The Astronomical Data Deluge: the Template Case of Photometric Redshifts

(Extended Abstract)

© Giuseppe Longo[1,2] © Massimo Brescia[2,1] © Stefano Cavuoti[1,2]

[1] Department of Physics, University Federico II, Napoli, Italy
[2] INAF Astronomical Observatory of Capodimonte, Napoli, Italy

longo@na.infn.it brescia@oacn.inaf.it cavuoti@na.infn.it

Abstract. Machine learning methods have become crucial to many aspects of astrophysics and cosmology. We focus on the evaluation of photometric redshifts as a template case of classification/regression problem in astronomical data mining. We discuss the general aspects of the problem and some recent work which tries to solve the issues posed by optimal feature selection, missing data and by the evaluation of probability distribution functions.

Keywords: data intensive domains, astrophysics big data, machine learning, photometric redshifts.

1 Introduction

Multiband, multi-epoch digital sky surveys are producing a tsunami of complex, high quality data, which is changing the landscape of astrophysical research. New generation survey telescopes such as the Large Synoptic Survey Telescope (LSST) and Euclid in the optical domain, or the Square Kilometer array (SKA) in the radio domain, will soon produce many tens of TB of processed data every day, and on the long term will provide hundreds of measured parameters for billions of sources. An unprecedented wealth of high quality, accurate and complex data – stored in distributed data centers - that on the long term is expected to revolutionize our understanding of the universe. In order to cope with this data overabundance, all steps of the data understanding chain – acquisition, reduction, analysis, visualization and interpretation – are being deeply transformed and machine learning methods (ML) are becoming crucial at every stage of the process. In particular, modern precision cosmology requires accurate information on both type and redshift (i.e. the distance) for very large (in the hundreds of millions) samples of galaxies. This task cannot be accomplished by means of traditional spectroscopic techniques and in recent years there has been an explosion of alternative methods based on the exploitation of the information contained in multiband photometry: the so called photometric redshifts (hereafter photo-z). A very effective and promising approach to the evaluation of photo-z relies on ML methods. Many different implementation have appeared in the specialized literature based on different flavors of (Multi Layer Perceptrons) MLP's [cf. 1, 2, 3], random forest [4], nearest neighbors [5], active learning [6], etc. all with their slight advantages and disadvantages. Therefore, rather than focusing on a specific method, we shall discuss the general aspects of the problems and some ongoing work addressing the main issues: characterization of the knowledge base, feature extraction and selection, missing data and evaluation of errors.

2 Photo-z with ML Methods

2.1 The Knowledge Base

From a ML point of view, the evaluation of photo-z is a classification/regression problem, where the chosen method learns how to estimate the redshift of a galaxy interpolating the knowledge available for a small but significant subsample of objects with known spectroscopic redshifts (knowledge base or KB). After training, the methods (and the underlying mapping function) can be applied to those objects for which the spectroscopic redshift is not available. Data augmentation techniques have been tested but did not lead to reliable results. More promising seems to be the combination of machine learning methods with other techniques, (such as, for instance, template fitting [7]).

This process has two implications, one rather obvious and the other much less so. First, the methods cannot be applied to objects outside of the parameter space sampled by the KB (for instance, fainter than the spectroscopic limit). Second, methods often fail to capture the properties of objects which, being intrinsically rare or peculiar, are not well represented in the KB. Given the complexity of the extragalactic zoo that spans over a very wide variety of observed and physical properties, understanding the properties of the KB becomes crucial. This will be particularly relevant if we take into account that almost all we know about systematic in photo-z comes from optically selected samples, while some surveys of the future will deal with radio (e.g. SKA) or X-ray (e.g. e-Rosita) selected samples. Some recent attempts have been made which are worth mentioning. In [8] a SOM was used to map the photometric space ex-

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

pected for the Euclid space mission in order also to define the optimal strategy to build the KB.

2.2 Features Extraction and Feature Selection

Digital surveys produce for each observed object many hundreds of parameters that are often highly correlated. These features (i.e. fluxes within a given aperture, radii, concentration indexes, etc.) are usually derived using recipes based on the expertise of astronomers. A pioneering work [9] based on a purely data driven approach, has recently shown that traditional features, almost always fail to capture the subtleties of the information contained in the raw data. This calls for a new way to access the information contained in the astronomical images. While this process is still in its infancy, there are clear signs that deep learning can be greatly beneficial (K. Polsterer, priv. comm.).

In any case, due to both computational constraints and to the need to optimize the dimensionality of the parameter space, feature selection remains a crucial problem that only recently has begun to be properly addressed within the astronomical community. At the moment, two approaches seem to be viable: a brute force approach, where all possible combinations of features are tried until a plateau in the performances (defined by some metrics) is reached [9, 10] and Cavuoti (priv. comm.). This approach, however, is computationally demanding and not very flexible. A different path to the identification of the optimal set of features, is currently being implemented by Brescia and collaborators (Brescia et al. 2017, in preparation).

2.3 Missing Data and Non Detection

Most ML methods do not deal effectively with “missing data” (or NAN) and in many cases incomplete data need to be rejected from the sample. This is no longer possible in many modern astronomical applications where incomplete data might affect a quite large fraction of the objects. Furthermore, we need to take into account that in astronomical applications we encounter two types of missing data: “true” missing data (e.g. objects in a region of the sky not observed in a specific band) and “non detection” (e.g. objects which are observed but not detected in one or more photometric band). Dealing with these two types of missing data obviously pose different problems since the latter contain some information (for instance: an upper limit to the flux) that needs to be taken into account. A new approach has been implemented and tested (Cavuoti et al. in preparation) that makes use of a nearest-neighbors approach, to optimize and reconstruct missing information. This approach has been validated on a variety of real data sets.

2.4 Probability Distribution Functions

In many real science applications of photometric redshifts (e.g. weak lensing and shear map reconstruction) one of the main requirements is the need to provide a PDF (Probability Distribution Function) for both the global distribution and the individual objects. Such requirement cannot be met in a trivial way using ML based techniques, since the analytical relation mapping the photometric parameters onto the redshift space is virtually unknown. The tool METAPHOR (*Machine-learning Estimation Tool for Accurate PHOtometric Redshifts*, [11]) was implemented as a modular workflow, whose internal engine for photo-z estimation makes use of MLPQNA (Multi Layer Perceptron with Quasi Newton Approximation; [1]), with the possibility to easily replace the specific machine learning model. METAPHOR takes into account all possible sources of error both internal to the method (e.g. initialization errors) and external (e.g. photometric errors). METAPHOR is independent on the specific ML method used to evaluate the photo-z (it has been extensively tested using several implementation of MLP’s and Random Forest algorithm. Recent tests on the KiDS (Kilo Degree survey; [12]) Third Data Release confirmed the robustness of the approach [13].

References

[1] Cavuoti, S. et al.: Photometric Redshifts with Quasi Newton Algorithm (MLPQNA). Results in the PHAT1 Contest, Astr. & Astroph., 546, p. 13 (2012)

[2] Brescia, M. et al: DAMEWARE: A Web Cyber-infra-structure for Astrophysical Data Mining, Publ. Astron. Soc. of Pacific, 126, p. 783 (2014)

[3] Sadeh, I. et al., ANNz2: Photometric Redshift and Probability Distribution Function Estimation using Machine Learning, Publ. Astron. Soc. of Pacific, 128 (2016)

[4] Carliles, R. et al.: Random Forests for Photometric Redshifts, Astrop. J., 712, p. 511 (2016)

[5] Sheldon, E. S. et al.: Photometric Redshift Probability Distributions for Galaxies in the SDSS DR8, Astrop. J. Suppl. Series, 2012, 32 (2012)

[6] Bo, H. et al.: Active Learning Applied for Photometric Redshift Estimation of Quasars, AAS, 2015IAUGA.2256851H (2015)

[7] Cavuoti, S. et al., A Cooperative Approach Among Methods for Photometric Redshifts Estimation: an Application to KiDS Data, MNRAS, 466, p. 2039 (2017)

[8] Masters, D. et al: Mapping the Galaxy Color-Redshift Relation: Optimal Redshift Calibration Strategies for Cosmological Surveys, Astrop. J., 813, p. 53 (2015)

[9] Polsterer, K. et al., Improving the Performance of Photometric Regression Models via Massive Parallel Feature Selection, Proc. of Astronomical Data Analysis Software and Systems XXIII, p. 425 (2013)

[10] D’Isanto, A. et al: An Analysis of Feature Relevance in the Classification of Astronomical Transients with Machine Learning Methods, MNRAS, 457, pp. 3119-3132 (2016)

[11] Cavuoti, S. et al.: METAPHOR: A Machine Learning Based Method for the Probability Density Estimation of Photometric Redshifts, MNRAS 465, 1969 (2016)

[12] de Jong, J. T. A. et al: The Third Data Release of the Kilo-degree survey and Associated Data Products, Astr. & Astrop. (arXiv:1703.02991) (2017)

[13] Amaro, V. et al: Machine Learning Based Photometric Probability Density Functions for the KiDS ESO DR3 Galaxies, MNRAS (2017)