Astronomy and Computing 16 (2016) 34–40

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

Measuring photometric redshifts using galaxy images and Deep Neural Networks

B. Hoyle

Universitaets-Sternwarte, Fakultaet fuer Physik, Ludwig-Maximilians Universitaet Muenchen, Scheinerstr. 1, D-81679, Muenchen, Germany
Excellence Cluster Universe, Boltzmannstr. 2, D-85748, Garching, Germany

ARTICLE INFO

Article history:
Received 27 April 2015
Accepted 30 March 2016

Keywords:
Astronomy
Machine learning
Cosmology

ABSTRACT

We propose a new method to estimate the photometric redshift of galaxies by using the full galaxy image in each measured band. This method draws from the latest techniques and advances in machine learning, in particular Deep Neural Networks. We pass the entire multi-band galaxy image into the machine learning architecture to obtain a redshift estimate that is competitive, in terms of the measured point prediction metrics, with the best existing standard machine learning techniques. The standard techniques estimate redshifts using post-processed features, such as magnitudes and colours, which are extracted from the galaxy images and are deemed to be salient by the user. This new method removes the user from the photometric redshift estimation pipeline. However we do note that Deep Neural Networks require many orders of magnitude more computing resources than standard machine learning architectures, and as such are only tractable for making predictions on datasets of size $\leq$50k before implementing parallelisation techniques.

1. Introduction

To maximise the cosmological information available from current and upcoming large scale galaxy surveys, one requires robust distance estimates to many galaxies. The distances to galaxies are inferred by the distance-redshift relation which relates how the galaxy light is stretched due to the expansion of the Universe as it travels from the galaxy to our detectors. This stretching leads to an energy loss of the photon and a shift towards redder wavelengths, which is known as the redshift. The further away the galaxy is from us, the longer the light has been passing through the expanding Universe, and the more it becomes redshifted.

Obtaining very accurate spectroscopic redshifts, which measures the redshifted spectral absorption and emission lines, requires very long exposure times on dedicated spectrographs and is typically only performed for a small sub-sample of all galaxies. Conversely, the measurement of multi-band photometric properties of galaxies is much cheaper. The compromise is then to attempt to extract less accurate redshift information from photometrically measured properties, but applied to a much larger galaxy sample.

Photometric redshift estimates are obtained from either template fitting techniques, machine learning techniques, or some hybrid of the two for example using data augmentation (Hoyle et al., 2015). The template methods are parametric techniques and are constructed from templates of the Spectral Energy Distribution of the galaxies. Some templates encode our knowledge of stellar population models which result in predictions for the evolution of galaxy magnitudes and colours. The parametric encoding of the complex stellar physics coupled with the uncertainty of the parameters of the stellar population models, combine to produce redshift estimates which are little better than many non-parametric techniques. See e.g., Hildebrandt et al. (2010), Dahlen (2013) for an overview of different techniques. Unlike non-parametric and machine learning techniques, the aforementioned template methods do not rely on training samples of galaxies, which must be assumed to be representative of the final sample of galaxies for which redshift estimates are required. Other template methods are generated either completely from, or in combination with, empirical data, however these templates both require tuning, and also rely upon representative training samples.

When an unbiased training sample is available, machine learning methods offer an alternative to template methods to estimate galaxy redshifts. The 'machine architecture' determines how to best manipulate the photometric galaxy input properties (or 'features') to produce a machine learning redshift. The machine attempts to learn the most effective manipulations to minimise the difference between the spectroscopic redshift and the machine learning redshift of the training sample.

E-mail addresses: hoyleb@usm.uni-muenchen.de, benhoyle1212@gmail.com.

http://dx.doi.org/10.1016/j.ascom.2016.03.006
2213-1337/

The field of machine learning for photometric redshift analysis has been developing since Tagliaferri et al. (2003) used artificial Neural Networks (aNNs). A plethora of machine learning architectures, including tree based methods, have been applied to the problem of point prediction redshift estimation (Sánchez and Photometric, 2014) or to estimate the full redshift probability distribution function (Gerdes et al., 2010; Carrasco Kind and Brunner, 2013; Bonnett, 2015; Rau et al., 2015). Machine learning architectures have also had success in other fields of astronomy such as galaxy morphology identification, and star & quasar separation (Lahav, 1997; Yeche et al., 0000).

The use of Deep Neural Networks (hereafter DNN) as the machine learning architecture has only recently been applied to problems in astrophysics. For example Dieleman et al. (2015) taught a DNN to replicate the detailed morphological classifications obtained by the citizen scientists answering questions within the Galaxy Zoo 2 project (Willett et al., 2013) and obtained an accuracy of up to 99% on some classification questions, and (Hála, 2014) examined the problem of spectral classification from Sloan Digital Sky Survey (Ahn et al., 2014) (hereafter SDSS) spectra.

Within the standard machine learning approach the choice of which photometric input features to train the machine architecture, from the full list of possible photometric features, is still left to the discretion of the user. The current author recently performed an analysis of 'feature importance' for photometric redshifts, which uses machine learning techniques to determine which of the many possible photometric features produce the most predictive power (Hoyle et al., 2015). The technique described in this paper is the most extreme example of feature importance possible. We no longer need to impose our prior beliefs upon which derived photometric features produce the best redshift predictive power, or even measure the photometric properties. By passing the entire galaxy image into the Deep Neural Network machine learning framework we completely remove the user from the photometric redshift estimation process.

Furthermore in order to use either the template or standard machine learning techniques to estimate redshifts, the magnitudes, colours, and other properties of the galaxies must be measured. The analysis presented in this paper, which uses the full image of the galaxy partially removes this requirement. However we do still currently need the galaxy to have been detected so that we can generate a postage stamp image.

The outline of the paper is as follows. In Section 2 we describe the galaxy images and the pre-processing steps to prepare the images for the Deep Neural Networks. We then introduce both of the machine learning architectures in Section 3, and present the analysis and results in Section 4. We conclude and discuss in Section 5.

2. Galaxy data and images

The galaxy data in this study are drawn from the SDSS Data Release 10 (Ahn et al., 2014). The SDSS I–III uses a 2.4 m telescope at Apache Point Observatory in New Mexico and has CCD wide field photometry in 5 bands (Gunn et al., 2006; Smith et al., 2002), and an expansive spectroscopic follow up programme (Eisenstein and D.J., 2011) covering π steradians of the northern sky. The SDSS collaboration has obtained 2 million galaxy spectra using dual fibre-fed spectrographs. An automated photometric pipeline performs object classification to a magnitude of $r \approx 22$ and measures photometric properties of more than 100 million galaxies. The complete data sample, and many derived catalogs such as the photometric properties, and 5 band FITS images are publicly available through the SDSS website.[1]

[1] sdss.org.

We obtain 64,647 sets of images from the SDSS servers for a random selection of galaxies which are chosen to pass the following photometric selection criteria; the angular extent must be less than 30 arc seconds as measured by the 'Exponential' and 'de' Vaucouleurs' light profiles in the r band; and that each g, r, i, z has magnitudes greater than 0. We further select galaxies which pass the following spectroscopic selection criteria; the error on the spectroscopic redshift to be less than 0.1 and the spectroscopic redshift must be below 2. We check that none of the selected galaxies have images with missing or masked pixel values. In detail we run the MySQL query as shown in the appendix in the CasJobs server.

We choose to obtain the galaxy image FITS files in the following four photometric bands; g, r, i, z. This enables a closer resemblance to the bands available in other photometric surveys, for example the Dark Energy Survey (The Dark Energy Survey Collaboration, 0000). Each pixel in the FITS file has a resolution of 0.396 arc seconds and contains the measured flux which has been corrected for a range of observational and instrument effects such as flat fielding and sky subtraction, in order to be suitable for astronomical analysis. All pixel fluxes are converted to pixel magnitudes following Lupton et al. (1999). We apply a further extinction correction to account for galactic dust using the maps of Schlegel et al. (1998) which is available from the photoObjAll table in the CasJobs server. The extinction corrections are subtracted from the value of magnitude in each pixel in the corresponding FITS files. We choose to use FITS images of size 72×72 pixels, corresponding to 28.5 arc seconds on a side. We have explored the use of other image dimensions (32×32) but do not find improvement in the obtained results. The chosen image size is motivated by, and closely follows earlier work using SDSS images (Dieleman et al., 2015), and ensures that the training times are tractable.

In the top row of Fig. 1 we show RGB jpeg images of three example galaxies with the following mappings; g band magnitude $\rightarrow$ R, r band magnitude $\rightarrow$ G, and the i band magnitude $\rightarrow$ B. All pixel magnitudes are further rescaled across the entire layer to be integers within the range 0 to 255 for viewing purposes only. We further modify these base images to be more suitable for photometric redshift analysis by producing pixel colours from the pixel magnitudes and map pixel colours to each RGB layer pixel. We map the pixel colours i–z to the R layer pixels, r–i to the G layer pixels, and g–r to the B layer pixels. Finally we pass the r band pixel magnitude into an additional Alpha layer to produce an RGBA image. The r band magnitude is often used in this way to act as a pivot point which provides an overall normalisation to the input data. This may be useful during training and is common practice in photometric redshift analysis using neural networks (see e.g., Brescia et al., 2014). Examples of these modified images are shown in the second row of Fig. 1, but we show only the RGB values for viewing purposes.

During the analysis we scale all of the images, such that the maximum pixel value of 255 corresponds to the largest value across all training and test images in each of the RGBA layers separately. Likewise the minimum pixel value of 0 is set to be the smallest value in each layer across all images.

For a comparison with standard machine learning architectures we obtain model magnitudes measured by the SDSS photometric pipeline for each of the galaxies. To produce a fair comparison with the image analysis, we choose to use the de-reddened model magnitudes in the g, r, i, z bands and the size of each galaxy measured by the Petrosian radius in the r band.

We randomly shuffle and subdivide the 64,647 galaxies into training, cross-validation and test samples of size 33,167, 4047, and 27,433. In what follows we train the machine learning architectures on the training sample. We then vary the hyper-parameters of the machine learning architecture and retrain a