A Web Application For Photometric Redshift Evaluation

O. Laurino[1], R.D'Abrusco[1], M.Brescia[2], S.Cavuoti[1], A.Corazza[1], G.D'Angelo[1], C.Donalek[3], G.Djorgovski[3], N.Deniskina[1], M.Fiore[1], M.Garofalo[1], G.Longo[1], A.Mahabal[3], F.Manna[1], A.Nocella[1], G.Riccio[1], B.Skordovski[1]

[1]Dipartimento di Scienze Fisiche, Università Federico II di Napoli
[2]Istituto Nazionale di AstroFisica – OACN
[3]California Institute of Technology

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 accurate determination of the photometric redshifts of QSOs based on multiwavelength photometry and 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 allows us to obtain photometric redshifts with an increased accuracy (up to 30%) with respect to the literature.

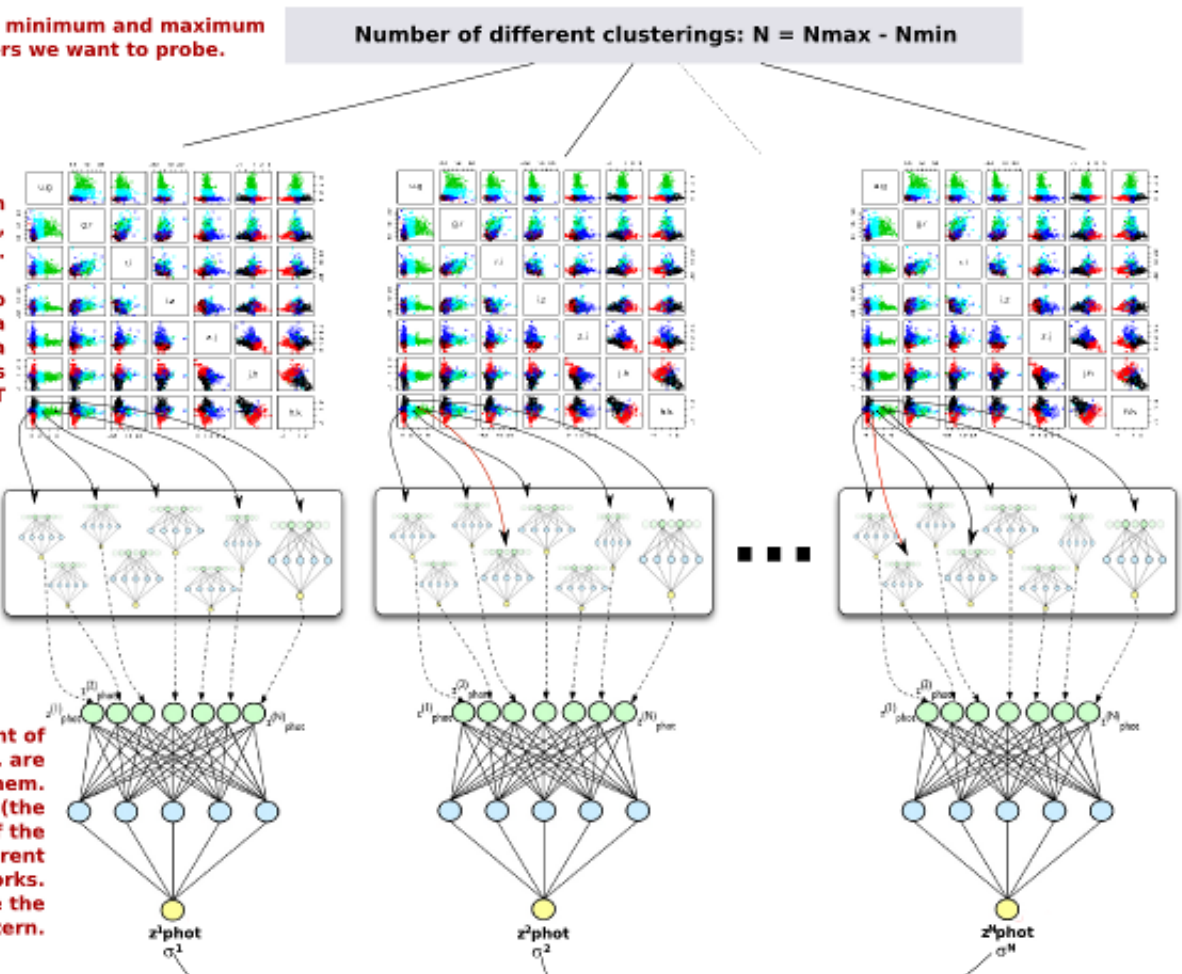
Photometric Redshifts

1) A method to evaluate distances when spectroscopic estimates become impossible due to either poor signal-to-noise ratio or to instrumental systematics, or to the fact that the sources under study are beyond the spectroscopic limit;

2) An economical way to obtain, at a relatively low price in terms of observing and computing time, redshift estimates for large samples of sources.

Clustering is an unsupervised method for partitioning the parameter space (quasars colours space) into different regions according to a given definition of distance. In our approach clustering is used to enhance the performance of the Neural Networks Regression training.
Since there is no preferable criterion to determin which clustering is better, we use a feedback approach, which is described below

1. We decide the minimum and maximum number of clusters we want to probe.



2. We apply a clustering algorithm N=Nmax-Nmin times. For each partition n clusters are produced, with n ranging from Nmin to Nmax.

Due to the sparseness of the data, we decide to use a fuzzy approach, so that each object has a non-zero membership to all the clusters. We set a threshold T so that each cluster actually contains only objects with a membership m>T

3. We train a different Neural Network (MLP) for each cluster. Thus, each Network is "expert" in each region of the colours space. Once trained, we "test" the Networks to determine their generalization level. The Networks "learn" the function that maps colours into redshifts.

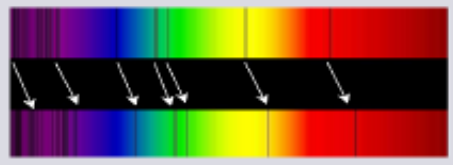
4. Since the clusters share a certain amount of objects, the networks trained at the step 3. are correlated among them.
We can then train a new Neural Network (the "Gate") to give the best estimation of the photometric redshift based on the different "opinions" of the input networks.
By means of a validation set we determine the overall error for a specific clustering pattern.

5. The smallest validation error gives the number of clusters that provides the most accurate results. By means of a test set, we can determine the generalization capability of the algorithm and the overall error.

Optimal number of clusters

A slightly different version of this algorithm makes use of a more traditional crispy clustering and is applied to galaxies, where the base of knowledge is vast and not as sparse as for a typical QSO sample

Photometric vs Spectroscopic Redshift



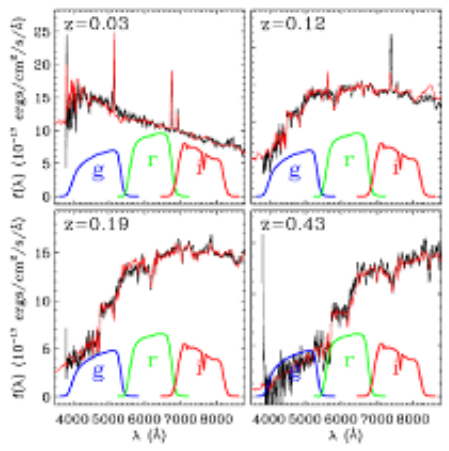
Extragalactic sources show a shift of their spectra due to the Doppler Effect induced by the expansion of the Universe and their motion with respect toobject this expansion. The Hubble law states that the further the source the larger the shift towards the red.
So, to calculate the redshift one usually needs the complete spectrum of the object.

But spectra are expensive, in terms of observation time, and we possess the spectra of just a small fraction of all the astronomic sources.

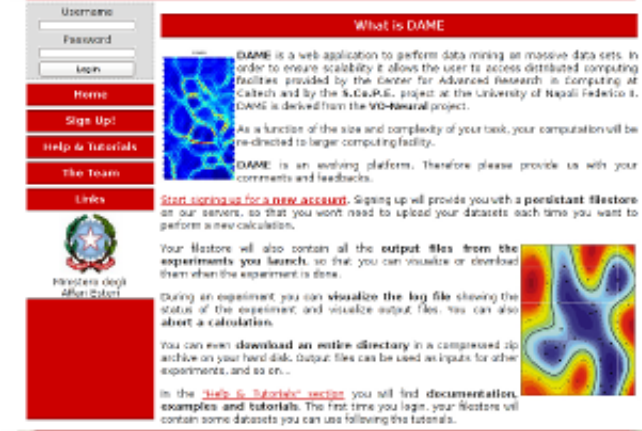
On the other hand, we have huge datasets full of multiwavelength photometric data.

The redshift can be calculated by means of this photometric information, but with a much lower accuracy.

We make use of Machine Learning Supervised techniques to train Artificial Neural Networks to map the photometric information of a source to its spectroscopic redshift based on a spectroscopic base of knowledge.



The Web Application

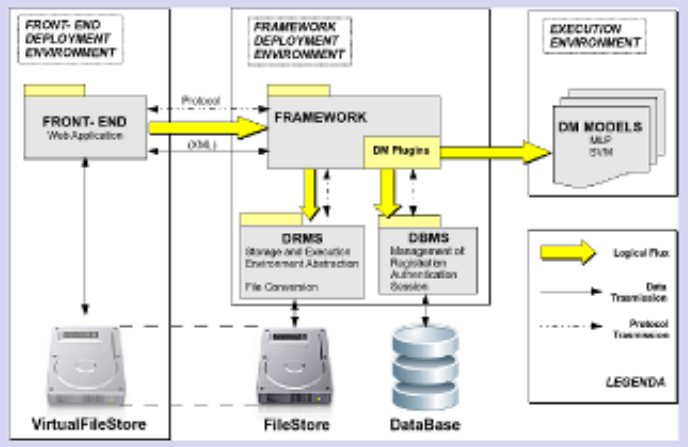


The Artificial Neural Networks Training is a computing intensive task when the number of training examples is large as it is in astronomical Data Mining.
In our approach, we have two different optimization problems: the ANN training process itself and the clustering optimization.
This yields to a tree of many ANN to train in parallel.
The GRID can easily boost this process.
The DAME/VONeural Framework makes extensive use of the GRID technology to store the data and perform calculations in a "Cloudy" fashion, by means of a "robot certificate" to allow the safe access to the GRID resources.

As it is shown in the picture above, the spectroscopic "features" shift down the photometric bands.
In order to avoid degeneracies and to increase accuracy one can federate different tables from different surveys.
The Virtual Observatory provides us with all the tools needed to build multiwavelength datasets on the fly.
The final goal of our approach is to create a fully automated application that queries the VO resources for all the photometric information available for a catalogue of extragalactic sources, so to provide the best estimation of the photometric redshift available with all the information at our disposal in the VO infrastructure.

The DAME/VONeural Framework

An integrated, web oriented, platform independent, extensible framework for Data Mining.
Integrated: the user can store his data and save his experiments and sessions. The developer will download a dedicated set of APIs.
Web oriented: the framework is based on a Service Oriented Architecture. We are also developing a dedicated client Web Application.
Platform independent: GRID is the main platform, but different "drivers" will be available.
Extensible: Data Mining models, complex tasks, middle-ware drivers and even the client application can all be extended and augmented by means of a consistent, Object Oriented plug-in architecture.



Exploiting the Virtual Observatory and GRID Computing Capabilities