A hybrid neural network approach to estimate galaxy redshifts from multi-band photometric survey.

Rafael Duarte Coelho dos Santos,[1] Felipe Carvalho de Souza,[1]
Amita Muralikrishna,[1,2] and Walter Augusto dos Santos Junior[1]

[1]*INPE - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, Brazil*

[2]*IFSP - Instituto Federal de São Paulo, Campus São José dos Campos, SP, Brazil*

Abstract.
Machine learning methods have been used in cosmological studies to estimate variables that would be hard or costly to measure precisely, like, for example, estimating redshifts from photometric data. Previous work showed good results for estimating photometric redshifts using nonlinear regression based on an artificial neural network (Multilayer Perceptron). In this work we explore a hybrid neural network approach that uses a Self-Organizing Map to separate the original data into different groups, then applying the Multilayer Perceptron to each neuron on the Self-Organizing Map to obtain different regression models for each group. Preliminary results indicate that in some cases better results can be achieved, although the computational cost may be increased.

1. Introduction

Formation of the Universe is studied through mapping its observed objects into different categories (e.g. by separating galaxies from stars) and by observing their spatial distributions, shapes and positions. For this redshifts can be used to estimate distances between galaxies and our own.

Photometry techniques have been used as an alternative estimate galaxies' redshifts. Information can be estimated about more objects, although with less accuracy. In this paper we explore a two-level neural network approach to estimate photometry redshifts, or photo-z.

2. Spectroscopic and Photometric Redshifts

Galaxies' redshifts can be used as a distance measure, which can be used to calculate galaxies's spatial distribution. The more accurate way to obtain the redshift is through the galaxies' spectra, what is called spectroscopic redshift (Santos (2012)). However, spectroscopy has some limitations: one of them is that the observed objects should have enough brightness that they can be detected, which limits large surveys (LINEA (2017))

Photometry is a technique that was taken as an alternative to the spectroscopy in those cases when there is a need to collect data from many objects in a survey. It

provides a quick approximation of the SED (Spectral Energy Distribution), with less accuracy, and applicable to more objects. One way to do this is through regression using neural networks (Santos (2012)), which we explored in our previous work (Muralikrishna et al. (2017)).

3. Data

For our tests, we used photometric information about 100,000 galaxies obtained from the SDSS survey (Abolfathi et al. (2017)), which has the respective spectroscopy redshifts also recorded in a column so we can compare the results of our approach with the reference value.

The initial data was further processed: only records with guaranteed photometric quality were kept, and as in the previous work, records with redshift values lower than 0.01 and with redshift errors values lower than 0 (zero) and larger than 0.0005 were removed. The preprocessed sample had 97,214 records with six columns: one for each photometric band - u, g, r, i and z and the spectroscopic redshift value.

4. Neural Networks

In our previous work we used a Multilayer Perceptron (MLP), with a supervised training algorithm (Fausett (1994)) to estimate the spectroscopic redshift from the photometric data through regression. Results were promising, but we consider that multiple MLPs, applied to different subsets of the data, may yield better results for the estimation of the regression coefficients.

In this paper we explore a two-level approach: a Self-Organizing Map (Kohonen (2011)) which "splits" the data into different sets, and one MLP trained and applied to each set. Figures 1 and 2 illustrates our approach. Figure 1 shows the UGRIZ values for all the almost 100,000 records in a Parallel Coordinates plot (Inselberg (2009)) – we can notice the variation between the values in the data set.

Figure 2 shows a set of Parallel Coordinates plots based on a Kohonen Self-Organizing Map, which each subplot corresponding to a SOM neuron that "captured" data points that were similar to data in the same neuron and somehow different from data in other neurons. All data from Figure 1 is distributed in the cells/neurons in Figure 2, but the cells' values are less spread.

We then apply the same MLP neural network to each of the data's subset defined by the SOM. We theorise that this second-level neural network will converge quicker and yield better results than our previous approach.

For each neuron on the SOM, a MLP was created and trained, with six different architectures (5, 10, 30, 50, 70 and 90 neurons on the hidden layer). The metric used during the training to measure the accuracy of the results was the Normalized Median Absolute Deviation (σ_{NMAD}) (Molino et al. (2017)), calculated as shown in Equation 1):

$$\sigma_{NMAD} = 1.48 \times median\left(\frac{|\delta_z - median(\delta_z)|}{1 + z_s}\right) \quad (1)$$

where z_b is the photometric redshift, z_s is the spectroscopic redshift and $\delta_z = (z_b - z_s)$.

Each MLP was trained 10 times and the σ_{NMAD} stored for each run. The σ_{NMAD} are shown in boxplots, in the same visual arrangement as the SOM neurons, in Figure 3.

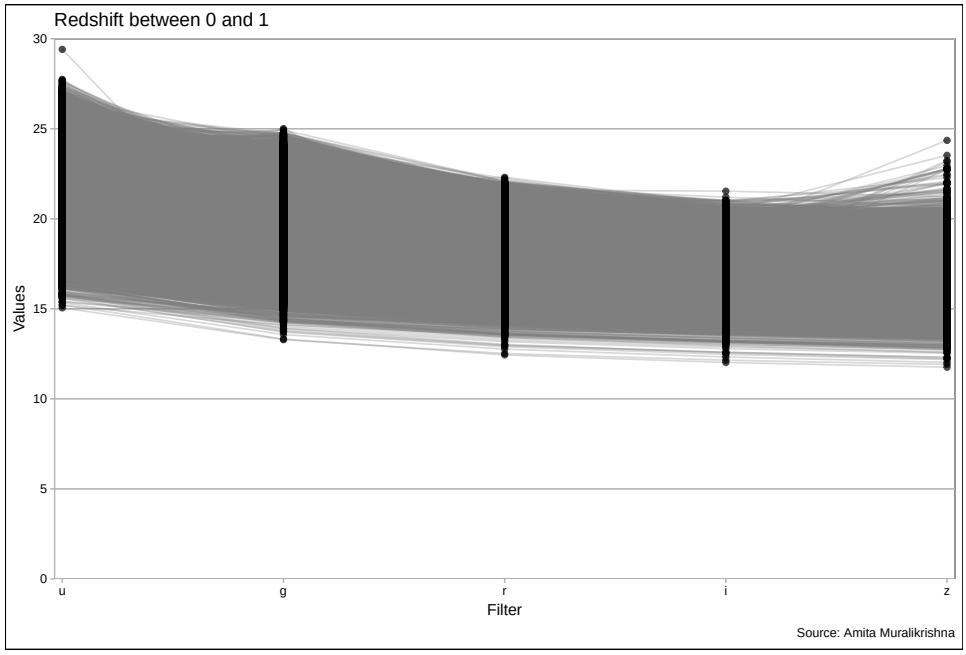

Figure 1. Around 97.000 UGRIZ data points.

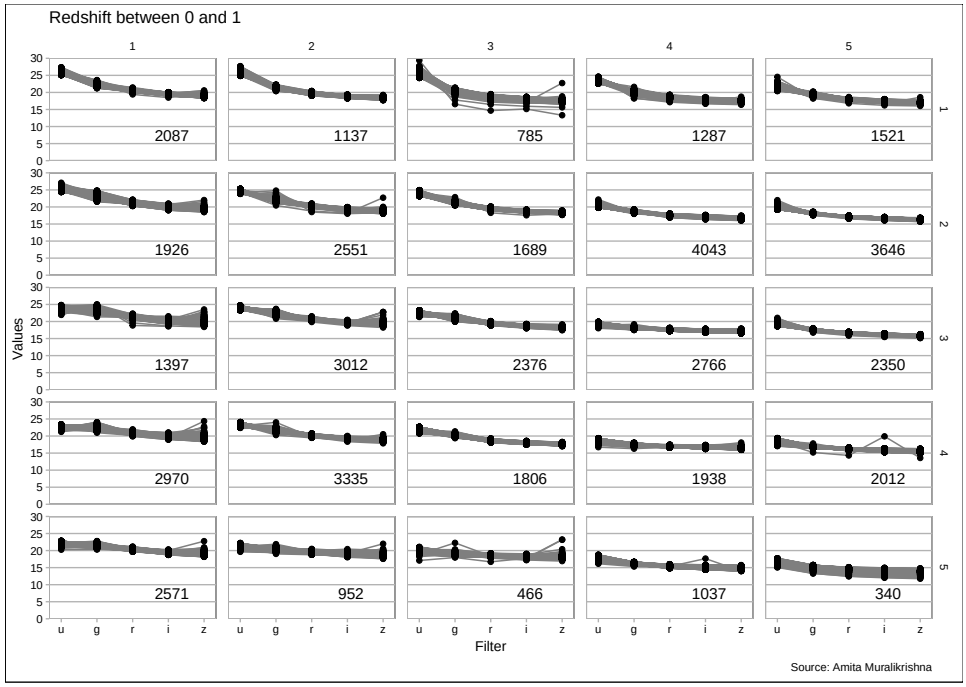

Figure 2. Around 97.000 UGRIZ data points, split in 5x5 subsets

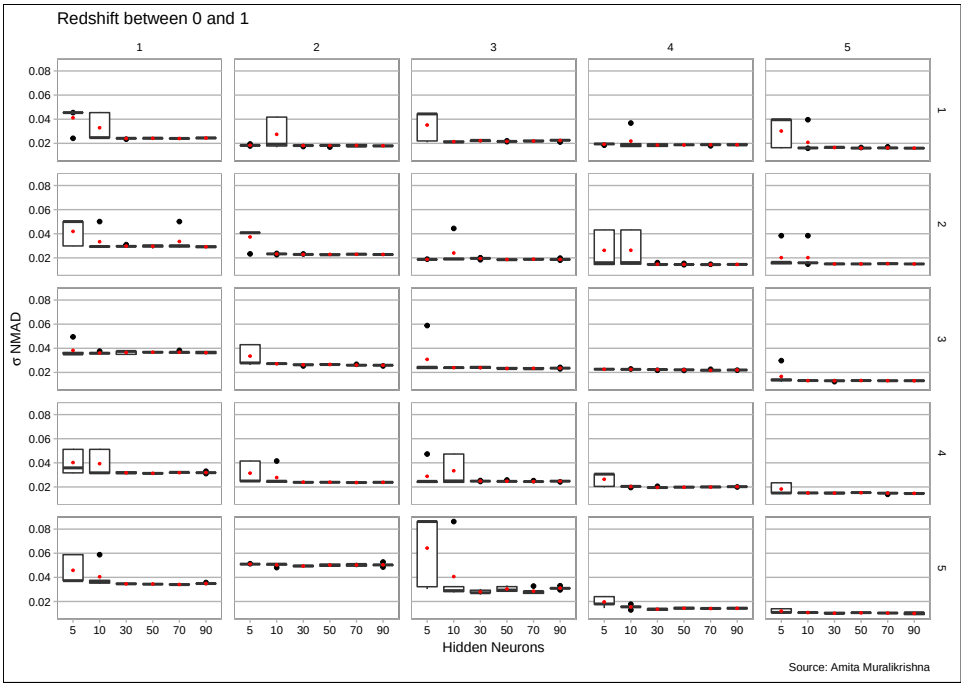

Figure 3. Boxplots for σ_{NMAD} for each of the SOM neurons

5. Evaluation and Conclusion

From Figure 3 we can see that many of the data subsets had smaller σ_{NMAD} values than we obtained in our previous work (0.022 with 90 neurons in the hidden layer) – some cells (eg. C3R5, C5R5) presented good results ($\sigma_{NMAD} = 0.009$) even with a smaller number of neurons in the second level MLP. On the other hands, some of the cells (ex. C2R5) presented worse results than the obtained before.

We consider that our approach was able to split the data set into "easy" and "hard" subsets. "Easy" ones can be used for estimation of photo-z, and "hard" ones can be further explored with different machine learning algorithms.

References

Abolfathi, B., et al. 2017, The Astrophysical Journal Supplement Series

Fausett, V. L. 1994, Fundamentals of Neural Networks: Architectures, Algorithms and Applications (PrenticeHall)

Inselberg, A. 2009, Parallel Coordinates – Visual Multidimensional Geometry and Its Applications (Springer)

Kohonen, T. 2011, Self-Organizing Maps (Springer), 3rd ed.

LINEA 2017, Laboratório Interinstitucional de e-Astronomia. URL `http://www.linea.gov.br/`

Molino, A., et al. 2017, Monthly Notices of the Royal Astronomical Society, 470, 95

Muralikrishna, A., Santos Junior, W., & Santos, R. D. C. 2017, in ADASS XXVII

Santos, W. A. 2012, Ph.D. thesis, Instituto de Astronomia, Geofísica e Ciências Atmosféricas (IAG), Departamento de Astronomia, Universidade de São Paulo (USP)