SUPPORT VECTOR MACHINES FOR PHOTOMETRIC REDSHIFT ESTIMATION FROM BROADBAND PHOTOMETRY

Dan Wang, Yanxia Zhang, and Yongheng Zhao

[*]*National Astronomical Observatories, Chinese Academy of Sciences, China*
*Email:*dwang@lamost.org

ABSTRACT

Photometric redshifts have been regarded as efficient and effective measures for studying the statistical properties of galaxies and their evolution. In this paper, we introduce SVM_Light, a freely available software package using support vector machines (SVM) for photometric redshift estimation. This technique shows its superiorities in accuracy and efficiency. It can be applied to huge volumes of datasets, and its efficiency is acceptable. When a large representative training set is available, the results of this method are superior to the best ones obtained from template fitting. The method is used on a sample of 73,899 galaxies from the Sloan Digital Sky Survey Data Release 5. When applied to processed data sets, the RMS error in estimating redshifts is less than 0.03. The performances of various kernel functions and different parameter sets have been compared. Parameter selection and uniform data have also been discussed. Finally the strengths and weaknesses of the approach are summarized.

Keywords: Galaxies, Distances, Redshifts, Support vector machine

1 INTRODUCTION

With the large and deep sky survey projects being carried out, studying the formation and evolution of galaxies has rapidly become a crucial goal of mainstream observational cosmology. In order to achieve this purpose, redshift, which is one of the most crucial factors, must be obtained. Most commonly, the redshifts of galaxies are determined spectroscopically. However, for those large and faint sets of galaxies, spectra of galaxies are not easy to obtain. Rather than observing narrow spectral features of galaxy spectra, the photometric redshift technique concentrates on medium- or broad-band color features. Because the photometric redshift measurement relies only on colors, the approach can be extended to high redshifts (Stephen, 1995). Moreover, the photometric redshift method is also the only way to estimate redshift beyond the spectroscopic limit. The chief disadvantage of using photometric redshifts is that they are less precise compared to spectroscopic ones. However, for determining properties of large numbers of galaxies in a statistical way, the uncertainty of photometric redshift can be tolerated.

Two kinds of photometric redshift methods are available: the template fitting approach and the training set approach. In template fitting, according to the known redshift and galaxy type, some templates are constructed in advance by minimizing the standard χ^2 to fit the observed photometric data with a set of spectral templates. No spectroscopic information is required, and this method can be extended beyond the redshift limit. Commonly used templates are derived either from real observation, such as CWW (Coleman, Wu, & Weedman, 1980) or from population synthesis models (e.g. Bruzual & Charlot, 1993). Although it is easy to implement, the

accuracy of this approach strongly depends on the templates.

The essence of the training set approach is to derive a function between the redshift and photometric data by using a large and representative training set of galaxies for which both photometry results and redshifts are known and then use this function to estimate the remainder of the galaxies with unknown redshifts. In the past few years, a large number of training set methods have been developed and used (Way & Srivastava, 2006). Examples include linear or non-linear fitting (Brunner, Connolly, Szalay, Bershady, 1997; Wang, Bahcall, & Turner, 1998; Budavari, Szalay, Charlot, Seibert, Wyder, Arnouts, et al. 2005); support vector machines (Wadadekar, 2005); artificial neural networks (Firth, Lahav, & Somerville, 2003; Ball, Loveday, Fukugita, Nakamura, Okamura, & Brinkman, 2004; Collister & Lahav, 2004; Vanzella, Cristiani, Fontana, Nonino, Arnouts, & Giallongo, 2004); and nearest neighbors and kd-trees (Csabai, Budavari, Connolly, Szalay, Gyory, & Benitez, 2003).

In this paper, we use support vector machines to estimate photometric redshifts using photometric data from the Sloan Digital Sky Survey and the Two-Micron All Sky Survey. The outline of the paper is as follows: Section 2 introduces support vector machines; Section 3 illustrates the data used in the study, and Section 4 describes and discusses the results. Our conclusions are summarized in Section 5.

2 SUPPORT VECTOR MACHINES

Support Vector Machines (SVMs) were developed by Vapnik (1995) and has be applied to solve classification and regression problems. The regression problem solution of SVMs is achieved by using an alternative loss function, which is modified to include a distance measure. The task of SVMs usually involves training and testing sets that consist of data instances. Each instance in the training set contains one "target value" and several "attributes." The goal of SVMs is to produce a model that predicts the target value of data instances in the testing set, which are given only the attributes.

Given a training set of training pairs $(x_1, y_1),...,(x_l, y_l)$, $x_i \in R^n$, $y \in R$,

with a linear function,

$$f(x) = < \omega , x > + b ,$$

the optimal regression function is given by the minimum of the function

$$\Phi(\omega, \xi) = \frac{1}{2} \omega.\omega + C \sum_i (\xi_i^- + \xi_i^+) .$$

Using a quadratic loss function,

$$L_{quad}(f(x) - y) = (f(x) - y)^2 ,$$

the solution is given by

$$\max_{\alpha, \alpha^*} W(\alpha, \alpha^*) = \max_{\alpha, \alpha^*} - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) < x_i, x_j >$$

$$+ \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) y_i - \frac{1}{2C} \sum_{i=1}^{l} (\alpha_i^2 + (\alpha_i^2)^2).$$

The resultant optimization problem is

$$\min_{\beta} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\beta_i\beta_j\langle x_i, x_i\rangle - \sum_{i=1}^{l}\beta_i y_i + \frac{1}{2C}\sum_{i=1}^{l}\beta_i^2$$

with constraints

$$\sum_{i=1}^{l}\beta_i = 0.$$

To generalize to a non-linear regression, we replace the dot product with a kernel function. More information can be found in Steve's tutorial (1998).

Because of their excellent generalization performance, SVMs have been widely applied in the area of machine learning, such as handwritten digit recognition and face detection. In astronomy, SVMs have been applied for identifying red variables (Williams, Wozniak, Vestrand, & Gupta, 2004), clustering astronomical objects (Zhang & Zhao, 2004), and classifying AGNs from stars and normal galaxies (Zhang, Cui, & Zhao, 2002).

Several software packages of the SVM algorithm are accessible on the web. Regarding its robustness, ability to handle large amounts of data, and the regression time, we use SVM_Light in our case study. SVM_Light is a fast, optimized SVM algorithm, which is implemented in C language. It can deal with many thousands of support vectors, handle hundreds of thousands of training examples, and provide several standard kernel functions. The details about SVM_Light can be found at http://www.cs.cornell.edu/People/tj/svm_light/.

3 DATA

The data we used for this paper is from the Sloan Digital Sky Survey (SDSS) and the Two-Micron All Sky Survey (2MASS). The general information of SDSS and 2MASS is as follows.

3.1 Sloan Digital Sky Survey

The Sloan Digital Sky Survey (SDSS) (York, Adelman, Anderson, Annis, Bahcall, et al., 2000) is an astronomical survey project, which covers more than a quarter of the sky, to construct the first comprehensive digital map of the universe in 3D, using a dedicated 2.5-meter telescope located in Apache Point, New Mexico. In its first phase of operations, it has imaged 8,000 square degrees in five bandpasses (u, g, r, i, z) and measured more than 675,000 galaxies, 90,000 quasars, and 185,000 stars. In its second stage, SDSS will carry out three new surveys in different research areas: the nature of the universe, the origin of galaxies and quasars, and the formation an evolution of the Milky Way.

3.2 Two-Micron All Sky Survey

The Two-Micron All Sky Survey (2MASS) uses two highly-automated 1.3-m telescopes; one is in Mt. Hopkins, Arizona, and the other is located in CTIO, Chile. Each telescope has three-channels, which can observe the entire sky simultaneously at three near-infrared bands (j, h, and k). Jarrett et al. (2000) has more detailed information on the extended source catalog.

We select all galaxies of known redshifts from SDSS Data Release Five, and cross-match the data with 2MASS

extended point catalog within a search radius of 3 times the SDSS positional errors. After cross-matching, we generate more than 150,000 galaxies. Using these data, we include more restrictions. All data should satisfy the following criteria:

1) The spectroscopic redshift confidence must be equal to or greater than 0.95.

2) The redshift warning flag is 0.

3) Each magnitude should be inside its limit magnitude, namely u≤22.0, g≤22.2, r≤22.2, i≤21.3, and z≤20.5.

These qualifications produce a sample of 73899 galaxies. Table 1 shows the broadband filters and their wavelength range.

Table 1. Survey filters and characteristics

Bandpass	Survey	λ_{eff} (Å)	$\Delta\lambda$ (Å)
u	SDSS	3551	600
g	SDSS	4686	1400
r	SDSS	6165	1400
i	SDSS	7481	1500
z	SDSS	8931	1200
j	2MASS	12500	1620
h	2MASS	16500	2510
k	2MASS	21700	2620

4 RESULT AND DISCUSSION

When implementing SVMs, we adopt default soft margin (c) and radial basis function (RBF) kernel, modulate the kernel parameter (ϒ) to obtain the optimal result. We randomly divide the sample into two parts: two thirds for training and one third for testing. The training set has 50,000 samples and the test set has 23,899 samples. The different parameter sets are selected, including model magnitudes (u, g, r, i, z) from SDSS, dereddening magnitudes (u', g', r', i', z') from SDSS, magnitudes (j, h, k) from 2MASS, and colors composed of these magnitudes. Applying the training set to train the SVMs and the test set to test the regression estimator, we obtain the performances of various parameter sets. The RMS scatters of photometric redshift are listed in Table 2. As Table 2 shows, the performance of colors is better than that of magnitudes; the results with input pattern based on dereddening magnitudes are superior to those based on model magnitudes; the more parameters used, the higher the precision of the redshift estimation. The best RMS error reduces to 0.028.

If using artificial neural networks (ANNs), one should be familiar with the network architecture and make a decision about how many input nodes or hidden lays they have. The more complex networks available, the more accurate the results will be. However, SVMs may use different kernel functions instead of different ANN networks. As long as the appropriate kernel function and parameters are chosen, the RMS scatter will decrease significantly. In this study, the Gaussian function is adopted. Moreover, some classic problems, such as multi-local minima, curse of dimensionality, and overfitting in ANNs, seldom occur in SVMs.

Table 2. Photometric redshift prediction rms errors for different kernel parameters

Kernel parameter(ϒ)	Input parameters	σ
0.1	u, g, r, i, z…….	0.0303
1.1	u-g, g-r, r-i, i-z…….	0.0293
0.35	u-g, g-r, r-i, i-z, z-j, j-h, h-k…….	0.0286
0.3	u-g, g-r, r-i, i-z, z-j, j-h, h-k, r….....	0.0286
0.3	u'-g', g'-r', r'-i', i'-z', z'-j, j-h, h-k…….	0.0283
0.3	u'-g', g'-r', r'-i', i'-z', z'-j, j-h, h- k, r'…	0.0280

5 CONCLUSION

We utilize Support Vector Machines (SVMs) to estimate photometric redshifts using cross-matched data from SDSS DR5 and 2MASS. Photometric redshift accuracy produced by SVMs is comparable to that of ANN, as good as linear or quadratic regression, and clearly much better than template fitting. In appropriate situations, SVMs will be highly competitive tools for determining photometric redshifts in terms of speed and application. However, they do depend on the existence of a large and representative training sample. As a part of empirical photometric redshift estimations, it is impossible to extrapolate SVMs to a region that is not well sampled by the training set. Moreover, a potential solution to the problem of increasing the photometric redshift accuracy is to add additional input parameters, such as r-band 50% and 90% petrosian flux radii. This may improve the accuracy of redshift estimation about 15% (Wadadekar, 2005). Another approach to the problem is to choose a more appropriate kernel function. In the future, we will consider the feature selection/extraction methods in the process of parameter selection.

6 ACKNOWLEDGEMENTS

This paper has made use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, Cambridge University, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington. This paper is funded by National Natural Science Foundation of China under grant No.90412016, 60603057 and 10778623.

7 REFERENCES

Ball, N. M., Loveday, J., Fukugita, M., Nakamura, O., Okamura, S., & Brinkman, J. (2004) Galaxy types in the Sloan Digital Sky Survey using supervised artificial neural networks. *Monthly Notices of the Royal Astronomical Society 348,* 1038-1046.

Brunner, R., Connolly, A., Szalay, A., & Bershady, M. (1997) Toward More Precise Photometric Redshifts: Calibration Via CCD Photometry. *Astrophysical Journal Letters 482*, 21.

Bruzual A., G., Charlot, S. (1993) Spectral evolution of stellar population using isochrone synthesis. *Astrophysical Journal, 405*, 538-553.

Budavari, T., Szalay, A., Charlot, S., Seibert, M., Wyder, T., Arnouts, S., et al. (2005) The Ultraviolet Luminosity Function of GALEX Galaxies at Photometric Redshifts between 0.07 and 0.25. *The Astrophysical Journal, 619*, 31-34.

Coleman, G. D., Wu, C. C., & Weedman, D. W. (1980) Colors and magnitudes predicted for high redshift galaxies. *Astrophysical Journal Supplement Series 43*, 393-416.

Collister, A. & Lahav, O. (2004) ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. *The publications of the Astronomical Society of the Pacific 116*, 345- 351.

Csabai, I., Budavari, T., Connolly, A., Szalay, A. Gyory, Z., & Benitez, N. (2003) The Application of Photometric Redshifts to the SDSS Early Data Release. *The Astronomical Journal 125*, 580-592.

Firth, A., Lahav, O., & Somerville, R. (2003) Estimating photometric redshifts with artificial neural networks. *Monthly Notice of the Royal Astronomical Society 339*, 1195-1202.

Jarrett, T. H, Chester, T., Cutri, R., Schneider, S., Skrutskie, M., & Huchra, J. P. (2000) 2MASS Extended Source Catalog: Overview and Algorithms. *Astronomical Journal 119*, 2498-2531.

Gwyn, S. (1995) *Photometric Redshifts of Galaxies*, Master thesis, University of Victoria, BC, Canada

Steve, R. G. (1998) *Support Vector Machines for Classification and Regression*. Tutorial, Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science, University of Southampton, UK

Vanzella, E., Cristiani, S., Fontana, A., Nonino, M., Arnouts, S., & Giallongo, E. (2004) Photometric redshifts with the Multilayer Perceptron Neural Network: Application to the HDF-S and SDSS. *Astronomy and Astrophysics 423*, 761-776.

Wadadekar, Y. (2005) Estimating Photometric Redshifts Using Support Vector Machines. *The Astronomical Society of the Pacific 117*, 79-85.

Wang, Y., Bahcall, N., & Turner, E. (1998) A Catalog of Color-based Redshift Estimates for Z<~4 Galaxies in the Hubble Deep Field. *The Astronomical Journal 116*, 2081-2085.

Way, M. J. & Srivastava, A. N. (2006) Novel Methods for Predicting Photometric Redshifts from Broadband Photometry Using Virtual Sensors. *The Astrophysical Journa, 647*, 102-115.

Williams, S. J., Wozniak, P. R., Vestrand, W. T., & Gupta, V. (2004) Identifying Red Variables in the Northern Sky Variability Survey. *The Astronomical Journal 128,* 2965-2976.

York, D. G., Adelaman, J., Anderson, J., Anderson, S., Annis, J., Bahcall, N., et al. (2000) The Sloan Digital Sky Survey: Technical Summary. *Astronomical Journal 120*, 1579-1587.

Zhang, Y. & Zhao, Y. (2004) Automated clustering algorithms for classification of astronomical objects. *Astronomy and Astrophysics 422*, 1113-1121.

Zhang, Y., Cui, C., & Zhao, Y. (2002) Classification of AGNs from stars and normal galaxies by support vector machines. *Proceeding of the SPIE 4847*, 371-378.