A Web Application for Photometric Redshift Estimation

O. Laurino[1]★, R. D'Abrusco[1,4], M. Brescia[2], A. Corazza[1], G. d'Angelo[1], C. Donalek[3], S.G. Djorgovski[3], N. Deniskina[1], M. Fiore[1], M. Garofalo[1], G. Longo[1,2], A. Mahabal[3], F. Manna[1], A. Nocella[1], B. Skordovski[1]
[1]Department of Physical Sciences, University Federico II, Naples, Italy
[2]INAF-OACN, National Institute of Astrophysics, Naples, Italy
[3]CALTECH, California Institute of Technology, Pasadena, California USA
[4]Department of Astronomy, University of Padua, Padua, Italy
★laurino@na.infn.it

***Abstract*—In the era of massive astronomical datasets, efficient identification of candidate quasars and the reconstruction of their three dimensional distribution in the Universe is a key requirement for constraining some of the main issues regarding the formation and evolution of QSOs. A method for the determination of photometric redshifts of QSOs based on multiwavelength photometry and on a combination of data mining techniques will be discussed. This procedure, specifically suited for accompanying the candidate selection method discussed in (D'Abrusco et al. 2008), makes use of specific tools developed under the EuroVO and NVO frameworks for data gathering, pre-processing and mining, while relying on the scaling capabilities of the computing grid. This method allowed us to obtain photometric redshifts with an increased accuracy (up to 30%) with respect to the literature.**

***Index Terms*—QSO, AGN, neural networks, photometric redshifts.**

I. Introduction

THE accurate knowledge of the shape and redshift evolution of the quasar luminosity function (QLF) is fundamental to many fields of modern cosmology [1]. For instance, the faint-end slope of the QLF, which is found to decrease with redshift (e.g. [2], [3], [4]), is important in determining the early formation history of Black Holes (BHs) and their contribution to reionization, as well as the possible connections between quasars and, e.g., the low-luminosity Seyfert galaxies seen at $z \simeq 0$. In traditional models, this trend requires a significant and rapid evolution in the shape of the distribution of the masses of the host galaxies, which cannot be accounted for in either semianalytical models or numerical simulations and is not consistent with a wide range of galaxy observations (e.g., [5]). These models, in fact, generally succeed at high redshift but do not explain the decrease in counts of bright quasars at low redshift (e.g., [6]), unless other ad hoc mechanisms are invoked such as, for instance, the suppression of the growth of high-mass spheroids (e.g., [7]), or the evolution in the BH accretion efficiency with redshift (e.g., [8]). In other words, these models cannot be extrapolated to low luminosities or to redshifts where the slope is undetermined. Observations at high redshifts are uncertain, and will remain so until large, uniformly selected samples will allow to measure the faint-end slope at both low ($z < 1$) and high ($z > 3$) redshifts. The construction of such samples is troublesome due to the necessity of obtaining spectra for large samples of mainly faint objects: this task which is both very demanding in terms of observing time and challenging in terms of signal to noise ratio. In recent years, however, the availability of deep, accurate and multiband digital surveys such as the Sloan Digital Sky Survey (SDSS, [15]), the Two Degree Fields (2dF), or the UKIDSS ([16]) has opened a new possibility: the search for

quasars in the photometric parameter space and the evaluation of their distances using photometric redshifts. This approach uses photometry to determine the redshift of the observed object relying upon the fact that the spectrum of radiation emitted by the sources possess very strong features that are detected in different filters of the photometric system. The photometric redshift of a source can then be calculated by establishing an empirical relation between the brightness in different filters and the actual redshift of the source, and finally applying such relations to sources for which only the photometric parameters have been measured.

Such approach, however, is constrained by the availability of a reliable (in terms of completeness and efficiency) sample of candidate quasars, and the accuracy with which it is possible to obtain an estimate of their distance. As it has been shown in several papers (cf. [9], [11]), the first task has already been successfully completed but the second step still poses few problems.

In what follows we shortly outline a new method based on Neural Networks which, by making use of the S.Co.P.E. computing infrastructure, offers to the Virtual Observatory (VObs) community the possibility to derive photometric redshifts for both galaxies and quasars, provided that a large enough spectroscopic base of knowledge on which to train the algorithms is available. The application of machine-learning techniques for the evaluation of photometric redshifts of galaxies has been widely explored in the literature (cf. [13] and references therein) but the problems posed by quasars are very different. In the case of these type of extragalactic sources, in fact, the base of knowledge (BoK) is characterized by a high degree of sparseness in the parameter space since quasars are observed in a very large interval of redshifts (from ~ 0.5 to ~ 6) and cover a large region of the photometric space so that the average density of BoK members can be very low. The application of this method to the catalogue of candidate quasars produced by [11] will lead to the evaluation of the luminosity function of candidate quasars in the optical magnitudes of the SDSS ([14], in prep.).

II. The Method

ONE possible approach to the estimation of the photometric redshifts is based on supervised machine learning algorithms such as the neural networks [12], trained on a subsample of the photometric data set for which spectroscopic redshifts are available (the BoK). The method is then capable of producing accurate estimates of the photometric redshifts for sources found in the same regions of the photometric parameter space spanned by the members of the BoK.

The photometric redshift estimation can be regarded as a regression problem: we seek out a functional mapping

$$f : c \in \mathcal{C} \rightarrow z_{phot} \in \mathcal{Z}$$

from the parameter space $\mathcal{C}$ of the source's colours to the target space $\mathcal{Z}$ of the redshifts.

A. The Multi-Layer Perceptron

To solve regression problems, the Multi-Layer Perceptron (MLP) is one of the most robust and reliable non-deterministic, machine learning techniques: i) it is a universal function approximator and does not require any assumptions on f; ii) it has a good response when dealing with noisy data and, iii) it has remarkable generalization properties. In a nutshell, the problem is the following: we need to *train* an MLP network to *learn* the mapping function f.

As for every supervised machine learning algorithm, it requires the BoK to be split into different subsets:

- *Training Set*. 60% of the objects will be used to provide the network with *examples* to induce the mapping function form;
- *Validation set*. We save 20% of the data to avoid overfitting: in principle (and in the absence of degeneracies), a neural network could approximate a function with arbitrary precision given a sufficient number of hidden nodes. We stop the training phase when the error on the validation set is minimum.
- *Test set*. The remaining objects (20%) are eventually used to test the overall performance of the network.

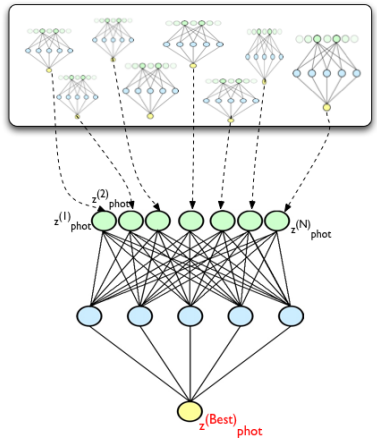


Fig. 1. Multilayer Perceptrons in ensemble

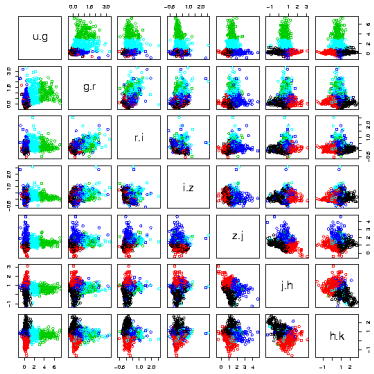


Fig. 2. Clustering of a multidimensional dataset with k-means

In the specific case of quasar redshifts, the regression problem suffers from the intrinsic deficiency of the data due to the noise induced by the *degeneracies*. So, it is preferable to train an MLP *ensemble*, where each network is trained (and optimized) on a specific partition of the parameter space. We wish to recall that MLP training is an NP-hard optimization problem, so it can be computationally expensive, with a computational cost that grows linearly with the number of input patterns and the dimensionality of the features space.

B. Partitioning The Parameter Space

Clustering is an unsupervised data mining method for organizing objects into groups according to a given definition of distance. We can also look at this data partitioning method, from a geometrical standpoint, as providing a way to divide the parameter space into different regions according to a given criterion. It has to be stressed that, as for all the unsupervised methods, there is not a unique criterion to decide whether a specific partition is better than another; neither we know the best number of clusters to partition the parameter space in. Again, we can exploit our Base of Knowledge to set a *feedback* criterion based on the final regression results to infer the best number of clusters.

k-means is one of the simplest, yet robust, clustering algorithms: given the number of clusters (k) and using an euclidean distance, k-means finds the k cluster centroids aggregating inputs with similar properties. When splitting the dataset into clusters we decimate the patterns the neural networks will be trained on. This could not be much of a problem when the parameter space is densely covered by the patterns and there is a large number of training objects. However, when dealing with QSO catalogues, the sparseness of the data does not allow to ignore the effects of this decimation on the training process, and it is necessary to find a different training strategy. This can be done by taking redundant, overlapping clusters so to introduce a coupling in the way network are trained. This coupling can be exploited with the *ensemble* approach itself, as it will be shown in the following sections.

To produce such redundant clustering we made use of a fuzzy variant of the k-means clustering algorithm: each point in the parameter space belongs to each cluster with a non-zero membership probability; we then set an arbitrary threshold T, so that a cluster actually contains only points with a membership larger that T. With an iterative procedure the fuzzy k-means algorithm is run several times with a variable number of clusters between a minimum and a maximum value. The optimal number of clusters

N is defined by the value for which the overall regression error is minimized.

C. Neural Networks ensemble

Given N different clusters, each cluster represents the BoK for a different MLP network, and these BoKs share a certain variable number of sources, with the effect of introducing a coupling in the training of the different networks. In other words, each of the N networks is an *expert* for the sub-domain of the parameter space it has been trained on, and these sub-domains are overlapping. Without the *ensemble* approach, in order to estimate the photometric redshift of a generic point (source) $c \in \mathcal{C}$, we should assign the point to one of the k clusters in the first place, and then run the correspondent MLP network. Alternatively, in our method, each input pattern c is presented to all the MLPs, allowing them all to provide their "opinion", and then these opinions (i.e. different guesses of the photometric redshift) are combined to provide our best estimation. According to the way we determined the clusters, the outputs of the single networks will be coupled and different networks will have a different reliability on different regions of the parameter space. The combination is then performed by means of a *gating* MLP network that *learns* how to combine the predictions coming from the first layer networks.

We measure the performance of the method by calculating the Mean Square Error over the test set. Each different clustering scheme will yield to a different MSE value. We define the best number of clusters N as the number that yields to the minimum value of the MSE error over the test set.

D. The method as a whole

The workflow we have briefly depicted so far, involves several optimization problems. Given the smallest and largest number of clusters we want to probe, N_m and N_M, for all k between N_m and N_M we have to:

- perform a k-means clustering to find the k centroids;
- train k different MLP networks.

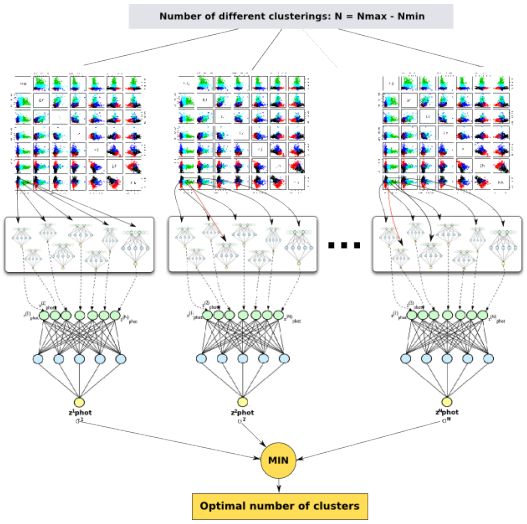


Fig. 3. A flow chart of the whole method.

The number of first layer MLPs to train is

$$N_{net} = \sum_{k=N_m}^{N_M} k = \frac{N_M(N_M+1)}{2} - \frac{N_m(N_m-1)}{2}$$

If $N_{clu} = N_M - N_m + 1$,, the total number of optimization problems is $N_p = N_{net} + 2N_{clu}$,, since we have to take into account the gating networks and the clusterings. For $N_m = 2$ and $N_M = 10$, we have 72 optimization problems. However, since the k-means's time complexity is linear in all the relevant factors (iterations, space dimensionality, number of clusters, number of points), we can neglect the computational cost of the clustering task and consider just the 63 MLP training tasks, i.e. 63 NP-hard optimization problems. All these tasks are obviously independent, so they can be launched in parallel on the computing GRID.

E. The data

We performed a set of experiments aimed at the estimation of photometric redshifts for quasars candidates extracted from the optical parameter space of SDSS [15] stellar sources

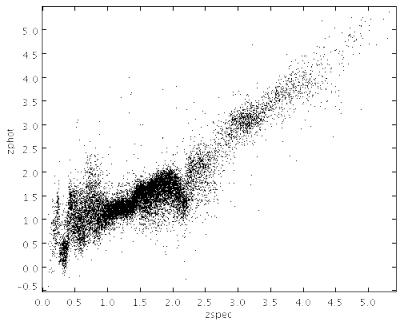


Fig. 4. z_{phot} vs z_{spec} scatter plot: $\sigma = 0.23$. The dataset is a catalogue of optical candidate quasars extracted from the SDSS. See text for details.

defined by the colours $(u-g, g-r, r-i, i-z)$. These candidates were obtained via an original method based on unsupervised clustering (cf. [11]) which has yielded a catalogue of QSO candidates in the SDSS (cf. the webpage at voneural.na.infn.it/qso.html). The colours of the candidate quasars were obtained using the *psfgMag* magnitudes retrieved from the $PhotoObjAll$ table of the SDSS DR7 database, after culling the dataset excluding all sources with missing magnitudes or NaN. As BoK, we used the spectroscopic redshifts obtained from the SDSS spectra for a subsample of photometric candidates down to a limiting magnitude g 17.7. In order to obtain a clean sample of confirmed quasars and to avoid the introduction of biases in the training process of the neural networks, only spectroscopic redshifts with confidence *zConf* > 0.90 were included.

F. A preliminary result

As first scientific application, we have obtained photometric redshifts for a sample of optical candidate quasars extracted from the SDSS, obtaining the accuracy (measured as the robust standard deviation of the $z_{phot} - z_{spec}$ variable for the quasars of the test set) $\sigma = 0.23$, which is better than the accuracies of most of the methods found in the literature (e.g. [9], [10]).

III. The method as a VObs Web Application

THE VObs aims at providing the world astronomical community with a comprehensive, consistent and interoperable infrastructure for sharing and federating the massive amounts of astronomical data observed or simulated by astronomers worldwide. By means of the VObs it has become possible to access a multidimensional parameter space and to retrieve the most complete Base of Knowledge available. For photometric redshifts this means that we can build rich multiwavelength BoKs significantly improving our prediction accuracy, at the price of increased computational cost. Supervised methods need to be trained on a well defined BoK, and their predictions are reliable as far as the input patterns fall in the same regions of the BoK they have been trained on. Photometric redshift estimation can obviously improve as new colours are added to the feature space, but in order to exploit the information carried by this extension of the parameter space, specific networks need to be trained on the new parameter space in the first place. This is the main reason why, since its start, the method presented above was conceived as a Virtual Observatory (VObs) tool to be offered to the community as a web application through the DaME/VOneural platform. DAME (Data Mining & Exploration) is described elsewhere in this volume [17]. In its final version, our application will:

- ask the user for a catalogue of extragalactic sources, defined by their names or their sky coordinates;
- for each source, gather all the relevant information within the VObs;
- dynamically call the properly trained models;
- if the spectroscopic redshift is available for a source, retrieve it;
- output the original catalogue with some columns added: the photometric redshift, the spectroscopic (if any) redshift and the respective uncertainties;
- optionally, retrieve from the VObs some general information about the source, and

display it to the user.

A prototype of this web application is already served along with the DAME/VONeural front end at the address http://dame.na.infn.it

IV. Conclusions

OUR method for photometric redshift estimation of quasars was designed to improve upon the accuracy of available methods. The key role is played by the way we handle the intrinsic deficiency of the base of knowledge: for quasars, the BoK can be very sparse, and degeneracies may likely be introduced by the spectroscopic features shifting off the photometric system filters used for the observations. To overcome this deficiency we apply, for the first time in astronomy, fuzzy clustering techniques to the original catalogue, in order to implement redundant training sets. We train an *ensemble* of artificial neural networks on these subsets, so that the networks become *expert* at specific overlapping regions of the feature space. Input patterns are then presented to all the experts and a gating neural network is trained on their outputs in order to combine them. The subsets are determined by the fuzzy k-means clustering algorithm. This method, already publicly available as a web application at the address http://dame.na.infn.it, exploits the computational capabilities of the S.Co.P.E. computing GRID and is fully compliant with the Virtual Observatory standards and infrastructure.

Acknowledgments

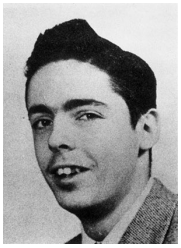
This work was funded by the S.Co.P.E. project, by the VO-Tech and VO-AIDA European projects and by the DAME project financed by the Ministry of Foreign Affairs as a "Great Relevance Bilateral Project" between Italy and U.S.A.

References

[1] Hopkins P.F., Hernquist L., Cox T.J., Robertson B., Di Matteo T. & Springel V., 2006, ApJ, 639, 700.
[2] Hunt, M. P., Steidel, C. C., Adelberger, K. L., & Shapley, A. E. 2004, ApJ, 605, 625
[3] Cirasuolo, M., Magliocchetti, M., & Celotti, A. 2005, MNRAS, 357, 1267
[4] Hasinger, G., Miyaji, T., & Schmidt, M. 2005, A&A, 441, 417
[5] Merloni, A. 2004, MNRAS, 353, 1035
[6] Enoki, M., Nagashima, M., & Gouda, N. 2003, PASJ, 55, 133
[7] Scannapieco, E., & Oh, S. P. 2004, ApJ, 608, 62
[8] Haiman, Z., & Menou, K. 2000, ApJ, 531, 42
[9] Richards, G. T. et al. 2008, ApJS, 180, 67-83.
[10] Weinstein, M. A., Richards, G. T. et al., 2004, ApJS, 155, 243
[11] R. D'Abrusco, G. Longo and N.A. Walton, accepted by MNRAS, 2008.
[12] R. D'Abrusco et al. 2007, ApJ, 632, 752.
[13] R. D'Abrusco, PhD Thesis, University of Napoli Federico II, 2007.
[14] O. Laurino et al., in preparation, 2009.
[15] C. Stoughton et al. 2002, AJ, 123, 485-548.
[16] A. Lawrence et al. 2007, MNRAS, 379, 1599-1617.
[17] Brescia M., et al. 2009, these proceedings.
[18] Bishop, C. M. 1995, Neural Networks for Pattern Recognition, New York, Oxford Univ. Press.

Omar Laurino is both a physics Master of Science student and an IT worker. He worked as a system administrator for both the Italian Institute of Nuclear Physics (INFN) and the Department of Physical Sciences of the University of Naples, and is now taking (at last) his *laurea* in physics with a thesis on the reconstruction of the three-dimensional distribution of candidate QSOs with photometric redshifts and the derivation of their luminosity function.

He is the Project Engineer of the DaME/VONeural project, a data mining framework for massive datasets (also in these proceedings).



Raffaele D'Abrusco is currently a post-doc researcher at the Department of Astronomy of the University of Padova, after having spent 1 year as post-doc at the University of Naples. He earned his PhD in Physics at the same University of Naples "Federico II" on 21st December 2007 defending a thesis titled "The Large Scale Structure of the Nearby Universe". His area of scientific interests is the extragalactic astronomy, in particular the study of the large scale distribution of galaxies, the observational characterization of AGNs and QSOs and the application of new data mining tools and statistical techniques to several astronomical topics. He collaborates with the DaME/VONeural group.