawn from the correlation of the VLA/FIRST radio survey with stellar 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 a decision tree classifier, 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 probabil-
ity 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 (Murthy, Kasif & Salzberg 1994) showed a very good performance, allowing one 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 activation functions used had the form $1/[1 + \exp(-z)]$ for the hidden nodes, and $2/[1 + \exp(-2z)] - 1$ for the output node. We applied the Levenberg-Marquardt optimization algorithm to minimize the mean of the squared errors $mse$, 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 modeling) 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. 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 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 splits to separate the training and validation sets. In order to choose the best ANN we first selected the splitting with better average of $mse$ for the training and the validation, in the sense that $mse$ 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 the values $p(Q)$ for the 1112 candidates.

2. RESULTS

Figure 1a shows the distribution of $p(Q)$ for the 7:2:1 ANN. The majority of the high-$p(Q)$ nonquasars are BL Lac objects. The reliability of quasar selection for $p(Q) > 0.85$ is 89% (372/418) and increases to 96% considering quasar or BL Lac selection (402/418). The corresponding completeness would be 59% (372/627) for quasars and 58% (402/694) for quasars or BL Lac. The ANN architecture 7:1 gives a probability distribution that is less peaked towards the extreme values, especially for high probabilities, but similar efficiency and completeness can be achieved.

© Teorines fizikos ir astronomijos institutas
Fig. 1. Left and central panels. Distribution of $p(Q)$ for the 7:2:1 ANN. The shaded distributions correspond to the objects of the indicated types. Right panel. Efficiency versus completeness for the 7:1 ANN (continuous line), the 7:2:1 ANN (dotted line) and the OC1 decision tree (dash-dotted line).

considering the candidates with $p(Q) > 0.75$, instead of 0.85. Both models find $p(Q) < 0.2$ for $\sim 40$ quasars and about 60% of these quasars have redshifts below 0.25.

The efficiency and completeness of the sample as a function of the quasar probability threshold is shown in Figure 1b 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.

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. The inspection of the NASA Extragalactic Database revealed that the sources J120354.7+371137 ($z = 0.401$, Appenzeller et al. 1998) and J125142.2+240435 ($z = 0.188$, Chen et al. 2002) have broad emission lines, and in fact we found for them high quasar probabilities ($\sim 0.94$ and $\sim 0.82$), reinforcing the high efficiency of the ANN models.

REFERENCES

Bishop C. M. 1995, Neural Networks for Pattern Recognition, Oxford Univ. Press
Tagliaferri R. et al. 2003, Neural Networks in Astronomy, Neural Networks, 16, 297

© Teorines fizikos ir astronomijos institutas