Astroinformatics
Proceedings IAU Symposium No. 325, 2016
M. Brescia, S.G. Djorgovski, E. Feigelson,
G. Longo & S. Cavuoti, eds.

doi:10.1017/S1743921317001296

Cooperative photometric redshift estimation

S. Cavuoti[1], C. Tortora[2], M. Brescia[1], G. Longo[3], M. Radovich[4], N. R. Napolitano[1], V. Amaro[3] and C. Vellucci[5]

[1]INAF - Astronomical Observatory of Capodimonte, via Moiariello 16, I-80131 Napoli, Italy
[2]Kapteyn Astronomical Institute, Univ. of Groningen, P.O. Box 800, 9700 AV Groningen, the Netherlands
[3]Department of Physics, University Federico II, Via Cinthia 6, I-80126 Napoli, Italy
[4]INAF - Astronomical Observatory of Padua, vicolo dell'Osservatorio 5, I-35122 Padova, Italy
[5] DIETI, University of Naples Federico II, Via Claudio,21 I-80125 Napoli, Italy

Abstract. In the modern galaxy surveys photometric redshifts play a central role in a broad range of studies, from gravitational lensing and dark matter distribution to galaxy evolution. Using a dataset of $\sim 25,000$ galaxies from the second data release of the Kilo Degree Survey (KiDS) we obtain photometric redshifts with five different methods: *(i)* Random forest, *(ii)* Multi Layer Perceptron with Quasi Newton Algorithm, *(iii)* Multi Layer Perceptron with an optimization network based on the Levenberg-Marquardt learning rule, *(iv)* the Bayesian Photometric Redshift model (or BPZ) and *(v)* a classical SED template fitting procedure (Le Phare). We show how SED fitting techniques could provide useful information on the galaxy spectral type which can be used to improve the capability of machine learning methods constraining systematic errors and reduce the occurrence of catastrophic outliers. We use such classification to train specialized regression estimators, by demonstrating that such hybrid approach, involving SED fitting and machine learning in a single collaborative framework, is capable to improve the overall prediction accuracy of photometric redshifts.

Keywords. methods: data analysis, methods: statistical, catalogs

1. Introduction

Photometric redshift produced through the modern multi-band digital sky surveys are crucial to provide a reliable distance estimation for a large number of galaxies in order to be used for several tasks in precision cosmology, to mention just a few: the weak gravitational lensing to constrain dark matter and dark energy, the identification of galaxy clusters and groups, the search of strong lensing and ultra-compact galaxies, as well as the study of the mass function of galaxy clusters. We can derive photometric redshift (hereafter photo-z) thanks to the existence of a hidden (and complex) correlation among the fluxes in the different broad bands, the spectral types of the object itself, and the real distance. Although hidden, the mapping function that can map the photometric space into the redshift one could be approximated in several ways, and the existing methods can be broadly divided into two main classes: theoretical and empirical.

In the previous work (Cavuoti *et al.* 2015a) we have already applied an empirical method, the Multi Layer Perceptron with Quasi Newton Algorithm, MLPQNA, (Cavuoti *et al.* 2012, Brescia *et al.* 2013, Brescia *et al.* 2014, Brescia *et al.* 2015, Cavuoti *et al.* 2017), to a dataset extracted from the Kilo Degree Survey (KiDS). Here we apply five different photo-z techniques to the same dataset and then we analyze the behavior of such methods with the aim at finding a way to combine their features in order to to optimize the accuracy of photo-z estimation; a similar, but reversed approach was followed recently by Fotopoulou *et al.* (2016).

2. The data

As stated before we used the photometric data from the KiDS optical survey (de Jong *et al.* 2015). The KiDS data releases consist of tiles which are observed in the *u, g, r*, and *i* bands. The sample of galaxies on which we performed our analysis is mostly extracted from KiDS-DR2 (de Jong *et al.* 2015), which contains 148 tiles observed in all filters during the first two years of survey regular operations. We added 29 extra tiles, not included in the DR2 at the time this was released, that will be part of the forthcoming KiDS data release, thus covering an area of 177 square degrees.

We used the multi-band source catalogs, based on source detection in the *r*-band images. While magnitudes are measured in all filters, the star-galaxy separation, as well as the positional and shape parameters are derived from the r-band data only, which typically offers the best image quality and seeing $\sim 0.65''$, thus providing the most reliable source positions and shapes. The KiDS survey area is split into two fields, KiDS-North and KiDS-South, KiDS-North is completely covered by the combination of SDSS and the 2dF Galaxy Redshift Survey (2dFGRS), while KiDS-South corresponds to the 2dFGRS south Galactic cap region. Further details about data reduction steps and catalog extraction are provided in de Jong *et al.* (2015) and Tortora *et al.* (2016).

Aperture photometry in the four *ugri* bands measured within several radii was derived using S-Extractor (Bertin & Arnouts 1996). In this work we use magnitudes `MAGAP_4` and `MAGAP_6`, measured within the apertures of diameters $4''$ and $6''$, respectively. These apertures were selected to reduce the effects of seeing and to minimize the contamination from mis-matched sources. The limiting magnitudes are: MAGAP_4_u = 25.17, MAGAP_6_u = 24.74, MAGAP_4_g = 26.03, MAGAP_6_g = 25.61, MAGAP_4_r = 25.89, MAGAP_6_r = 25.44, MAGAP_4_i = 24.53, MAGAP_6_i = 24.06. To correct for residual offsets in the photometric zero points, we used the SDSS as reference: for each KiDS tile and band we matched bright stars with the SDSS catalog and computed the median difference between KiDS and SDSS magnitudes (*psfMag*). For more details about data preparation and pre-processing see de Jong *et al.* (2015) and Cavuoti *et al.* (2015a).

In order to build the spectroscopic Knowledge Base (KB) we cross-matched the KiDS data with the spectroscopic samples available in the GAMA data release 2, (Liske *et al.* 2015), and SDSS-III data release 9 (Ahn *et al.* 2012). The detailed procedure adopted to obtain the data used for the experiments was as follow: *(i)* we excluded objects having low photometric quality (i.e., with flux error higher than one magnitude); *(ii)* we removed all objects having at least one missing band (or labeled as Not-a-Number or NaN), thus obtaining the cleaned catalogue used to create the training and test sets, in which all photometric and spectroscopic information required is complete for all objects; *(iii)* we performed a randomly shuffled splitting into a training and a blind test set, by using the 60%/40% percentages, respectively; *(iv)* we applied the cuts on limiting magnitudes (see Cavuoti *et al.* 2015b for details): *(v)* we selected objects with `IMA_FLAGS` equal to zero in the *g, r* and *i* bands, i.e., sources that have not been flagged because located in proximity of saturated pixels, star haloes, image border or reflections, or within noisy areas, see de Jong *et al.* (2015). The *u* band is not considered in such selection since the masked regions relative to this band are less extended than in the other three KiDS bands.

The final KB consists of $15,180$ objects to be used as training set and $10,067$ for the test set.

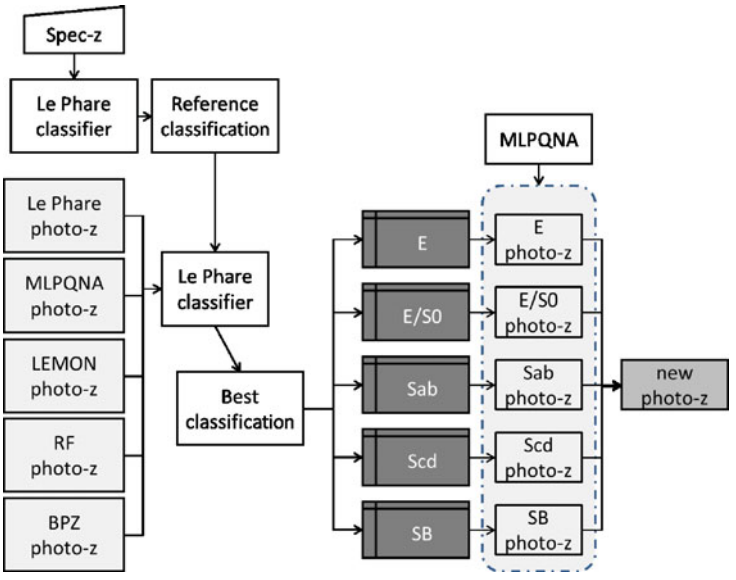

Figure 1. Workflow of the method implemented to combine SED fitting and ML models to improve the overall photo-z estimation quality. See text for details.

3. The methods

We chose three machine learning methods, among the ones which are publicly available in the DAta Mining & Exploration Web Application REsource or simply DAMEWARE (Brescia *et al.* 2014) web-based infrastructure: the Random Forest (RF; Breiman 2001), and two versions of the Multi Layer Perceptron with different optimization methods, i.e., the Quasi Newton Algorithm (Byrd *et al.* 1994) and the Levenberg-Marquardt rule (Nocedal & Wright 2006), respectively; furthermore we made use of a SED fitting method: Le Phare (Ilbert *et al.* 2006) and BPZ (Benitez 2000), a Bayesian photo-z estimation based on a template fitting method which is the last method involved in our experiments. The results were evaluated using only the objects of the blind test set by calculating the following set of standard statistical estimators for the quantity $\Delta z = (z_{\rm spec} - z_{\rm phot})/(1 + z_{\rm spec})$: *(i)* bias: defined as the mean value of the residuals Δz; *(ii)* σ: the standard deviation of the residuals; *(iii)* σ_{68}: the radius of the region that includes 68% of the residuals close to 0; *(iv)* NMAD: Normalized Median Absolute Deviation of the residuals, defined as $NMAD(\Delta z) = 1.48 \times Median(|\Delta z|)$; *(v)* fraction of outliers with $|\Delta z| > 0.15$.

4. Experiments

After a preliminary evaluation of the photometric redshifts, based on each of the five methods, by analyzing the results on the basis of the spectral type classification performed by Le Phare (i.e., the class of the template which shows the best fitting), we noticed that ML methods have a better performance, although strongly dependent from the spectral type itself. Therefore, we decide to exploit the capability of Le Phare to produce such spectral type classifications to train a specific regressor for each class. The workflow is described in Fig. 1. It goes without saying that the training of a specific regression model for each class can be effective only if the subdivision itself is as accurate as possible.

After having obtained the preliminary results, we started by creating a reference spectral type classification of data objects through Le Phare model. By bounding the fitting

Class	*Exptype*	*Datasize*	*bias*	σ	$NMAD$	*out.*(%)	$\sigma 68$
E	**hybrid**	**638**	**-0.0009**	**0.020**	**0.016**	**0.00**	**0.017**
E	standard	638	0.0130	0.029	0.022	0.31	0.028
E/S0	**hybrid**	**2858**	**-0.0005**	**0.016**	**0.012**	**0.10**	**0.012**
E/S0	standard	2858	-0.0059	0.022	0.014	0.31	0.014
Sab	**hybrid**	**1383**	**-0.0003**	**0.015**	**0.015**	**0.00**	**0.014**
Sab	standard	1383	-0.0032	0.018	0.016	0.00	0.016
Scd	**hybrid**	**3900**	**-0.0011**	**0.024**	**0.019**	**0.18**	**0.019**
Scd	standard	3900	0.0006	0.025	0.020	0.23	0.020
SB	**hybrid**	**1288**	**-0.0014**	**0.038**	**0.021**	**0.70**	**0.022**
SB	standard	1288	0.0027	0.038	0.022	0.85	0.023
ALL	**hybrid**	**10067**	**-0.0008**	**0.023**	**0.016**	**0.19**	**0.016**
ALL	standard	10067	-0.0007	0.026	0.018	0.31	0.018

Table 1. Photo-z estimation results based on MLPQNA model for each spectral type subset of the test set, classified by Le Phare by bounding the fit through the photo-z's predicted by RF model, which provided the best classification. The term *hybrid* refers to the results obtained by the workflow discussed here and based on the combined approach, while *standard* refers to the results obtained on the same objects but through the standard approach.

procedure with the spec-z's, Le Phare provided the templates with the best fit. In this way it was possible to assign a specific spectral type class to each object. Afterwards, by replacing the spec-z's with the photo-z's estimated by the preliminary experiment and by alternating the 5 photo-z estimates (one for each applied model) as redshift constraint for the fitting procedure, the Le Phare model was used to derive five different spectral type classifications for each object of the KB.

A normalized confusion matrix has been used to find the best classification, as the class of the best fit template derived from Le Phare which resulted from the experiment using the photometric redshifts produced by the random forest. By comparing the five matrices, the case of RF model presents the best behavior for all classes. Therefore, we considered as the best classification the one obtained by using the photo-z's provided by the RF model. We then subdivided the KB on the base of the five spectral type classes, thus obtaining five different subsets used to perform distinct training and blind test experiments, one for each individual class.

The final stage of the workflow consisted into the combination of the five subsets to produce the overall photo-z estimation, which was compared with the preliminary experiment in terms of the statistical estimators described in Sec. 3. The combined statistics was calculated on the whole datasets, after having gathered together all the objects of all classes and are reported in the last two rows of the Table 1. As with single classes, all the statistical estimators show an improvement in the combined approach case, with the exception of a slightly worst bias. As Table 1 shows, the proposed combined approach induces an estimation improvement for each class, as well as for the whole dataset.

5. Discussion and conclusions

In this work we described an original workflow designed to improve the photo-z estimation accuracy through a combined use of theoretical (SED fitting) and empirical (machine learning) methods. The data sample used for the analysis was extracted from the ESO KiDS DR2 photometric galaxy data, using a knowledge base derived from the SDSS and GAMA spectroscopic samples. For a catalog of about 25,000 galaxies with spectroscopic redshifts, we estimated photo-z's using five different methods: *(i)* Random

Forest; *(ii)* MLPQNA (Multi Layer Perceptron with the Quasi Newton learning rule); *(iii)* LEMON (Multi Layer Perceptron with the Levenberg-Marquardt learning rule); *(iv)* Le Phare SED fitting and *(v)* the bayesian model BPZ. The results obtained with the MLPQNA on the complete KiDS DR2 data have been discussed in Cavuoti *et al.* (2015a), and further details are provided there.

The spectral type classification provided by the SED fitting method allows to derive also for ML models the statistical errors as function of spectral type, thus leading to a more accurate characterization of the errors. Therefore, it is possible to assign a specific spectral type attribute to each object and to evaluate single class statistics. This fact by itself, can be used to derive a better characterization of the errors. Furthermore, as it has been shown, the combination of SED fitting and ML methods allows also to build specialized (i.e., expert) regression models for each spectral type class, thus refining the process of redshift estimation.

Although the spec-z's are in principle the most accurate information available to bound the SED fitting techniques, this would make impossible to produce a wide catalogue of photometric redshifts, that would also include objects not observed spectroscopically. Thus, it appears reasonable to identify the best solution by making use of predicted photo-z's to bound fitting, in order to obtain a reliable spectral type classification for the widest set of objects. This approach, having also the capability to use arbitrary ML and SED fitting methods, makes the proposed workflow widely usable in any survey project.

By looking at Table 1, our procedure shows clearly how the MLPQNA regression method benefits from the knowledge contribution provided by the combination of SED fitting (Le Phare in this case) and machine learning (RF in the best case. This allows to use a set of regression experts based on MLPQNA model, specialized to predict redshifts for objects belonging to specific spectral type classes, thus gaining in terms of a better photo-z estimation.

By analyzing the results of Table 1 in more detail, the improvement in photo-z quality is significant for all classes and for all statistical estimators. Only the two classes *Scd* and *SB* show a less evident improvement, since their residual distributions appear almost comparable in both experiment types, as confirmed by their very similar values of statistical parameters σ and σ_{68}. This leads to obtain a more accurate photo-z prediction by considering the whole test set.

The only apparent exception is the mean (column *bias* of Table 1), which suffers the effect of the alternation of positive and negative values in the *hybrid* case, that causes the algebraic sum to result slightly worse than the *standard* case (the effect occurs on the fourth decimal digit, see column *bias* of the last two rows of Table 1). This is not statistically relevant because the bias is one order of magnitude smaller than σ and σ_{68}, therefore negligible.

We note that in some cases, the *hybrid* approach leads to the almost complete disappearance of catastrophic outliers. This is the case, for instance of the E type galaxies. The reason is that for the elliptical galaxies the initial number of objects is lower than for the other spectral types in the KB. In the *standard* case, i.e., the standard training/test of the whole dataset, such small amount of E type representatives is mixed together with other more populated class objects, thus causing a lower capability of the method to learn their photometric/spectroscopic correlations. Instead, in the *hybrid* case, using the proposed workflow, the possibility to learn E type correlations through a regression expert increases the learning capabilities, thus improving the training performance and the resulting photo-z prediction accuracy.

The confusion matrices allow us to compare classification statistics. The most important statistical estimators are: *(i)* the *purity* or *precision*, defined as the ratio between

the number of correctly classified objects of a class (the block on the main diagonal for that class) and the number of objects predicted in that class (the sum of all blocks of the column for that class); *(ii)* the *completeness* or *recall*, defined as the ratio between the number of correctly classified objects in that class (the block on the main diagonal for that class) and the total number of (true) objects of that class originally present in the dataset (the sum of all blocks of the row for that class); *(iii)* the *contamination*, automatically defined as the reciprocal value of the *purity*.

Scd and *SB* spectral type classes are well classified by all methods. This is also confirmed by their statistics, since the *purity* is on average on all five cases around 88% for *Scd* and 87% for *SB*, with an averaged *completeness* of, respectively, 91% in the case of *Scd* and 82% for *SB*.

Moreover the three classifications based on the machine learning models maintain a good performance in the case of *E/S0* spectral type class, reaching on average a *purity* and a *completeness* of 89% for both estimators.

In the case of *Sab* class, only the RF-based classification is able to reach a sufficient degree of efficiency (78% of *purity* and 85% of *completeness*). In particular, for the two cases based on photo-z's predicted by SED fitting models, for the *Sab* class the BPZ-based results are slightly more *pure* than those based on Le Phare (68% vs 66%) but much less *complete* (49% vs 63%).

Finally, by analyzing the results on the *E* spectral type class, the classification performance is on average the worst case, since only the RF-based case is able to maintain a sufficient compromise between *purity* (77%) and *completeness* (63%). The classification based on Le Phare photo-z's reaches a 69% of completeness on the *E* class, but shows an evident high level of contamination between *E* and *E/S0*, thus reducing its purity to the 19%. We also note that the intrinsic major difficulty to separate *E* objects from *E/S0* class is due to the partial co-presence of both spectral types in the class *E/S0*, that may partially cause wrong evaluations by the classifier.

Furthermore, the fact that later Hubble types are less affected may be easily explained by considering that their templates are, on average, more homogeneous than for early type objects.

All the above considerations lead to the clear conclusion that the classification performed by Le Phare model and based on RF photo-z's achieves the best compromise between purity and completeness of all spectral type classes. Therefore, its spectral classification has been taken as reference throughout the further steps of the workflow.

At the final stage of the proposed workflow, the photo-z quality improvements obtained by the expert MLPQNA regression estimators on single spectral types of objects induce a reduction of σ from 0.026 to 0.023 and of σ_{68} from 0.018 to 0.016 for the overall test set, in addition to a more significant improvement for the E class (σ from 0.029 to 0.020 and of σ_{68} from 0.028 to 0.017). This is mostly due to the reduction of catastrophic outliers. This result, together with the generality of the workflow in terms of choice of the classification/regression methods, demonstrates the possibility to optimize the accuracy of photo-z estimation through the collaborative combination of theoretical and empirical methods.

Acknowledgments

CT is supported through an NWO-VICI grant (project number 639.043.308). MB and SC acknowledge financial contribution from the agreement ASI/INAF I/023/12/1. MB acknowledges the PRIN-INAF 2014 *Glittering kaleidoscopes in the sky: the multifaceted nature and role of Galaxy Clusters*.

References

Ahn, C. P., Alexandroff, R., Allende Prieto, C., *et al.* 2012, *ApJS*, 203, 21
Benitez, N., 2000, *ApJ*, 536, 571
Bertin, E. & Arnouts, S., 1996, *A&AS*, 117, 393
Breiman, L., 2001, *Machine Learning*, Springer Eds., 45, 1, 25-32
Brescia, M., Cavuoti, S., Longo, G., & De Stefano, V., 2015, *A&A*, 568, A126.
Brescia, M., Cavuoti, S., Longo, G., *et al.*, 2014, *PASP*, 126, 942, 743-797
Brescia, M., Cavuoti, S., D'Abrusco R., Mercurio, A., & Longo, G., 2013, *ApJ*, 772, 140
Byrd, R. H., Nocedal, J., & Schnabel, R. B., 1994, *Mathematical Programming*, 63, 129-156
Cavuoti, S., *et al.*, 2017, *MNRAS* 465 (2): 1959-1973.
Cavuoti, S., Brescia, M., Tortora, C., *et al.* 2015, *MNRAS*, 452, 3, 3100-3105
Cavuoti, S., *et al.*, 2015, *Experimental Astronomy*, Springer, Vol. 39, Issue 1, 45-71
Cavuoti, S., Brescia, M., Longo, G., & Mercurio, A., 2012, *A&A*, 546, 13
de Jong, J. T. A., Verdoes Kleijn, G. A., Boxhoorn, D. R., *et al.*, 2015, *A&A*, 582, A62
Fotopoulou, S., *et al.* submitted to MNRAS
Ilbert, O., Arnouts, S., McCracken, H. J., *et al.*, 2006, *A&A*, 457, 841
Liske, J., Baldry, I. K., Driver, S. P., *et al.*, 2015, *MNRAS*, 452, 2, 2087-2126
Nocedal, J. & Wright, S. J., 2006, *Numerical Optimization*, 2nd Edition. Springer
Tortora, C., La Barbera, F., Napolitano, N. R., *et al.*, 2016, *MNRAS*, 457, 3, 2845-2854