CAN SELF-ORGANIZING MAPS ACCURATELY PREDICT PHOTOMETRIC REDSHIFTS?

M.J. Way\textsuperscript{1,2,3}, C.D. Klose\textsuperscript{4}

\textit{Draft version January 6, 2012}

\section*{ABSTRACT}

We present an unsupervised machine learning approach that can be employed for estimating photometric redshifts. The proposed method is based on a vector quantization approach called Self-Organizing Mapping (SOM). A variety of photometrically derived input values were utilized from the Sloan Digital Sky Survey’s Main Galaxy Sample, Luminous Red Galaxy, and Quasar samples along with the PHAT0 data set from the PHoto-z Accuracy Testing project. Regression results obtained with this new approach were evaluated in terms of root mean square error (RMSE) to estimate the accuracy of the photometric redshift estimates. The results demonstrate competitive RMSE and outlier percentages when compared with several other popular approaches such as Artificial Neural Networks and Gaussian Process Regression. SOM RMSE results (using $\Delta z = z_{\text{phot}} - z_{\text{spec}}$) for the Main Galaxy Sample are 0.023, for the Luminous Red Galaxy sample 0.027, Quasars are 0.418, and PHAT0 synthetic data are 0.022. The results demonstrate that there are non-unique solutions for estimating SOM RMSEs. Further research is needed in order to find more robust estimation techniques using SOMs, but the results herein are a positive indication of their capabilities when compared with other well-known methods.

\textit{Subject headings:} methods: data analysis, methods: statistical, galaxies: distances and redshifts

\section*{1. INTRODUCTION}

There is a pressing need for accurate estimates of galaxy photometric redshifts (photo-$z$s) as demonstrated by the increasing number of papers on this topic and especially by recent attempts to objectively compare methods (e.g. Hildebrandt et al. 2010; Abdalla et al. 2011). The need for photo-$z$s will only increase as larger and deeper surveys such as Pan-STARRS$^5$(Kaiser 2004), LSST$^6$(Ivezic et al. 2008) and Euclid (Sorba & Savicki 2011) come on-line in the coming decade. The photometric-only surveys (Pan-STARRS, LSST) will have relatively small numbers of follow-up spectroscopic redshifts and will rely upon either template fitting methods such as Bayesian Photo-$z$’s (Benitez 2000), Le Phare (Ilbert et al. 2006), or training-set methods such as those discussed herein. The Euclid mission may include a slitless spectrograph offering far more training set galaxies.

A diverse set of regression techniques using training-set methods have been applied to the problem of estimating photometric redshifts in the past 10 years. These include Artificial Neural Networks (Firth et al. 2003; Tagliaferri et al. 2003; Ball et al. 2004; Collister & Lahav 2004; Vanzella et al. 2004), Decision Trees (Suchkov et al. 2005), Gaussian Process Regression (Way & Srivastava 2006; Foster et al. 2009; Way et al. 2009; Bonfield et al. 2010; Way 2011), Support Vector Machines (Wadadakar 2005), Ensemble Modeling (Way et al. 2009), Random Forests Carliies et al. (2008), and Kd-Trees (Csaibai et al. 2003) to name but a few.

On the other hand, even though Self-Organizing Maps (SOMs) have been used extensively in a number of other scientific fields (the paper that opened the field, Kohonen 1982, currently has over 2000 citations) they have been used sparingly thus far in Astronomy (e.g. Mahdi 2011; Naim et al. 1997; Way, Gazis & Scargle 2011), and only this year in estimating photometric redshifts (Geach 2011).

In this work we attempt to use SOMs to estimate photometric redshifts for several Sloan Digital Sky Survey (SDSS, York et al. 2000) derived catalogs of different galaxy types, including Quasars along with the PHAT0 data set of Hildebrandt et al. (2010). In Section 2 we describe the input data sets used, in Section 3 we give an overview of SOMs, and some conclusions in Section 4.

\section*{2. DATA}

Three different data sets derived from the SDSS Data Release Seven (DR7, Abazajian et al. 2009) were used. They include the Main Galaxy Sample (MGS, Strauss et al. 2002) the Luminous Red Galaxy Sample (LRG, Eisenstein et al. 2001), and the Quasar sample (QSO, Schneider et al. 2007). Data from the Galaxy Zoo\textsuperscript{7} (Lintott et al. 2008) Data Release 1 (Lintott et al. 2011) survey results were used to segregate galaxies as Spiral or Elliptical in the case of the MGS and LRG samples. Details of how this was done are given in Way (2011). Dereddened magnitudes ($u,g,r,i,z$) were used as inputs in all scenarios. The same SDSS photometric and redshift quality flags on the input variables were used as in Way (2011). In addition we used the simulation-based PHAT0 data set (see Hildebrandt et al. 2010) which was constructed to test a variety of different photo-$z$ estimation methods. The PHAT0 data set consists of 5 SDSS like filters ($u,g,r,i,z$) used on MEGACAM at CFHT (Boulade et al. 2009).

\begin{footnotesize}
\begin{itemize}
\item\textsuperscript{1} NASA Goddard Institute for Space Studies, 2880 Broadway, New York, New York 10025, USA
\item\textsuperscript{2} NASA Ames Research Center, Space Sciences Division, MS 247-9, Moffett Field, California 94035, USA
\item\textsuperscript{3} Department of Astronomy and Space Physics, Uppsala, Sweden
\item\textsuperscript{4} Think Geohazards, 205 Vernon Street, Suite A Roseville, CA 95678, USA
\item\textsuperscript{5} Panoramic Survey Telescope \& Rapid Response System
\item\textsuperscript{6} Large Synoptic Survey Telescope
\item\textsuperscript{7} http://www.galaxyzoo.org
\end{itemize}
\end{footnotesize}
procedure as feature vectors drawn from a training data set are presented to the input layer of the SOM network (Figure 1a). These feature vectors are also referred to as input vectors. Neurons of the Kohonen layer compete to see which neuron will be activated by the weight vectors that connect the input neurons and Kohonen neurons. In other words, the weight vectors identify which input vector can represented by a single Kohonen neuron. Hence, they are used to determine only one activated neuron in the Kohonen layer after the winner-takes-all principle (Figure 1b).
The SOM is considered as trained after learning, at which time the weights of the neurons have stored the inter-relations of all 5-dimensional $u, g, r, i, z$ feature vectors. Then, known spectroscopic redshift values for all input vectors of a test data set that are represented by a single Kohonen neuron are averaged (Fig.1b). The redshift mean value represents all 5-D $u, g, r, i, z$ vectors that are similar to the weight vector of the activated Kohonen neuron. The more Kohonen neurons there are the more precisely each input vector can be represented by a weight vector. However, the total number of Kohonen neurons are optimized for each data set (see Figure 2). A practical overview about the learning/training process is described by Klose (2006); Klose et al. (2008, 2010) and in much greater detail by Kohonen (2001).

After training, the $u, g, r, i, z$ input vectors of a test data set are presented to a trained SOM. At the end of a classification step, every Kohonen neuron approximates an input vector whereby similar/dissimilar input data were represented by neighboring/distant neurons. One neuron could even classify several input vectors, if these input vectors were very similar compared to the other given input vectors. Results from the photometric redshift approximations are then compared to known spectroscopic redshift data. Regression performance is estimated based on the root mean square error (RMSE) of the predicted photometric redshifts and the known spectroscopic redshifts ($\Delta z=z_{\text{phot}}-z_{\text{spec}}$). To reiterate, during the training phase, each Kohonen neuron identifies a certain number of galaxies that are characterized by similar $u, g, r, i, z$ intensities. Photometric redshift data were then averaged for these intensity values.

The SOM approximates the input feature space and maps it into an output space. The output space shows the SOM approximation as a 2-D map (Haykin 2009). Best results can be obtained with an optimization scheme such that the RMSE of the test data set is minimal as illustrated in Figure 2. Accuracy (e.g. RMSE) depends on the size of the Kohonen map. The number of neurons in the Kohonen map can be considered a regularization parameter ($\xi$) as shown in Figure 2.

Figure 2 shows that RMSE is high when the number of Kohonen neurons is too small ($\xi < 2000$) or too large ($\xi > 10000$) and hence that the 5-dimensional $u, g, r, i, z$-input space is underfit or overfit. Theoretically, a global minimum of the RMSE-curve might exist. However, the input feature space for the given photometric redshift problem shows a very rough RMSE-curve (Figure 2) with at least 2 local minima. In this case it is clear that SDSS redshift estimation tends to have several local minima, which makes it important to chose the right optimization method to determine the SOM network size. On the other hand, the smoother the RMSE-curve is the better gradient methods can be utilized. Evolution strategies or genetic programming could be applied to rougher curves with many local minima. This in turn can make it cumbersome to find fast back-propagation Artificial Neural Network (ANN) structures, especially when data sets are small.

Another advantage of SOMs in comparison to ANNs is that there is no need to optimize the structure of SOMs (e.g., number of hidden layers), since it is based on unsupervised learning.

Only the size of the Kohonen map needs to be optimized for each data set. SOMs also allow non-experts to visualize the redshift estimates in relation to the multi-dimensional input space. This eliminates the often criticized "black box" problem of ANNs. As mentioned previously, SOMs approximate the input feature space while ANNs typically separate them into sub-regions. Finally, SOMs are known to be powerful when very small data sets are available for training (see, Haykin 2009).

4. CONCLUSION
SOMs offer a competitive choice in terms of low RMSE, algorithm comprehension (also see Göpbert & Rosenstiel [1993]) and percentage of outliers. The final results are presented in Table 1 and plots for the LRG–ELL data set for the SOM, ANNz and GPR methods are shown in Figure 3.

As mentioned previously, obtaining the global minimum is important and, not surprisingly, can affect the results. Figure 2 shows the two local minima ($\xi=4100$ and 5100) listed for the LRG–ELL (Luminous Red Galaxies classified as Ellipticals by GalaxyZoo) data set in Table 1. Clearly there are a number of other $\xi$-values and the RMSE will be greatly affected by the choice as seen on the y-axis of Figure 2 for a given $\xi$-value. Given these facts, the roughness of the RMSE cost function in Figure 2 shows that traditional gradient based optimization strategies, e.g., deterministic annealing, might yield suboptimal solutions. Other methods, such as, genetic programming might find the “global” minimum much faster.

if a global minima exists with respect to the uncertainties of the RMSE.

During completion of this manuscript another paper using SOMs for classification and photometric estimation was released (Geach 2011). Our work differs in that we mostly focus on a wider variety of low-redshift samples drawn from the SDSS, while (Geach 2011) focuses more on the higher redshift samples akin to those used in Hildebrandt et al. (2010).

We have shown that SOMs are a powerful tool for estimating photometric redshifts and that with additional work they are sure to be useful in future surveys with limited numbers of follow-up spectroscopic redshifts.

M.J.W. would like to thank the Astrophysics Department at Uppsala University for their generous hospitality while part of this work was completed.

C.D.K. thanks Think Geohazards for providing the computational resources needed for estimating photometric redshifts via Self–Organizing Mapping.

Thanks goes to Joe Bredekamp and the NASA Applied Information Systems Research Program for support and encouragement.

Funding for the SDSS has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS Web site is http://www.sdss.org/

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are The University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, Los Alamos National Laboratory, the Max–Planck–Institute for Astronomy, the Max–Planck–Institute for Astrophysics, New Mexico State University, University of Pittsburgh, Princeton University, the United States Naval Observatory, and the University of Washington.

This research has made use of NASA’s Astrophysics Data System Bibliographic Services.

This research has also utilized the viewpoints (Gazis, Levit, & Way 2010) software package.

REFERENCES

Carilles, S., et al. 2008, ASPC, 394, 521

Klose, C.D. 2006, Computational Geosciences; 10(3), 265-277

Fig. 3.—Results from the three methods using SDSS u-g-r-i-z dereddened magnitudes as inputs for the SDSS DR7 Luminous Red Galaxies classified as ellipticals by the GalaxyZoo team. The bottom two plots show the SOM results for the two local minima described in Section 3 and shown in Figure 2.
Maldini, B. 2011, arXiv:1108.0514