RAA 2019 Vol. **19** No. **12**, 175(12pp) doi: 10.1088/1674–4527/19/12/175

http://www.raa-journal.org *http://iopscience.iop.org/raa*

***R**esearch in*
***A**stronomy and*
***A**strophysics*

A new strategy for estimating photometric redshifts of quasars

Yan-Xia Zhang[1], Jing-Yi Zhang[1,2], Xin Jin[1,2] and Yong-Heng Zhao[1]

[1] Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China; *zyx@bao.ac.cn*

[2] University of Chinese Academy of Sciences, Beijing 100049, China

Received 2019 February 26; accepted 2019 June 12

Abstract Based on the SDSS and SDSS-WISE quasar datasets, we put forward two schemes to estimate the photometric redshifts of quasars. Our schemes are based on the idea that the samples are firstly classified into subsamples by a classifier and then a photometric redshift estimation of different subsamples is performed by a regressor. Random Forest is adopted as the core algorithm of the classifiers, while Random Forest and kNN are applied as the key algorithms of regressors. The samples are divided into two subsamples and four subsamples, depending on the redshift distribution. The performances based on different samples, different algorithms and different schemes are compared. The experimental results indicate that the accuracy of photometric redshift estimation for the two schemes generally improves to some extent compared to the original scheme in terms of the percents in $\frac{|\Delta z|}{1+z_i} < 0.1$ and $\frac{|\Delta z|}{1+z_i} < 0.2$ and mean absolute error. Only given the running speed, kNN shows its superiority to Random Forest. The performance of Random Forest is a little better than or comparable to that of kNN with the two datasets. The accuracy based on the SDSS-WISE sample outperforms that based on the SDSS sample no matter by kNN or by Random Rorest. More information from more bands is considered and helpful to improve the accuracy of photometric redshift estimation. Evidently, it can be found that our strategy to estimate photometric redshift is applicable and may be applied to other datasets or other kinds of objects. Only talking about the percent in $\frac{|\Delta z|}{1+z_i} < 0.3$, there is still large room for further improvement in the photometric redshift estimation.

Key words: astronomical databases: catalogs — (galaxies:) quasars: general — methods: statistical — techniques: miscellaneous

1 INTRODUCTION

With the development of large photometric survey projects (e.g., 2MASS, GALEX, the Sloan Digital Sky Survey (SDSS), Pan-STARRS, LSST), we face a photometric data deluge, which is the best test bed for various algorithms. Among them, photometric redshift estimation is an important issue. Research on this aspect focuses on celestial objects, such as galaxies, quasars, supernovas, gamma-ray bursts and so on. The study of photometric redshifts is of great importance to the large scale structure of the Universe, the formation and evolution of galaxies, clustering of galaxies, distance measurement and so on. There are lots of works on the photometric redshift measurement of distant objects including quasars, and especially galaxies. Furthermore, a large number of algorithms and tools on photometric redshift estimation are in development. The algorithms are grouped into two kinds: template-fitting and machine learning — for instance, Bayesian method (Benítez 2000; Edmondson et al. 2006; Mortlock et al. 2012), color-redshift relation (Richards et al. 2001; Wu et al. 2004; Ball et al. 2007), k-Nearest Neighbors (kNN; Ball et al. 2007; Zhang et al. 2013), Gaussian process regression (Way & Srivastava 2006; Way et al. 2009; Bonfield et al. 2010), sparse Gaussian process regression (Almosallam et al. 2016b,a), Artificial Neural Networks (ANNs; Firth et al. 2003; Zhang et al. 2009; Yèche et al. 2010; Cavuoti et al. 2012; Brescia et al. 2013; Cavuoti et al. 2017), kernel regression (Wang et al. 2007), spectral connectivity analysis (SCA; Freeman et al. 2009), Random Forests (RFs) (Carliles et al. 2010; Schindler et al. 2017), ArborZ (Gerdes et al. 2010), the extreme deconvolution technique (Bovy et al. 2012), the Directional Neighborhood Fitting (DNF) algorithm (De Vicente et al. 2016), a hybrid technique (Beck et al. 2016), Self-Organizing Map (SOM; Way & Klose 2012; Carrasco Kind & Brunner 2014), Clustering aided Back propagation Neural network (CuBANz; Samui & Samui Pal 2017) and Support Vector Machine (SVM; Jones & Singal 2017; Schindler et al. 2017).

To improve the accuracy of photometric redshift estimation, researchers have considered new approaches or combined several methods. Wolf (2009) combined χ^2 template fits and empirical training sets into a single framework, applied it to the SDSS Data Release 5 (DR5) quasars, and improved the accuracy of photometric redshift estimation. Laurino et al. (2011) put forward Weak Gated Experts (WGE) to derive photometric redshifts of galaxies and quasars through a combination of data mining techniques. Gorecki et al. (2014) investigated different approaches and combined a template-fitting method and a neural network method for photometric redshifts of galaxies. Han et al. (2016) integrated kNN and SVM for photometric redshift estimation of quasars. Hoyle (2016) proposed Deep Neural Networks to estimate the photometric redshift of galaxies by using the full galaxy image in each measured band. Leistedt & Hogg (2017) presented a new method for inferring photometric redshifts in deep galaxy and quasar surveys, which combines the advantages of both machine learning methods and template fitting methods by building template spectral energy distributions (SEDs) directly from the spectroscopic training data. Wolf et al. (2017) investigated the photometric redshift performance of several empirical and template methods, and kernel-density estimation (KDE) was the best for their case. Jouvel et al. (2017) explored different techniques to reduce the photometric redshift outliers fraction with a comparison between the template fitting, neural networks and RF methods. Speagle & Eisenstein (2017a,b) derived photometric redshifts using fuzzy archetypes and SOMs, and demonstrated that statistical robustness and flexibility can be gained by combining template-fitting and machine-learning methods, and can provide useful insights into how astronomers may further exploit the color-redshift relation. Since large numbers of images are available, it is applicable to directly use image data and save time by preprocessing image data. D'Isanto & Polsterer (2018) probed deep learning to derive probabilistic photometric redshift directly from multi-band imaging data, rendering pre-classification of objects and feature extraction obsolete.

Although a large number of algorithms have been employed in this aspect, algorithms that perform well on galaxies may be not necessarily be applicable for quasars. Because the accuracy of the photometric redshift estimation of quasars is not too satisfactory, there is still large room for improvement. Therefore, we have designed a new strategy to estimate the photometric redshifts of quasars. The sample used for photometric redshift estimation is described in Section 2. Then, the adopted methods are briefly introduced in Section 3. Based on the SDSS and SDSS-WISE samples, the different schemes of photometric redshift estimation of quasars by kNN and RF are depicted in detail and compared in Section 4. The discussion is presented in Section 5. Finally we summarize the results of this paper in Section 6.

2 SAMPLE

The SDSS (York et al. 2000) has been one of the most successful surveys in the history of astronomy. In particular, it has created the most detailed three-dimensional maps of the Universe ever compiled, with deep multi-color images of one-third of the sky, and spectra for more than three million astronomical objects. We adopt the quasar sample from the Data Release 14 Quasar catalog (DR14Q) of SDSS-IV/eBOSS (Pâris et al. 2018). The DR14Q contains 526 356 unique quasars, of which 144 046 are new discoveries since the beginning of SDSS-IV. The catalog also includes previously spectroscopically-confirmed quasars from SDSS-I, II and III. Spectroscopic observations of quasars were performed over 9376 deg^2 for SDSS-I/II/III and are available over 2044 deg^2 for new SDSS-IV. Removing the records which contain default SDSS magnitudes, $z\mathrm{Warning} = -1$ and full magnitude errors larger than 5, the number of entries in the SDSS quasar sample reduces to 445 958. When further ruling out the records with default $W1$ and $W2$, the number of entries in the SDSS-WISE quasar sample amounts to 324 333. In this paper, we adopt AB magnitudes and convert the SDSS u-band and z-band magnitudes with $u_{\rm AB} = u' - 0.04$ mag and $z_{\rm AB} = z' + 0.02$. All magnitudes are corrected for Galactic extinction with the extinction values from DR14Q. The $W1$ (3.4 μm) and $W2$ (4.6 μm) of the *Wide-field Infrared Survey Explorer* (*WISE*; Mainzer et al. 2011) are directly obtained from DR14Q and converted to AB magnitudes using $W1_{\rm AB} = W1 + 2.699$ and $W2_{\rm AB} + 3.339$, and then extinction-corrected by the extinction coefficients $\alpha_{\rm W1}, \alpha_{\rm W2} = 0.189, 0.146$ with the extinction values from SDSS photometry. The AB magnitude conversion and extinction correction process are similar to the work of Schindler et al. (2017).

3 METHODS

The kNN method belongs to the lazy learning family, which delays its learning until prediction. Its principle of operation is to find the k training samples closest in distance to the new point, and predict the label from these. For the classification problem, the new point is labeled according to the majority of the k closest neighbors. For applications involving regression, the prediction is the average of the k closest neighbors. In general, the distance can be any metric measure, and standard Euclidean distance is the most common choice. To improve query speed, a fast indexing structure such as a Ball Tree or KD Tree is adopted.

RF (Breiman 2001) is based on bagging models built using the Random Tree method, in which classification trees are grown on a random subset of descriptors (e.g.,

Gao et al. 2009). Each tree in the ensemble is constructed from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. When splitting a node in the process of tree construction, the chosen split is no longer the best split among all of the features. However, this split is the best split among a random subset of the features. Based on this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree). However, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model. Therefore, RF uses the average to improve the predictive accuracy and control overfitting. Compared to Breiman (2001), the scikit-learn implementation of RF combines classifiers by averaging their probabilistic prediction, instead of letting each classifier vote for a single class.

For these two methods, we use KNeighborsRegressor, RandomForestRegressor and RandomForestClassifier from the Python module scikit-learn (Pedregosa et al. 2013).

4 PHOTOMETRIC REDSHIFT ESTIMATION

The redshift distribution of this sample is depicted in Figure 1. Because the quasar colors change with redshift and the dominating features appear in different bands with different redshifts, we divide the quasar sample into two classes and four classes according to the redshift range. The two classes are one with $0 < z \leq 2.2$ and the other with $2.2 < z$. The four classes are "vlowz" with $0 < z \leq 1.5$, "lowz" with $1.5 < z \leq 2.2$, "midz" with $2.2 < z \leq 3.5$ and "highz" with $3.5 < z$ similar to Schindler et al. (2017). At the first break of $z = 1.5$, the Lyman-alpha (Lyα) emission line stays blueward of the u-band and the CIV emission line still remains in the g-band. Because the second break is at $z = 2.2$, the Lyα emission line is just leaving the u-band. At $z = 3.5$, a strong flux decreases in the u-band while the Lyα forest absorbs flux blueward of the Lyα line. We apply these two classes and four classes to label the SDSS and SDSS-WISE quasar samples for the classification problem. In the following experiments, for the SDSS quasar sample, r, $u-g$, $g-r$, $r-i$, $i-z$ are taken as the input pattern, while for SDSS-WISE quasar sample, r, $u-g$, $g-r$, $r-i$, $i-z$, $z-W1$, $W1-W2$ are adopted. The whole quasar samples from SDSS and SDSS-WISE are randomly separated into two-thirds for training and one-third for testing.

The problem of photometric redshift estimation belongs to the regression task of data mining. Thus, the algorithm's fit for regression can be applied for photometric redshift estimation. When the sample is specified, a choice of approaches is needed. Comparison of different regressors depends on different regression metrics, such as the residual between the spectroscopic and photometric redshifts, $\Delta z = z_{\rm spec} - z_{\rm photo}$, and the mean absolute error σ. Another metric to determine the goodness of photometric redshift estimation is the fraction of test samples that satisfy $\mid \triangle z \mid = |z_i - \widehat{z_i}| < e$ (Schindler et al. 2017 and references therein).

The definition of mean absolute error σ is as follows

$$\sigma = \frac{1}{n}\sum_{i=0}^{n-1} |z_i - \widehat{z_i}|\,, \tag{1}$$

where z_i is the true redshift, $\widehat{z_i}$ is the the predicted redshift value and n is the sample size.

The fraction of test samples that satisfies $\mid \triangle z \mid = |z_i - \widehat{z_i}| < e$ is usually used to evaluate the redshift estimation, where e is a given residual threshold,

$$f_{|\triangle z|<e} = \frac{N(|z_i - \widehat{z_i}| < e)}{N_{\rm total}}\,. \tag{2}$$

The typical values of e are 0.1, 0.2 and 0.3. However, the redshift normalized residuals are often adopted,

$$\delta_e = \frac{N(|z_i - \widehat{z_i}| < e(1+z_i))}{N_{\rm total}}\,. \tag{3}$$

Once the methods have been chosen, the next important task is to determine an algorithm's hyperparameters, which indicate how the machine learns. For kNN, the model parameter is only k when taking Euclidean distance as metric measure and KD Tree as index. But for RF, main parameters contain the maximum depth of individual trees (max depth) and the number of trees in the forest (n estimators). The goal is to find the optimal model parameters which optimize the algorithm's performance. In reality we do not know these values in advance. Therefore, a grid search is performed with K-fold cross-validation, which means that the training sample is split into K subsamples such that one subsample is left to estimate the algorithm's performance while the remaining subsamples are utilized to train the algorithm and construct the classifier/regressor. This process is done K times and finally the average performance is kept. The entire process is repeated for every combination of hyperparameters in the grid space and values that optimize the performance are output. The grid for kNN has k values from 10 to 30. The grid for RF has the hyperparameters: $n\_{\rm estimators} = [50, 100, 200, 300]$ and $\max\_{\rm depth} = [15, 20, 25]$ (12 combinations).

4.1 Photometric Redshift Estimation with One Sample

For the SDSS sample and the SDSS-WISE sample, the two samples are randomly divided into two parts: two-thirds as training sets and one-third as test sets. All of the model constructions are performed by 10-fold cross-validation on the full training sets while the other test sets are applied to test the regressors (kNN and RF). Their performance,

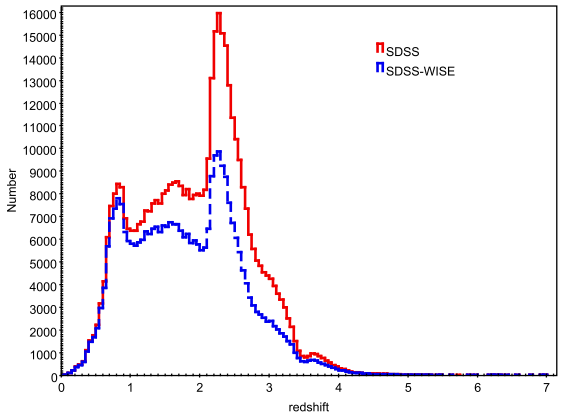

Fig. 1 Spectroscopic redshift distribution of the SDSS quasars (the *red solid line histogram*) and the SDSS-WISE quasars (the *blue dashed line histogram*). The bin size is $\Delta z = 0.05$.

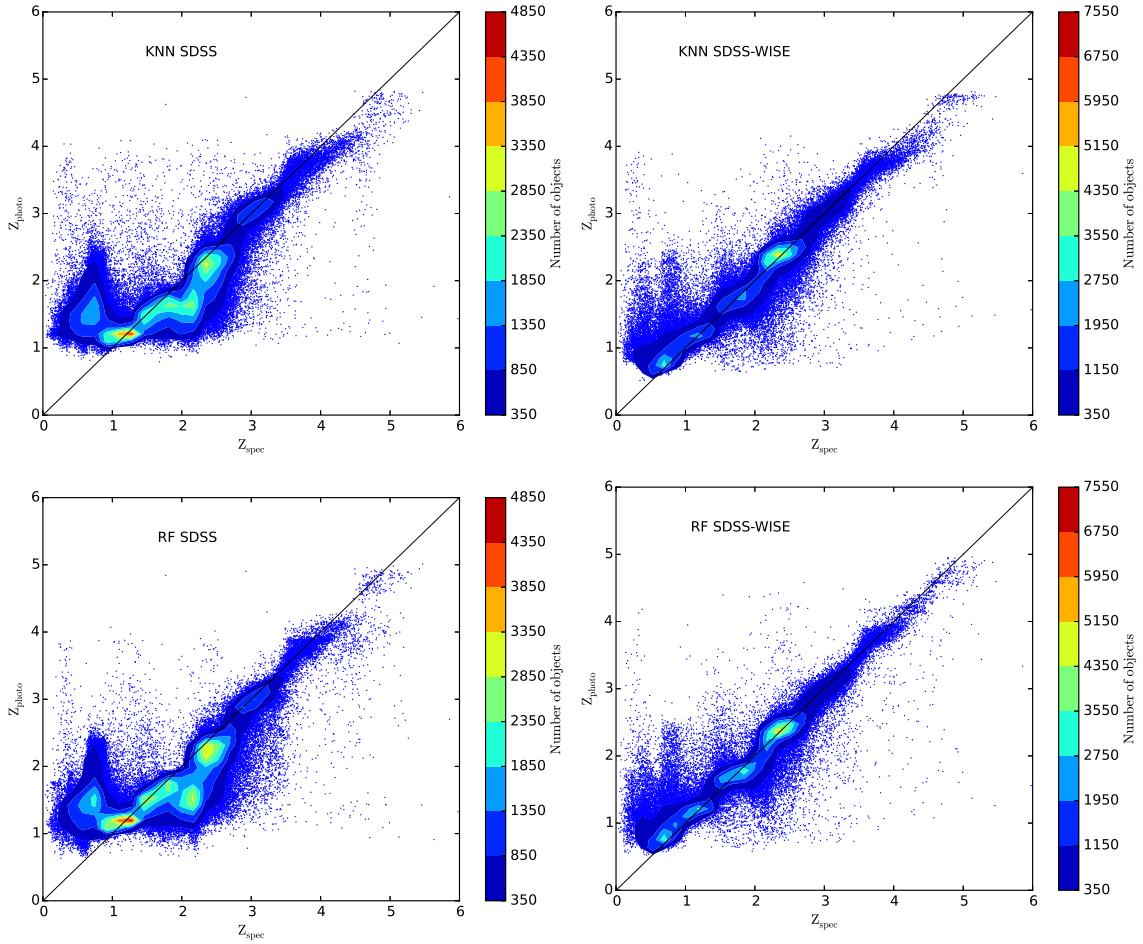

Fig. 2 Predicted photometric redshifts vs. spectroscopic redshifts. The *color bars* signify the number of objects per rectangular bin. The upper-left panel is based on the SDSS test sample by kNN; the upper-right panel is based on the SDSS-WISE test sample by kNN; the lower-left panel is based on the SDSS test sample by RF; the lower-right panel is based on the SDSS-WISE test sample by RF.

optimal model parameters and running time of the two algorithms for model construction and predicting photometric redshifts of quasars are written in Table 1. A comparison of photometric redshift estimation with spectroscopic redshifts by different methods is displayed in Figure 2. As listed in Table 1, for SDSS sample with kNN, the percents ($\delta_{0.1}$, $\delta_{0.2}$ and $\delta_{0.3}$) in different $\frac{|\Delta z|}{1+z_i}$ intervals and the mean absolute error σ are 62.53%, 80.13%, 87.17% and 0.3326, respectively; for the SDSS-WISE sample with kNN, $\delta_{0.1}$, $\delta_{0.2}$ and $\delta_{0.3}$ and σ are 79.40%, 91.37%, 95.28% and

Table 1 Performance of Photometric Redshift Estimation for Different Datasets with kNN and RF

Data Set	Algorithm	Model Parameters	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ	Time (s)
SDSS	kNN	$k = 30$	62.53	80.13	87.17	0.3326	242
SDSS-WISE	kNN	$k = 30$	79.40	91.37	95.28	0.1931	420
SDSS	RF	$n\_\text{estimators} = 300$ $\text{max\_depth} = 15$	63.34	80.48	87.34	0.3271	37628
SDSS-WISE	RF	$n\_\text{estimators} = 300$ $\text{max\_depth} = 20$	79.87	91.37	95.23	0.1907	36762

Table 2 Performance of photometric redshift estimation for different datasets with kNN after classifying one sample into two subsamples by RF.

Data Set (Test set)	Algorithm	Model Parameters	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ
SDSS (T1)	RF_kNN	$k = 30$	54.07	71.79	84.44	0.3574
SDSS (T2)	RF_kNN	$k = 30$	84.58	89.31	90.12	0.2856
SDSS (T1+T2)			67.11	79.28	86.87	0.3267
SDSS-WISE (T1)	RF_kNN	$k = 30$	74.80	89.37	94.73	0.2037
SDSS-WISE (T2)	RF_kNN	$k = 20$	93.05	96.40	97.00	0.1710
SDSS-WISE (T1+T2)			80.97	91.75	95.50	0.1926

Table 3 Performance of photometric redshift estimation for different datasets with RF after classifying one sample into two subsamples by RF.

Data Set (Test set)	Algorithm	Model Parameters	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ
SDSS (T1)	RF_RF	$n\_\text{estimators} = 300$ $\text{max\_depth} = 15$	55.08	72.07	84.36	0.3550
SDSS (T2)	RF_RF	$n\_\text{estimators} = 300$ $\text{max\_depth} = 15$	84.77	89.55	90.31	0.2810
SDSS (T1+T2)			67.74	79.52	86.90	0.3235
SDSS-WISE (T1)	RF_RF	$n\_\text{estimators} = 300$ $\text{max\_depth} = 15$	75.77	89.55	94.52	0.2022
SDSS-WISE (T2)	RF_RF	$n\_\text{estimators} = 300$ $\text{max\_depth} = 20$	93.01	96.40	96.97	0.1660
SDSS-WISE (T1+T2)			81.60	91.87	95.35	0.1900

0.1931, separately; for the SDSS sample with RF, they are respectively 63.34%, 80.48%, 87.34% and 0.3271; for the SDSS-WISE sample with RF, they are respectively 79.87%, 91.37%, 95.23% and 0.1907. For kNN, the running time of model construction and prediction is 242 s with the SDSS sample and 420 s with the SDSS-WISE sample; while for RF, the running time is 37 628 s and 36 762 s, respectively. No matter for kNN or for RF, the accuracy of photometric redshift estimation improves apparently with both optical and infrared information compared to only the optical information. For the SDSS sample, the performance of RF is a little superior to that of kNN. Meanwhile, for the SDSS-WISE sample, the performance of RF is better than that of kNN except for $\delta_{0.2}$ and $\delta_{0.3}$, although their accuracy is comparable. If only considering speed, kNN shows its superiority.

4.2 Photometric Redshift Estimation with Two Subsamples

Considering the redshift distribution due to the physical properties of quasars, quasars may be separated into different groups. To improve the performance of photometric redshift estimation, we put forward a scheme of first classification and second regression for photometric redshift estimation, specifically any new source is classified by a classifier in advance and subsequently its photometric redshift is predicted by a regressor.

For the detailed steps of photometric redshift estimation with two subsamples, see Figure 3. First, the quasar samples of SDSS and SDSS-WISE are divided into two subsamples: one with $0 < z \leq 2.2$ and the other with $2.2 < z$. For the two subsamples, they are randomly segmented into two parts: two-thirds for training (A with redshift from 0 to 2.2 and B with redshift from 2.2 to 6) and one-third for testing (T1 with redshift from 0 to 2.2 and T2 with redshift from 2.2 to 6). With the training sets A and B, the classifier is created by 10-fold cross-validation. The testing sets T1 and T2 are applied as inputs for the classifier and then classified as T1A and T2A with redshift from 0 to 2.2 and T1B and T2B with redshift from 2.2 to 6. Second, the samples A and B are used as training sets to train regressors and represented as regressor_A and regressor_B, respectively. The testing samples T1A and T2A are tested by regressor_A while the testing samples T1B and

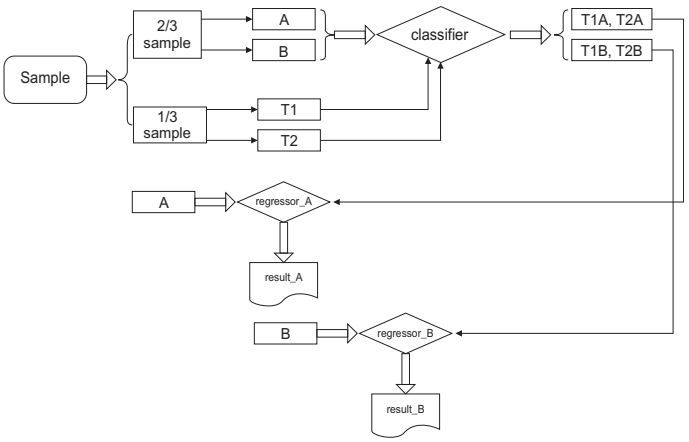

Fig. 3 Flow chart of photometric redshift estimation based on two subsamples.

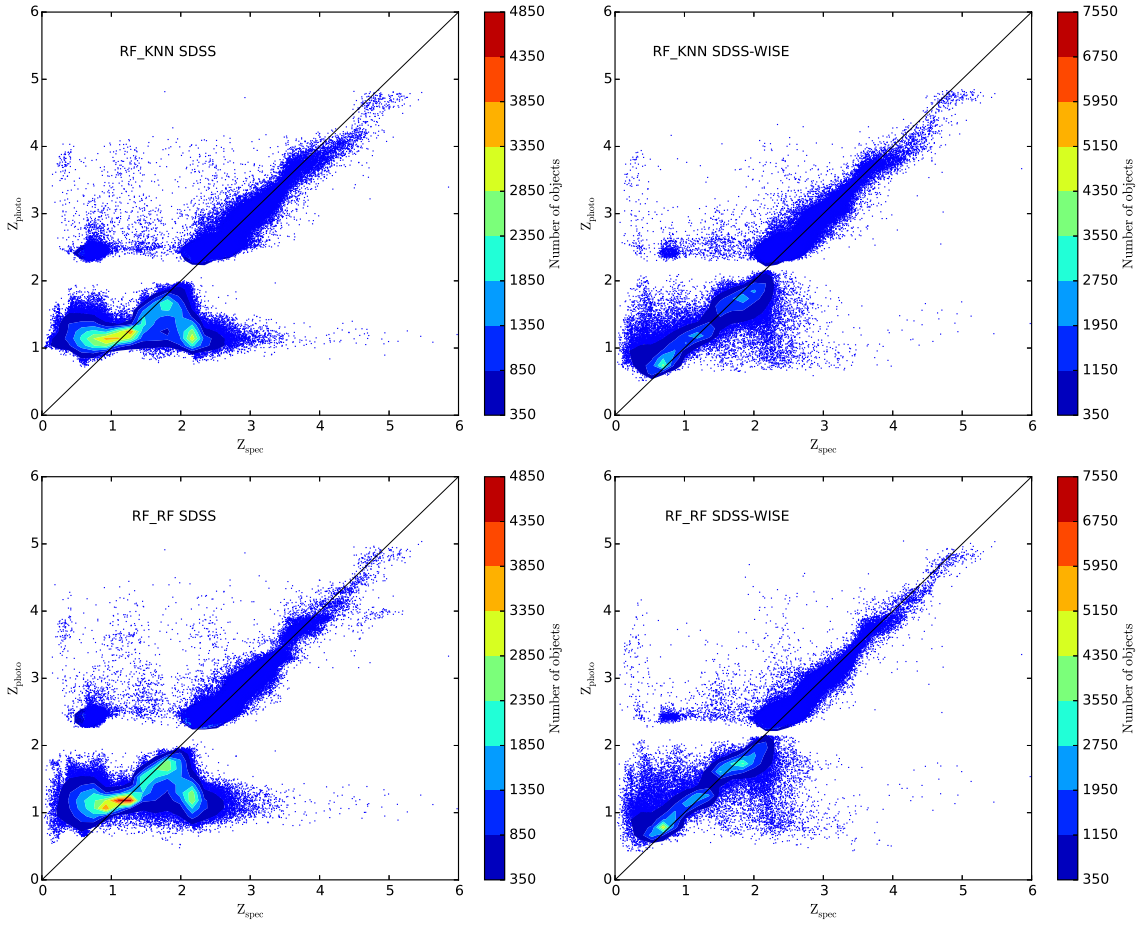

Fig. 4 Predicted photometric redshifts vs. spectroscopic redshifts for two subsamples. The *color bars* signify the number of objects per rectangular bin. The upper-left panel is based on SDSS test sample by kNN; the upper-right panel is based on SDSS-WISE test sample by kNN; the lower-left panel is based on SDSS test sample by RF; the lower-right panel is based on SDSS-WISE test sample by RF.

T2B are tested by regressor_B. Finally, the predicted results are obtained as result_A and result_B.

In brief, the core algorithm of the classifier is RF, while the regressors adopt kNN and RF. For the SDSS and SDSS-WISE samples, they are randomly segmented into two-thirds for training and one-third for testing. The RF classifier is constructed by 10-fold validation with the full training set. The regressors (kNN and RF) are also built by 10-fold validation with the full training set. For convenience, RF is utilized for classification and kNN is for regression, RF_KNN for short; RF is employed both for classification and regression, RF_RF for short. For

the SDSS sample with RF_KNN, the optimal parameters of the RF classifier are $n\_\mathrm{estimators} = 100$ and $\mathrm{max\_depth} = 15$; while for the SDSS-WISE sample with RF_KNN, $n\_\mathrm{estimators} = 300$ and $\mathrm{max\_depth} = 25$. For the SDSS sample with RF_RF, optimal parameters of the RF classifier are $n\_\mathrm{estimators} = 300$ and $\mathrm{max\_depth} = 15$; while for the SDSS-WISE sample with RF_RF, $n\_\mathrm{estimators} = 300$ and $\mathrm{max\_depth} = 20$. For different subgroups, the performance of photometric redshift estimation for the SDSS and SDSS-WISE samples with kNN after classifying one sample into two subsamples by RF is indicated in Table 2, while the performance with RF is shown in Table 3. Comparison of photometric redshift estimation with spectroscopic redshifts by different methods is indicated in Figure 4. No matter if implementing RF_KNN or RF_RF, adding the infrared information is helpful to improve the accuracy of photometric redshift estimation, and the performance based on T2 is better than that based on T1. Comparing the results in Table 2 with those in Table 3, it is found that the performance of RF_KNN is a little inferior to that of RF_RF except for $\delta_{0.3}$ of SDSS T1 and SDSS-WISE T1, $\delta_{0.3}$ of SDSS-WISE T1+T2 and three δ values of SDSS-WISE T2. Considering the entire test sets (SDSS T1+T2 and SDSS-WISE T1+T2), RF_RF manifests slightly better performance than RF_KNN.

4.3 Photometric Redshift Estimation with Four Subsamples

Similar to Section 4.2, we put forward another scheme of first classification and second regression for photometric redshift estimation. To be different, the quasar samples of SDSS and SDSS-WISE are separated into four subsamples: "vlowz" with $0 < z \leq 1.5$, "lowz" with $1.5 < z \leq 2.2$, "midz" with $2.2 < z \leq 3.5$ and "highz" with $3.5 < z$. These four subsamples are randomly broken up into two parts: two-thirds for training (a with redshift from 0 to 1.5, b with redshift from 1.5 to 2.2, c with redshift from 2.2 to 3.5 and d with redshift from 3.5 to 6) and one-third for testing ($t1$ with redshift from 0 to 1.5, $t2$ with redshift from 1.5 to 2.2, $t3$ with redshift from 2.2 to 3.5 and $t4$ with redshift from 3.5 to 6). Based on training sets a, b, c and d, the classifier is created. Then the classifier separates testing sets $t1$, $t2$, $t3$ and $t4$ into $t1a$, $t2a$, $t3a$ and $t4a$ with redshift from 0 to 1.5, $t1b$, $t2b$, $t3b$ and $t4b$ with redshift from 1.5 to 2.2, $t1a$, $t2c$, $t3c$ and $t4c$ with redshift from 2.2 to 3.5, and $t1d$, $t2d$, $t3d$ and $t4d$ with redshift from 3.5 to 6.0. Next, samples a, b, c and d are used for training regressors, and four regressors are obtained, represented as regressor_a, regressor_b, regressor_c and regressor_d, respectively. The testing samples $t1a$, $t2a$, $t3a$ and $t4a$ are tested by regressor_a, $t1b$, $t2b$, $t3b$ and $t4b$ by regressor_b, $t1c$, $t2c$, $t3c$ and $t4c$ by regressor_c, and $t1d$, $t2d$, $t3d$ and $t4d$ by regressor_d. In the end, the predicted results are obtained as result_a, result_b, result_c and result_d, respectively. The detailed steps of photometric redshift estimation with four subsamples are shown in Figure 5.

In the whole process, RF is still adopted as the classification algorithm, while RF and kNN are utilized as the regression algorithms. The performance of photometric redshift estimation for the SDSS and SDSS-WISE samples with kNN after classifying one sample into four subsamples by RF is indicated in Table 4, while the performance with RF is shown in Table 5. Comparison of photometric redshift estimation with spectroscopic redshifts by different methods is indicated in Figure 6. For the SDSS sample with RF_KNN, the optimal parameters of RF classifier are $n\_\mathrm{estimators} = 300$ and $\mathrm{max\_depth} = 15$; while for the SDSS-WISE sample with RF_KNN, $n\_\mathrm{estimators} = 200$ and $\mathrm{max\_depth} = 20$. For the SDSS sample with RF_RF, the optimal parameters of RF classifier are $n\_\mathrm{estimators} = 300$ and $\mathrm{max\_depth} = 15$; while for the SDSS-WISE sample with RF_RF, $n\_\mathrm{estimators} = 300$ and $\mathrm{max\_depth} = 25$. To compare the results in Table 4 with those in Table 5, given that there are only three δ values for the test sets (SDSS t4 and SDSS-WISE t4), the accuracy of RF_KNN is better than RF_RF, while only considering σ, RF_RF is superior to RF_KNN; given that there are only $\delta_{0.2}$ and $\delta_{0.3}$ for the test set SDSS t1, RF_KNN shows better performance than RF_RF while considering $\delta_{0.1}$ and σ, RF_RF displays superiority over RF_KNN; for the test set SDSS-WISE t1, RF_RF is better than RF_KNN except for $\delta_{0.3}$. In terms of the entire test sets (SDSS t1+t2+t3+t4 and SDSS-WISE t1+t2+t3+t4), RF_RF achieves slightly better results than RF_KNN.

5 DISCUSSION

The above results are summarized and compared in Table 6. The experimental results indicate that the accuracy of photometric redshift estimation can be generally improved by dividing the sample into subsamples and the accuracy of four subsamples is superior to that of two subsamples when not considering the percent in $\frac{|\Delta z|}{1+z_i} < 0.3$; the performance with information both from optical and infrared bands enhances compared to that with only optical information; the four estimation metrics ($\delta_{0.1}$, $\delta_{0.2}$, $\delta_{0.3}$ and σ) all only improve with the SDSS-WISE sample divided into two subsamples. Therefore, the scheme of dividing the sample is indeed effective and the accuracy of four subsamples is better than that of two subsamples. It is evident that the accuracy is rather satisfying if accurately knowing the redshift range of new objects in advance. The accuracy further improves through the classification system and the improvement in accuracy depends on the accuracy of classification into subsamples. In reality, we do not know the redshift range of new objects beforehand. If we want to better estimate redshift of new objects, then we

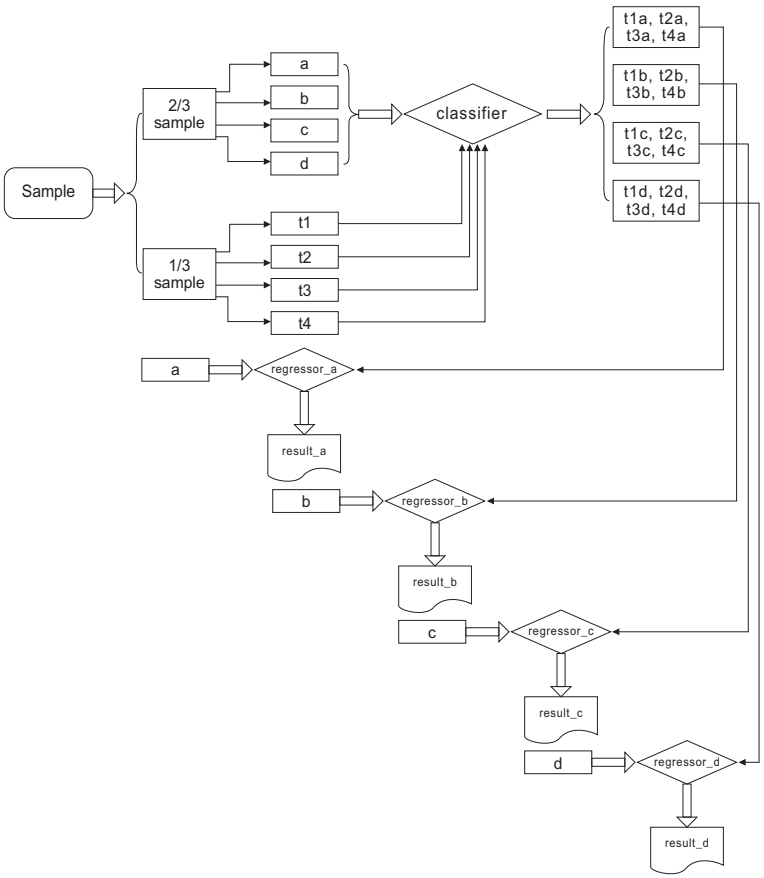

Fig. 5 Flow chart of photometric redshift estimation based on four subsamples.

Table 4 Performance of photometric redshift estimation for different datasets with kNN after classifying one sample into four subsamples by RF.

Data Set (Test set)	Algorithm	Model Parameters	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ
SDSS (t1)	RF_kNN	$k = 30$	65.36	75.91	80.47	0.3719
SDSS (t2)	RF_kNN	$k = 30$	72.96	84.83	86.88	0.2905
SDSS (t3)	RF_kNN	$k = 30$	81.79	86.92	87.90	0.3181
SDSS (t4)	RF_kNN	$k = 10$	95.57	96.66	96.82	0.1948
SDSS (t1+t2+t3+t4)			75.16	83.48	85.73	0.3235
SDSS-WISE (t1)	RF_kNN	$k = 20$	78.96	89.80	93.56	0.1949
SDSS-WISE (t2)	RF_kNN	$k = 30$	82.91	94.05	95.72	0.1885
SDSS-WISE (t3)	RF_kNN	$k = 30$	91.77	95.35	96.09	0.1833
SDSS-WISE (t4)	RF_kNN	$k = 10$	98.02	98.56	98.76	0.1498
SDSS-WISE (t1+t2+t3+t4)			84.63	92.95	95.08	0.1885

need to judge their redshift range. Therefore, it is necessary to construct a classification system before estimating photometric redshifts. More information from more bands leads to performance of a classifier or a regressor becoming better. In addition, there is a lot of room for improvement from the perspective of percents in different redshift ranges since the percent in $\frac{|\Delta z|}{1+z_i} < 0.3$ does not improve by the two schemes for most situations.

When the sample is classified into two/four subsamples, a discontinuity in the photometric redshift distribution exists due to misclassification near the cutoff. Because the accuracy of the classifier is much higher, the degree of discontinuity is much lower. Therefore, we may reduce the discontinuity by improving the accuracy of the classifier. In addition, when we utilize the photometric redshift catalog for further scientific study, we may adopt the estimated redshift value from two or four samples far from the cutoff, and keep the estimated value from one sample near the cutoff. Taking the SDSS-WISE sample into four subsamples by RF_RF for example, we adopt the estimated redshift value from the regressor and keep the estimated value from one sample by RF near the three cutoff points (± 0.3), then

Table 5 Performance of photometric redshift estimation for different datasets with RF after classifying one sample into four subsamples by RF.

Data Set (Test set)	Algorithm	Model Parameters	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ
SDSS (t1)	RF_RF	$n\_\text{estimators} = 200$ $\text{max\_depth} = 15$	65.65	75.88	80.27	0.3760
SDSS (t2)	RF_RF	$n\_\text{estimators} = 300$ $\text{max\_depth} = 15$	73.81	85.05	87.07	0.2854
SDSS (t3)	RF_RF	$n\_\text{estimators} = 300$ $\text{max\_depth} = 15$	82.04	87.16	88.16	0.3131
SDSS (t4)	RF_RF	$n\_\text{estimators} = 50$ $\text{max\_depth} = 15$	95.35	96.49	96.68	0.1935
SDSS (t1+t2+t3+t4)			75.56	83.62	85.82	0.3213
SDSS-WISE (t1)	RF_RF	$n\_\text{estimators} = 300$ $\text{max\_depth} = 15$	80.43	90.16	93.35	0.1916
SDSS-WISE (t2)	RF_RF	$n\_\text{estimators} = 300$ $\text{max\_depth} = 15$	83.35	93.69	95.45	0.1860
SDSS-WISE (t3)	RF_RF	$n\_\text{estimators} = 300$ $\text{max\_depth} = 15$	91.95	95.48	96.26	0.1770
SDSS-WISE (t4)	RF_RF	$n\_\text{estimators} = 200$ $\text{max\_depth} = 20$	97.31	98.30	98.52	0.1420
SDSS-WISE (t1+t2+t3+t4)			85.33	93.01	94.97	0.1843

Table 6 Performance of Photometric Redshift Estimation with Different Datasets for Different Schemes

Data Set	Scheme	Algorithm	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ
SDSS	one sample	kNN	62.53	80.13	87.17	0.3326
SDSS	two subsamples	RF_kNN	67.11	79.28	86.87	0.3267
SDSS	four subsamples	RF_kNN	75.16	83.48	85.73	0.3235
SDSS-WISE	one sample	kNN	79.40	91.37	95.28	0.1931
SDSS-WISE	two subsamples	RF_kNN	80.97	91.75	95.50	0.1926
SDSS-WISE	four subsamples	RF_kNN	84.63	92.95	95.08	0.1885
SDSS	one sample	RF	63.34	80.48	87.34	0.3271
SDSS	two subsamples	RF_RF	67.74	79.52	86.90	0.3235
SDSS	four subsamples	RF_RF	75.56	83.62	85.82	0.3213
SDSS-WISE	one sample	RF	79.87	91.37	95.23	0.1907
SDSS-WISE	two subsamples	RF_RF	81.60	91.87	95.35	0.1900
SDSS-WISE	four subsamples	RF_RF	85.33	93.01	94.97	0.1843

the metrics ($\delta_{0.1}$, $\delta_{0.2}$, $\delta_{0.3}$ and σ) are 85.76%, 93.28%, 95.19% and 0.1699, respectively. As a result, this method is applicable. To compare the performance of photometric redshift estimation, the true redshifts and estimated redshifts by different methods are depicted in Figure 7.

In general, there are many factors influencing the accuracy of photometric redshift estimation, among which the adopted techniques and selected features are most important. However, other factors are also not neglected. For example, Singal et al. (2011) presented the effects of including galaxy morphological parameters in photometric redshift estimation with an artificial neural network method. Way (2011) found that the broad bandpass photometry of the SDSS in combination with precise knowledge of galaxy morphology was helpful to improve the accuracy of estimating photometric redshifts for galaxies. Soo et al. (2018) studied the effects of incorporating galaxy morphology information in photometric redshift estimation, and found that the inclusion of quasar redshifts and associated object sizes in training improved the quality of photometric redshift catalogs and morphological information can mitigate biases and scatter due to bad photometry. Gomes et al. (2018) investigated improving photometric redshift estimation using GPZ by size information, post-processing and improved photometry. All these factors may be considered in our strategy in future work.

Considering improving the robustness, flexibility and automation of approaches for photometric redshift estimation, various tools in this aspect are in development, such as IMPZ (Babbedge et al. 2004), EAZY (Brammer et al. 2008, 2010), ArborZ (Gerdes et al. 2010), BPZ (Benítez 2011), Hyperz (Bolzonella et al. 2000, 2011), LePHARE (Arnouts & Ilbert 2011), ANNz (Collister & Lahav 2004; Niemack et al. 2009; Lahav & Collister 2012), PhotoZ (Saglia et al. 2013), XDQSO (Bovy et al. 2013), TPZ (Carrasco Kind & Brunner 2013), SOMz (Carrasco Kind & Brunner 2014), PhotoRApToR (Cavuoti et al. 2015a), GAz (Hogan et al. 2015), DAMEWARE (Cavuoti et al. 2015b), TailZ (Granett 2016), CuBANz (Samui & Pal 2016), GPZ (Almosallam et al. 2016a), ANNz2 (Sadeh et al. 2016) and Photo-z-SQL (Beck et al. 2017a,b). Abdalla et al. (2011) compared six photometric redshift codes (ANNz, HyperZ, SDSS, LePHARE, BPZ and ZEBRA) for 1.5 million luminous red galaxies (LRGs) in SDSS DR6. Therefore algo-

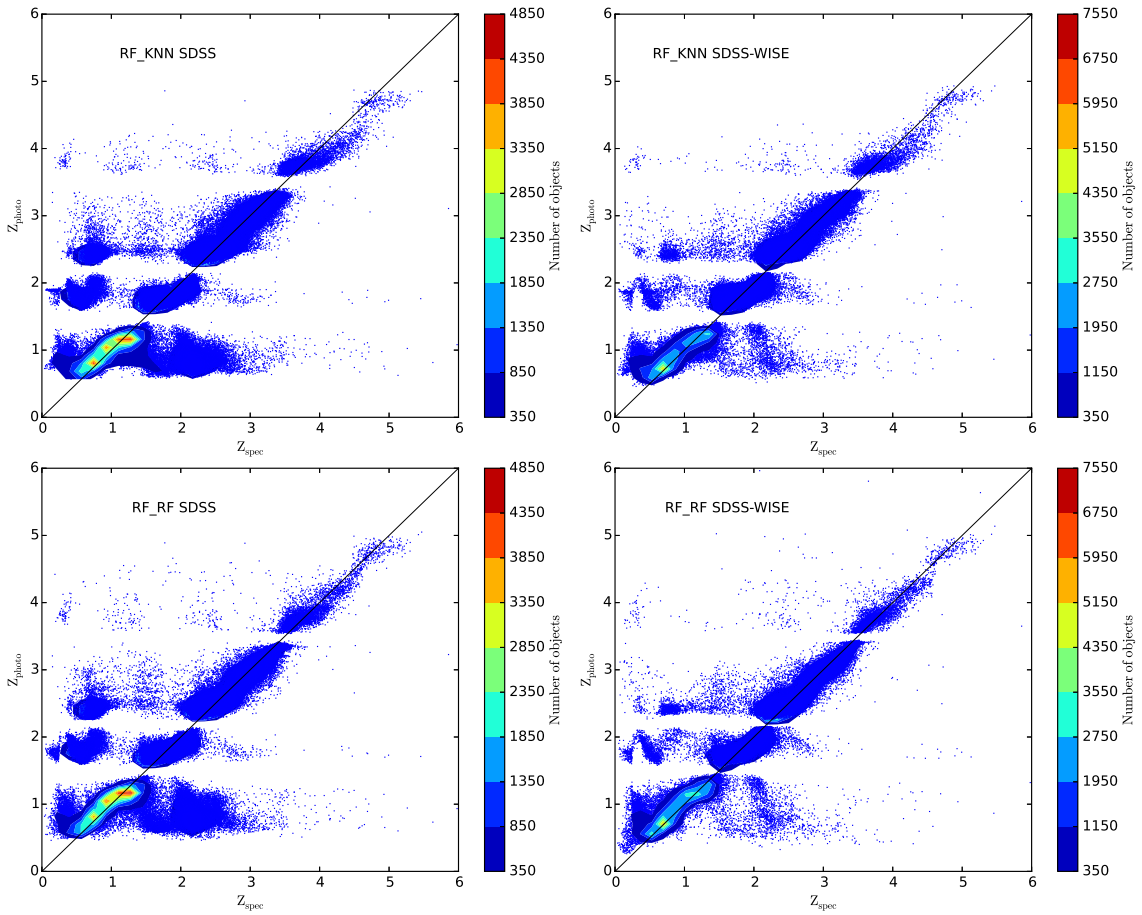

Fig. 6 Predicted photometric redshifts vs. spectroscopic redshifts for four subsamples. The *color bars* signify the number of objects per rectangular bin. The upper-left panel is based on SDSS test sample by *k*NN; the upper-right panel is based on SDSS-WISE test sample by *k*NN; the lower-left panel is based on SDSS test sample by RF; the lower-right panel is based on SDSS-WISE test sample by RF.

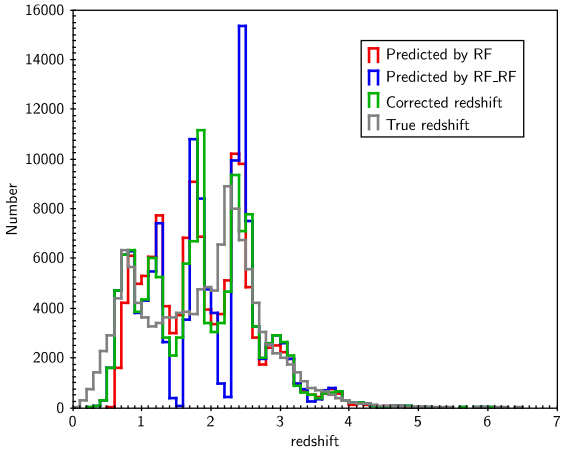

Fig. 7 Redshift distribution. *Grey line* represents true redshift; *red line* for estimated redshift from one sample by RF; *blue line* for estimated line from four samples by RF_RF; *green line* for estimated redshift from four samples by RF_RF and corrected near the cutoff.

rithms turning into automated tools are of great value once they are successfully applied in a specified issue. This is important and necessary for astronomers with such convenient tools in the big data era (Zhang & Zhao 2015).

6 CONCLUSIONS

In general, the work on accuracy improvement of photometric redshift estimation of quasars focuses on algorithm choice and feature selection. We design two schemes of

photometric redshift estimation by first classification and then regression, comparing the performance of RF and kNN with the SDSS and SDSS-WISE samples for the two schemes to the original scheme. We explore how to deal with the sample itself, and how the sample segmentation influences the estimation accuracy. Considering the experimental results, we are able to improve the estimation accuracy of photometric redshifts through first classification and then regression. In most of our experiments, the performance of dividing the sample into four subsamples is better than that of two subsamples with the two algorithms for the two samples, moreover the accuracy of both schemes improves compared to the original scheme, except for the percent in $\frac{|\Delta z|}{1+z_i} < 0.3$. In addition, for the SDSS-WISE dataset, no matter for RF or kNN, all the four metrics of performance criterion improve based on the sample divided into two parts compared to the one sample or four subsamples (see Table 6). RF shows a little better performance than kNN but its speed is slower than kNN since kNN is based on KD-tree index. The accuracy with the SDSS-WISE sample is superior to that with the SDSS sample when the same method is adopted. For the case of the SDSS-WISE sample divided into four subsamples by RF_RF, the estimated redshifts are adopted from the regressor and the estimated redshifts by RF with one sample replace them near the three cutoff points (± 0.3), then the metrics ($\delta_{0.1}$, $\delta_{0.2}$, $\delta_{0.3}$ and σ) amount to 85.76%, 93.28%, 95.19% and 0.1699, respectively. In other words, the strategy we put forward is effective. The accuracy of the classification system directly influences the performance of regression. The classification and regression also depend on the available information. As a result, information added from more bands is necessary to improve the accuracy of photometric redshift estimation and classification. In our next work, we will apply the databases (Pan-STARRS, future LSST, etc.) for this issue. Photometric redshift estimation of galaxies or other objects may be also improved by a similar strategy.

Acknowledgements We are very grateful to the constructive suggestions of the referee. This paper is funded by the 973 Program (2014CB845700) and the National Natural Science Foundation of China (Grant Nos. 11873066 and U1731109). We acknowledge the SDSS and WISE databases. Funding for the Sloan Digital Sky Survey (SDSS) IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is *www.sdss.org*. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) /University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University. This work makes use of data products from the *Wide-field Infrared Survey Explorer* (*WISE*), which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration.

References

Abdalla, F. B., Banerji, M., Lahav, O., et al. 2011, MNRAS, 417, 1891

Almosallam, I. A., Jarvis, M. J., & Roberts, S. J. 2016a, MNRAS, 462, 726

Almosallam, I. A., Lindsay, S. N., Jarvis, M. J., et al. 2016b, MNRAS, 455, 2387

Arnouts, S., & Ilbert, O. 2011, LePHARE: Photometric Analysis for Redshift Estimate, Astrophysics Source Code Library, ascl:1108.009

Babbedge, T. S. R., Rowan-Robinson, M., Gonzalez-Solares, E., et al. 2004, MNRAS, 353, 654

Ball, N. M., Brunner, R. J., Myers, A. D., et al. 2007, ApJ, 663, 774

Beck, R., Dobos, L., Budavári, et al. 2016, MNRAS, 460, 1371

Beck, R., Dobos, L., Budavári, T., et al. 2017a, Astronomy and Computing, 19, 34

Beck, R., Dobos, L., Budavári, T., et al. 2017b, Photo-z-SQL: Photometric Redshift Estimation Framework, Astrophysics Source Code Library, ascl:1704.009

Benítez, N. 2000, ApJ, 536, 571

Benítez, N. 2011, BPZ: Bayesian Photometric Redshift Code, Astrophysics Source Code Library, ascl:1108.011

Bolzonella, M., Miralles, J.-M., & Pelló, R. 2000, A&A, 363, 476

Bolzonella, M., Miralles, J.-M., & Pelló, R. 2011, Hyperz: Photometric Redshift Code, Astrophysics Source Code Library, ascl:1108.010
Bonfield, D. G., Sun, Y., Davey, N., et al. 2010, MNRAS, 405, 987
Bovy, J., Myers, A. D., Hennawi, J. F., et al. 2012, ApJ, 749, 41
Bovy, J., Hennawi, J. F., Hogg, D. W., et al. 2013, XDQSO: Photometic Quasar Probabilities and Redshifts, Astrophysics Source Code Library, ascl:1302.016
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2010, EAZY: A Fast, Public Photometric Redshift Code, Astrophysics Source Code Library, ascl:1010.052
Breiman, L. 2001, Machine Learning, 45, 5
Brescia, M., Cavuoti, S., D'Abrusco, R., et al. 2013, ApJ, 772, 140
Carliles, S., Budavári, T., Heinis, S., et al. 2010, ApJ, 712, 511
Carrasco, K. M., & Brunner, R. J. 2013, MNRAS, 432, 1483
Carrasco, K. M., & Brunner, R. J. 2014, MNRAS, 438, 3409
Cavuoti, S., Brescia, M., Longo, G., et al. 2012, A&A, 546, A13
Cavuoti, S., Brescia, M., De Stefano, V., et al. 2015a, Experimental Astronomy, 39, 45
Cavuoti, S., Brescia, M., Tortora, C., et al. 2015b, MNRAS, 452, 3100
Cavuoti, S., Amaro, V., Brescia, M., et al. 2017, MNRAS, 465, 1959
Collister, A. A., & Lahav, O. 2004, PASP, 116, 345
De Vicente, J., Sánchez, E., & Sevilla-Noarbe, I. 2016, MNRAS, 459, 3078
D'Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111
Edmondson, E. M., Miller, L., & Wolf, C. 2006, MNRAS, 371, 1693
Firth, A. E., Lahav, O., & Somerville, R. S. 2003, MNRAS, 339, 1195
Freeman, P. E., Newman, J. A., Lee, A. B., et al. 2009, MNRAS, 398, 2012
Gao, D., Zhang, Y.-X., & Zhao, Y.-H. 2009, RAA (Research in Astronomy and Astrophysics), 9, 220
Gerdes, D. W., Sypniewski, A. J., McKay, T. A., et al. 2010, ApJ, 715, 823
Gomes, Z., Jarvis, M. J., Almosallam, I. A., et al. 2018, MNRAS, 475, 331
Gorecki, A., Abate, A., Ansari, R., et al. 2014, A&A, 561, A128
Granett, B. R. 2016, TailZ: Redshift Distributions Estimator of Photometric Samples of Galaxies, Astrophysics Source Code Library, ascl:1602.013
Han, B., Ding, H.-P., Zhang, Y.-X., et al. 2016, RAA (Research in Astronomy and Astrophysics), 16, 74
Hogan, R., Fairbairn, M., & Seeburn, N. 2015, MNRAS, 449, 2040
Hoyle, B. 2016, Astronomy and Computing, 16, 34
Jones, E., & Singal, J. 2017, A&A, 600, A113
Jouvel, S., Delubac, T., Comparat, J., et al. 2017, MNRAS, 469, 2771
Lahav, O., & Collister, A. A. 2012, ANNz: Artificial Neural Networks for Estimating Photometric Redshifts, Astrophysics Source Code Library, ascl:1209.009
Laurino, O., D'Abrusco, R., Longo, G., et al. 2011, MNRAS, 418, 2165
Leistedt, B., & Hogg, D. W. 2017, ApJ, 838, 5
Mainzer, A., Bauer, J., Grav, T., et al. 2011, ApJ, 731, 53
Mortlock, D. J., Patel, M., Warren, S. J., et al. 2012, MNRAS, 419, 390
Niemack, M. D., Jimenez, R., Verde, L., et al. 2009, ApJ, 690, 89
Pâris, I., Petitjean, P., Aubourg, É., et al. 2018, A&A, 613, A51
Pedregosa, F., Gramfort, A., Michel, V., et al. 2013, Journal of Machine Learning Research, 12, 2825
Richards, G. T., Weinstein, M. A., Schneider, D. P., et al. 2001, AJ, 122, 1151
Sadeh, I., Abdalla, F. B., & Lahav, O. 2016, PASP, 128, 104502
Saglia, R. P., Snigula, J., Senger, R., & Bender, R. 2013, Experimental Astronomy, 35, 337
Samui, S., & Pal, S. S. 2016, CuBANz: Photometric Redshift Estimator, Astrophysics Source Code Library, ascl:1609.010
Samui, S., & Samui Pal, S. 2017, New Astron., 51, 169
Schindler, J.-T., Fan, X., McGreer, I. D., et al. 2017, ApJ, 851, 13
Singal, J., Shmakova, M., Gerke, B., et al. 2011, PASP, 123, 615
Soo, J. Y. H., Moraes, B., Joachimi, B., et al. 2018, MNRAS, 475, 3613
Speagle, J. S., & Eisenstein, D. J. 2017a, MNRAS, 469, 1186
Speagle, J. S., & Eisenstein, D. J. 2017b, MNRAS, 469, 1205
Wang, D., Zhang, Y. X., Liu, C., & Zhao, Y. H. 2007, MNRAS, 382, 1601
Way, M. J. 2011, ApJ, 734, L9
Way, M. J., Foster, L. V., Gazis, P. R., et al. 2009, ApJ, 706, 623
Way, M. J., & Klose, C. D. 2012, PASP, 124, 274
Way, M. J., & Srivastava, A. N. 2006, ApJ, 647, 102
Witten, I. H., & Frank, E. 2005, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco
Wolf, C. 2009, MNRAS, 397, 520
Wolf, C., Johnson, A. S., Bilicki, M., et al. 2017, MNRAS, 466, 1582
Wu, X.-B., Zhang, W., & Zhou, X. 2004, ChJAA (Chin. J. Astron. Astrophys.), 4, 17
Yèche, C., Petitjean, P., Rich, J., et al. 2010, A&A, 523, A14
York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579
Zhang, Y., Li, L., & Zhao, Y. 2009, MNRAS, 392, 233
Zhang, Y., Ma, H., Peng, N., et al. 2013, AJ, 146, 22
Zhang, Y., & Zhao, Y. 2015, Data Science Journal, 14, 11