RAA 2016 Vol. **16** No. **5**, 74 (10pp) doi: 10.1088/1674–4527/16/5/074
http://www.raa-journal.org *http://iopscience.iop.org/raa*

Research in Astronomy and Astrophysics

Photometric redshift estimation for quasars by integration of KNN and SVM

Bo Han[1], Hong-Peng Ding[1], Yan-Xia Zhang[2] and Yong-Heng Zhao[2]

[1] International School of Software, Wuhan University, Wuhan 430072, China
[2] Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China; *zyx@bao.ac.cn*

Received 2015 July 28; accepted 2016 January 5

Abstract The massive photometric data collected from multiple large-scale sky surveys offer significant opportunities for measuring distances of celestial objects by photometric redshifts. However, catastrophic failure is an unsolved problem with a long history and it still exists in the current photometric redshift estimation approaches (such as the k-nearest neighbor (KNN) algorithm). In this paper, we propose a novel two-stage approach by integration of KNN and support vector machine (SVM) methods together. In the first stage, we apply the KNN algorithm to photometric data and estimate their corresponding $z_{\rm phot}$. Our analysis has found two dense regions with catastrophic failure, one in the range of $z_{\rm phot} \in [0.3, 1.2]$ and the other in the range of $z_{\rm phot} \in [1.2, 2.1]$. In the second stage, we map the photometric input pattern of points falling into the two ranges from their original attribute space into a high dimensional feature space by using a Gaussian kernel function from an SVM. In the high dimensional feature space, many outliers resulting from catastrophic failure by simple Euclidean distance computation in KNN can be identified by a classification hyperplane of SVM and can be further corrected. Experimental results based on the Sloan Digital Sky Survey (SDSS) quasar data show that the two-stage fusion approach can significantly mitigate catastrophic failure and improve the estimation accuracy of photometric redshifts of quasars. The percents in different $|\Delta z|$ ranges and root mean square (rms) error by the integrated method are 83.47%, 89.83%, 90.90% and 0.192, respectively, compared to the results by KNN (71.96%, 83.78%, 89.73% and 0.204).

Key words: catalogs — galaxies: distances and redshifts — methods: statistical — quasars: general — surveys — techniques: photometric

1 INTRODUCTION

Photometric redshifts are obtained by images or photometry. Compared to spectroscopic redshifts, they show the advantages of high efficiency and low cost. Especially with the operation of multiple ongoing multiband photometric surveys, such as the Sloan Digital Sky Survey (SDSS), the UKIRT Infrared Deep Sky Survey (UKIDSS) and the Wide-Field Infrared Survey Explorer (WISE), a huge volume of photometric data is now being collected, which is larger than the corresponding spectroscopic data by two or three orders of magnitude. The massive photometric data offer significant opportunities for measuring the distances of celestial objects by photometric redshifts. However, photometric redshifts show the disadvantages of low accuracy compared to spectroscopic redshifts, and they require more sophisticated estimation algorithms to overcome the problem. Many researchers worldwide have investigated photometric redshift estimation techniques in recent years. Basically, these techniques are categorized into two types: template-fitting models and data mining approaches. The template-fitting model is the traditional approach for estimating photometric redshifts in astronomy. It extracts features from celestial observational information, such as multiband values, and then matches them with the designed templates constructed by theoretical models or real observations. With feature matching, researchers can estimate photometric redshifts. For example, Bolzonella et al. (2000) estimated photometric redshifts through a standard spectral energy distribution (SED) fitting procedure, where SEDs were obtained from broadband photometry. Wu et al. (2004) estimated the photometric redshifts of a large sample of quasars with the χ^2 minimization technique by using derived theoretical color-redshift relation templates. Rowan-Robinson et al. (2008) proposed an approach using fixed galaxy and quasar templates applied to data at 0.36–4.5 μm, and on a set of four infrared emission templates fitted to infrared excess data at 3.6–170 μm. Ilbert et al. (2009) applied a template-fitting method (Le PHARE) to calculate photometric redshifts in the 2-deg^2 COSMOS field. The experimental results from the above template-fitting methods have shown that their estimation accuracy relied on templates constructed by either simulation or real observational data.

Data mining approaches apply statistics and machine learning algorithms to a set of training samples, and they automatically learn complicated functional correlations between multiband photometric observations and their corresponding high confidence redshift parameters. These algorithms are data-driven approaches rather than template-driven approaches. The experimental results have shown that they have achieved much accurate photometric estimations in many applications. For example, Ball et al. (2008) applied a nearest neighbor algorithm to estimate photometric redshifts for galaxies and quasars using SDSS and the Galaxy Evolution Explorer (GALEX) data sets. Abdalla et al. (2008) estimated photometric redshifts by using a neural network method. Freeman et al. (2009) proposed a non-linear spectral connectivity analysis for transforming photometric colors to a simpler, more natural coordinate system wherein they applied regression to make redshift estimations. Gerdes et al. (2010) developed a boosted decision tree method, called ArborZ, to estimate photometric redshifts for galaxies. Way & Klose (2012) proposed an approach based on a Self-Organizing Map (SOM) to estimate photometric redshifts. Bovy et al. (2012) presented the extreme deconvolution technique for simultaneous classification and redshift estimation of quasars and demonstrated that the addition of information from UV and NIR bands was of great importance to photometric quasar-star separation and, essentially, the redshift degeneracies for quasars were resolved. Carrasco Kind & Brunner (2013) presented an algorithm using prediction trees and the random forest techniques for estimating photometric redshifts, incorporating measurement errors into the calculation while also efficiently dealing with missing values in the photometric data. Brescia et al. (2013) applied a MultiLayer Perceptron with Quasi Newton Algorithm (MLPQNA) to evaluate photometric redshifts of quasars with the data set from four different surveys (SDSS, GALEX, UKIDSS and WISE).

Although template-fitting approaches and data mining approaches can roughly estimate photometric redshifts, they both suffer from a problem of catastrophic failure in estimating photometric redshifts of quasars when the spectroscopic redshift is less than 3 (Richards et al. 2001; Weinstein et al. 2004; Wu et al. 2004). Zhang et al. (2013) practically demonstrated that with cross-matched multiband data from multiple surveys, such as SDSS, UKIDSS and WISE, the k-nearest neighbor (KNN) algorithm can largely solve the catastrophic failure problem and improve photometric redshift estimation accuracy. The method becomes more important with the development of multiple large photometric sky surveys and the coming age of astronomical big data. However, during the data preparation process, we need to cross-match multiband information of quasars from multiple photometric surveys. The number of matched quasar records is far less than the original number of quasars in a single survey. For example, there are 105 783 quasar samples available in SDSS DR7. However, the number of cross-matched samples from SDSS, WISE and UKIDSS is only 24 089. The number of cross-matched samples is around one fourth of the SDSS quasar data. This shortcoming greatly limits the scope of application for this estimation approach to only a small portion of cross-matched quasars observed by all surveys.

In this paper, we propose a novel two-stage photometric redshift estimation approach, i.e. the integration of KNN and support vector machine (SVM) approaches, to mitigate catastrophic failure for quasars by using relatively few band attributes from only a single survey. The rest of this paper is organized as follows. Section 2 describes the data used. Section 3 presents a brief overview of KNN, SVM and KNN+SVM. Section 4 gives the experimental results by KNN+SVM. The conclusions and discussions are summarized in Section 5.

2 DATA

Our experiments are based on a data set generated from SDSS (York et al. 2000), which labels highly reliable spectroscopic redshifts and has been widely used in photometric redshift estimation. The data set was constructed by Zhang et al. (2013) for estimating photometric redshifts of quasars. They used the samples of the Quasar Catalog V (Schneider et al. 2010) from SDSS DR7, which included 105 783 spectrally confirmed quasars. In each quasar record, five band features of u, g, r, i, z were provided. Similar to Zhang et al. (2013), in our experiments, we use these five attributes of $u-g, g-r, r-i, i-z, r$ ($4C+r$ for short) as the input and the corresponding spectroscopic redshift as a regression output.

3 METHODOLOGY

First, we study the characteristics of catastrophic failure for quasars and observe that the outliers by KNN are clustered into two groups: one group's spectroscopic redshift $z_{\rm spec}$ is between 0.2 and 1.1, while its photometric redshift $z_{\rm phot}$ is between 1.2 and 2.1, and the other group's $z_{\rm spec}$ is between 1.4 and 2.3, while its $z_{\rm phot}$ is between 0.3 and 1.2 (as shown in Figure 1). Some points with $z_{\rm phot}$ falling into Group 1 actually have $z_{\rm spec}$ close to the range of Group 2, but they are wrongly estimated by KNN and are mixed into Group 1, and vice versa. The two groups look almost 180-degree rotationally symmetric along the 45-degree diagonal line in the $z_{\rm phot}$ vs. $z_{\rm spec}$ diagram. The two outlier clusters show that the KNN method cannot effectively distinguish outliers from points that have good estimation using Euclidean distance in the two regions. Next, we propose a two-stage integrated approach by the fusion of KNN and SVM methods. In the first stage, we apply the KNN algorithm on photometric data and estimate their corresponding $z_{\rm phot}$. In the second stage, we map the photometric multi-band input pattern of points falling into the two ranges with $z_{\rm phot} \in [0.3, 1.2]$ and $z_{\rm phot} \in [1.2, 2.1]$ from an original attribute space into a high dimensional feature space by a Gaussian kernel function in SVM. In the high dimensional feature space, many outliers can be identified by a classification hyperplane in SVM and can be further corrected. Since most points resulting from catastrophic failure have been identified and corrected, our integrated approach can improve the accuracy of photometric redshift estimation.

The KNN algorithm generally applies Euclidean distance of attributes (shown in Eq. (1)) to compute distance between point m and point n,

$$d_{m,n} = \Big[(f_{m,1} - f_{n,1})^2 + (f_{m,2} - f_{n,2})^2 + ... \\ +(f_{m,k} - f_{n,k})^2\Big]^{1/2}, \qquad (1)$$

where $f_{m,j}(f_{n,j})$ denotes the jth attribute among the $4C + r$ input pattern for the mth (nth) point and k represents the total number of attributes. The points in Group 1 and Group 2 show that those outlier quasars cannot be correctly identified in a Euclidean space. In other words, we cannot have a simple plane as a useful separation criterion between points in Group 1 and Group 2. Based on the present data, the information provided is not enough to give a good estimation of the outliers. Now there is a question of whether those outliers can be linearly separable in a high-dimensional non-Euclidean feature space? Thereby, we explore the kernel function in SVM and map features into a high dimensional space and test if we can correctly classify outliers in Group 1 or Group 2. From the analysis above, we propose a two-stage integrated approach by fusion of estimation with KNN and classification with SVM.

3.1 Estimation with KNN

The KNN algorithm is a lazy predictor which requires a training set for learning. It first finds the nearest neighbors by comparing distances between a test sample and training samples that are similar to it in a feature space. Next, it assigns the average value of the nearest neighbors to the test sample as its prediction value. In general, the distance is computed as Euclidean distance described in Equation (1). In the era of big data, we have been collecting more data than ever before and KNN achieves very accurate predictions (Zhang et al. 2013). Thereby, we also use KNN in our research. One disadvantage of KNN is the high computational cost. We apply KD-tree to efficiently implement the KNN algorithm.

3.2 Classification with SVM

SVM is an effective classification algorithm based on the principle of structural risk minimization proposed by Vapnik (1995). Given a training data set with n records, where each record has the pattern (x_i, y_i) for $i = 1, 2, ..., n$, we aim to build a linear classifier with the following Equation (2),

$$f(x) = w \cdot x + b. \qquad (2)$$

Here, w and b are the weight vector and bias respectively.

Figure 2 illustrates that several lines can separate two categories of points. In SVM, for minimizing the classification error risk for other test data sets, we aim to find lines (shown as the dot-dashed lines in Fig. 2) with the maximized margin that can separate the two classes of points. In many classification tasks this principle gives SVM a better classification accuracy than other competing machine learning models.

Sometimes, a classification task is hard and not linearly solvable. The left panel in Figure 3 shows one such case. In this case, by using Vapnik-Chervonenk dimension theory, SVM applies a kernel function that transforms ordinary original flat space into inner products that are more effective in this type of classification. By using the theory of reproducing kernels, we can map the original Euclidean feature space to the high-dimensional non-Euclidean feature space with the SVM classification algorithm. Thereby, some non-linear problems in the original low-dimensional feature space $\Re^d$ become linearly solvable in high-dimensional space $\Re^D$. The right panel in Figure 3 illustrates how the mapping by a kernel function solves the problem. Therefore, Equation (2) can be transformed to the following form by a feature mapping function $\emptyset$,

$$f(x) = w \cdot \emptyset(x) + b. \qquad (3)$$

In this way, we have the following objective function and constraints for an SVM classifier, and minimize ,

$$|| \, w \, ||^2 + C \sum_{i=1}^{n} \epsilon_i$$

subject to

$$y_i \cdot (< w, \emptyset(x_i) > +b) \geq 1 - \epsilon_i, \qquad (4)$$

where C is a regularization parameter and ϵ_i is a slack variable.

By using the represented theorem, we have,

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \emptyset(x_i)^T \emptyset(x) + b, \qquad (5)$$

where α_i is a parameter with the constraint that $\alpha_i \geq 0$. For solving Equation (5), SVM introduces a kernel function defined as,

$$K(x_i, x) = \emptyset(x_i)^T \emptyset(x). \qquad (6)$$

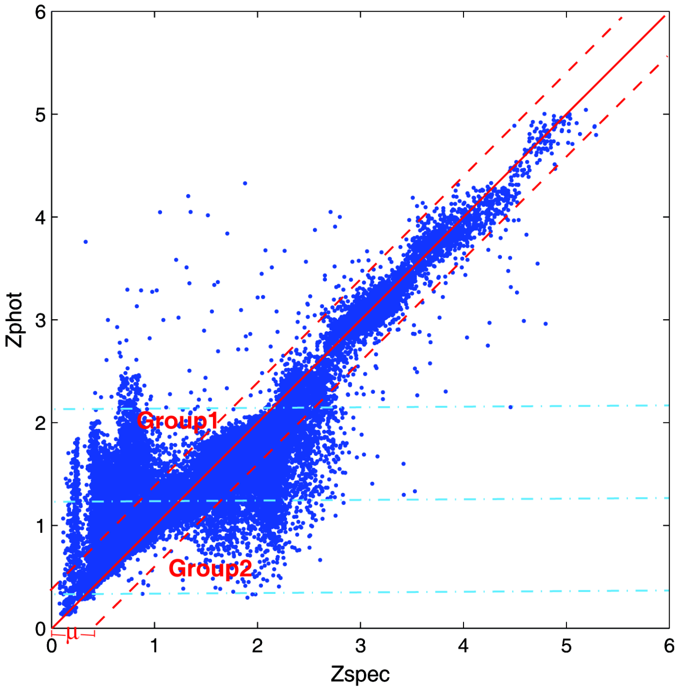

Fig. 1 Photometric redshift estimation by KNN estimation. The points in Group 1 and Group 2 are outliers. μ is the parameter representing the error tolerance interval. The slanted dashed lines show the corrected estimation range of photometric redshifts. The horizontal dashed lines define the zones corresponding to Group 1 and Group 2.

In this paper, we practically apply a Gaussian kernel (shown in Eq. (7)) to achieve the non-linear classification.

$$K(x_i, x) = e^{-\frac{||x_1 - x_2||^2}{2\sigma^2}}, \qquad (7)$$

where x_1, x_2 represent vectors of multiband attributes or input patterns observed from a single survey, and σ is a free parameter.

In this way, we aim to apply an SVM classifier to distinguish the mixture of points in Group 1 from other points around the minor diagonal with $z_{\rm phot} \in [1.2, 2.1]$ in the $z_{\rm phot}$ vs. $z_{\rm spec}$ diagram. Similarly, we can distinguish points in Group 2 from other points with $z_{\rm phot} \in [0.3, 1.2]$.

3.3 Integration of KNN and SVM for Photometric Redshift Estimation

The photometric redshift estimation algorithm that integrates KNN and SVM is presented in the following. To obtain a robust accuracy measure for our integrated approach, we repeat the experiments for num trials. In each trial, the initialization step, KNN step, SVM training step, SVM test step, correction step and evaluation step will be applied to the data sets. In the initialization step, we randomly divide the SDSS data set into a separate training set, validation set and test set. In the KNN step, we apply the KNN algorithm ($k = 17$) to estimate $z_{\rm phot-validation}$ and $z_{\rm phot-test}$ based on the training set and the union of the training set and validation set respectively. In the SVM training step, we aim to build two SVM classifiers: SVM1 and SVM2 to distinguish good estimations from outliers with $z_{\rm phot-validation} \in [1.2, 2.1]$ and $z_{\rm phot-validation} \in [0.3, 1.2]$, respectively. The good identification of outliers is defined by the following Equation (8),

$$\begin{cases} \mid z_{\rm spec} - z_{\rm phot} \mid \; \leq \; \mu \quad \text{good} \\ \mid z_{\rm spec} - z_{\rm phot} \mid \; > \; \mu \quad \text{bad}\,. \end{cases} \qquad (8)$$

Here, μ is the parameter which means the error tolerance interval derived from the validation set.

Visually, good estimation points will fall into an area close to a 45-degree diagonal line in the diagram, while the outliers will fall into Group 1 or Group 2 in Figure 1.

Specifically, we use those outliers with $z_{\rm phot-validation} \in [1.2, 2.1]$ and $z_{\rm phot-validation} \in [0.3, 1.2]$ to construct data sets Group1_trainingdata and Group2_trainingdata, respectively. In the two data sets, inputs are patterns $4C + r$ and $z_{\rm phot}$ directly from KNN, and the output is $z_{\rm spec}$.

In the SVM test step, we apply classifiers SVM1 and SVM2 to identify outliers.

In the correction step, we use the KNN algorithm based on Group1_data to compute $z_{\rm phot}$ for those outliers distinguished by SVM1 in test data. Since Group1_data and those outliers have a similar pattern but the output of Group1_trainingdata is $z_{\rm spec}$, the KNN algorithm can improve the $z_{\rm phot}$ estimation. Similarly, we can use Group

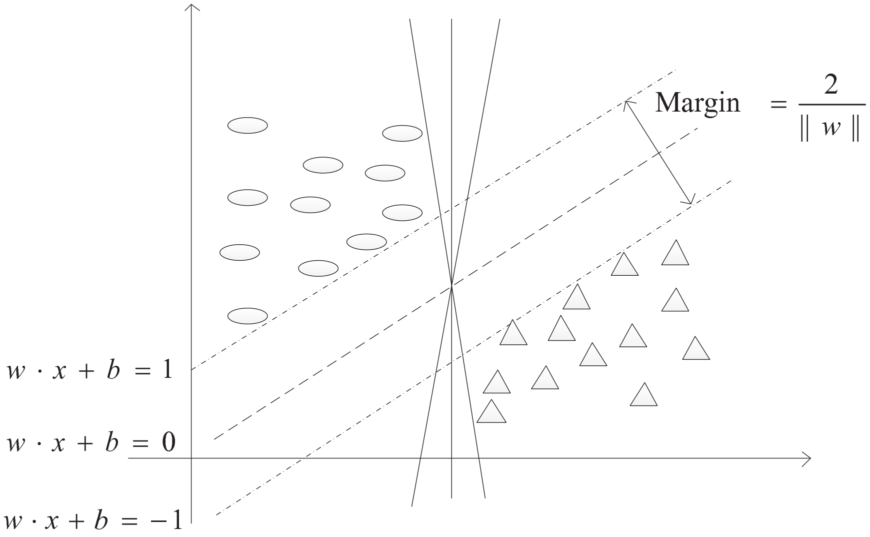

Fig. 2 Maximizing the classification margin is the aim in SVM. The points on the dot-dashed lines are called support vectors. The distance between the two dot-dashed lines is called the margin. When the margin is maximized, the classification accuracy achieves its best performance.

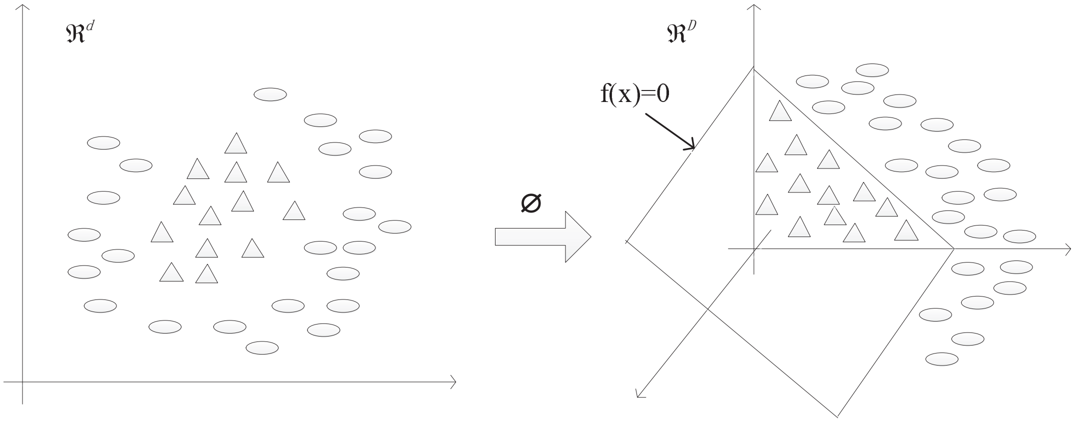

Fig. 3 Linearly indistinguishable points in a low dimensional space ($\Re^d$) can be separated in a high dimensional space ($\Re^D$) by the application of a kernel function ($\emptyset$) in SVM. $f(x) = 0$ represents the hyperplane that separates the two classes.

2 to train data and then correct outliers distinguished by SVM2 in test data.

In the evaluation step, we apply the percents in different $|\Delta z|$ ranges and root mean square (rms) error of Δz to test our photometric redshift estimation approach. The definition of Δz is listed in Equation (9).

$$\Delta z = \frac{z_{\rm spec} - z_{\rm phot}}{1 + z_{\rm spec}}. \tag{9}$$

The detailed steps of the two-stage method are as follows. Also to be clear, a flow chart of the whole process is shown in Figure 4.

LoopId= 1;
Do while LoopId≤ num;
Initialization Step:

Randomly select a 1/3 sample from the SDSS quasar sample as the training set, another 1/3 sample as the validation set and the remaining 1/3 sample as the test set.

KNN Step:

(1) Based on the training set, we apply the KNN ($k = 17$) algorithm to estimate $z_{\rm phot-validation}$ for each sample in the validation set;

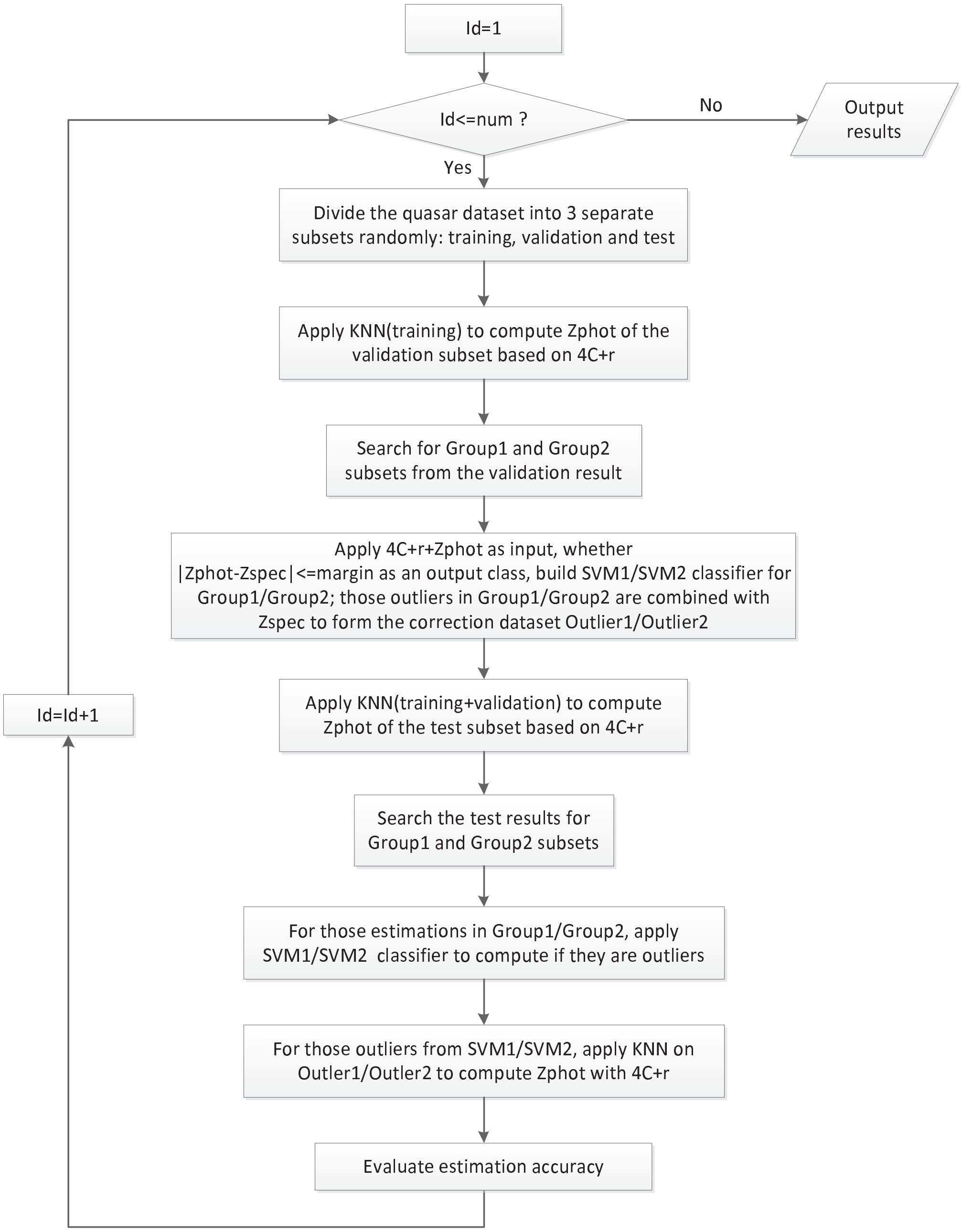
Id=1
Id<=num ?
No
Output results
Yes
Divide the quasar dataset into 3 separate subsets randomly: training, validation and test
Apply KNN(training) to compute Zphot of the validation subset based on 4C+r
Search for Group1 and Group2 subsets from the validation result
Apply 4C+r+Zphot as input, whether |Zphot-Zspec|<=margin as an output class, build SVM1/SVM2 classifier for Group1/Group2; those outliers in Group1/Group2 are combined with Zspec to form the correction dataset Outlier1/Outlier2
Apply KNN(training+validation) to compute Zphot of the test subset based on 4C+r
Id=Id+1
Search the test results for Group1 and Group2 subsets
For those estimations in Group1/Group2, apply SVM1/SVM2 classifier to compute if they are outliers
For those outliers from SVM1/SVM2, apply KNN on Outler1/Outler2 to compute Zphot with 4C+r
Evaluate estimation accuracy

Fig. 4 The flow chart of photometric redshift estimation by the integration of KNN and SVM.

(2) Based on the union of the training set and validation set, we apply the KNN ($k = 17$) algorithm to estimate $z_{\rm phot-test}$ for each sample in the test set.

SVM Training Step:

(1) For those samples with $z_{\rm phot-validation} \in [1.2, 2.1]$ in the validation set, we train a classifier SVM1 with a Gaussian kernel, which distinguishes good estimations from outliers by Equation (8). With those outliers, we build a data set Group1_trainingdata, which is composed of $4C + r$ and $z_{\rm phot}$ as the input and $z_{\rm spec}$ as the output;

(2) Similarly, for those samples with $z_{\rm phot-validation} \in [0.3, 1.2]$ in the validation set, we train a classifier SVM2 with a Gaussian kernel, which distinguishes good estimations from outliers. With those outliers, we build a data set Group2_trainingdata, which is composed of $4C + r$ and $z_{\rm phot}$ as the input and $z_{\rm spec}$ as the output.

SVM Test Step:

(1) For those samples with $z_{\rm phot-test} \in [1.2, 2.1]$ in the test set, we apply the classifier SVM1 to distinguish a good estimation from outliers;
(2) Similarly, for those samples with $z_{\rm phot-test} \in [0.3, 1.2]$ in the test set, we apply the classifier SVM2 to distinguish a good estimation from outliers.

Correction Step:

(1) For those outliers with $z_{\rm phot-test} \in [1.2, 2.1]$ in the test set, we apply the KNN algorithm based on the data set Group1_trainingdata;
(2) For those outliers with $z_{\rm phot-test} \in [0.3, 1.2]$ in the test set, we apply the KNN algorithm based on the data set Group2_trainingdata.

Evaluation Step:

By comparing $z_{\rm phot-test}$ and $z_{\rm spec}$ for all of the samples in the test set, we compute the popular measures of accuracy for the redshift estimation and the associated rms error of Δz.

LoopId = LoopId+1;

End do.

Output the mean and standard error for the percents in different $|\Delta z|$ ranges and rms error of Δz to evaluate the accuracy of our proposed integrated approach of KNN+SVM.

4 EXPERIMENTAL RESULTS

In the experiments, we adopt the input patterns $4C + r$ as attributes, which are widely accepted by recent researches on photometric redshift estimation. In our designed algorithm, we practically set $num = 10$ and repeat the experiments 10 times.

For classification, we apply the widely used tool LIBSVM (Chang & Lin 2011). By using a Gaussian kernel function, we train classifiers SVM1 and SVM2 with the samples with $z_{\rm phot-validation} \in [1.2, 2.1]$ and with $z_{\rm phot-validation} \in [0.3, 1.2]$ in the validation set, respectively. To optimize the estimation accuracy, we adjust two parameters controlling the Gaussian kernel in SVM, a cost coefficient C that corrects for imbalance in the data and a factor γ that takes the shape of the high dimensional feature space into account. Other parameters are set to their default values in LIBSVM. In order to obtain the best model parameters, a grid search is adopted. The grid search in SVM1 and SVM2 is indicated in Figure 5.

For SVM1, the optimal model parameter C is 2 and γ is 8, while the classification accuracy is 94.12%. For SVM2, the best model parameter C is 128, γ is 0.5, and the classification accuracy amounts to 90.04%.

With the optimized parameters and the union of the training set and the validation set as a new training set, we compare the estimation accuracy between the original KNN ($k = 17$) algorithm and our integrated approach of KNN+SVM. The parameter μ is a factor to determine whether a point has a good estimation or not. We change the value of μ to check its influence on the estimation accuracy.

The results are listed in Table 1. For KNN, the proportions of $|\Delta z| < 0.1, 0.2, 0.3$ and rms error of predicted photometric redshifts are 71.96%, 83.78%, 89.73% and 0.204, respectively; for KNN+SVM, these optimal measures are 83.47%, 89.83%, 90.90% and 0.192, respectively, when $\mu = 0.3$, which are shown as bold in Table 1. Obviously, these criteria for photometric redshift estimation are all significantly improved with the new method. This suggests that the integrated approach can effectively correct those outliers with $z_{\rm phot} \in [1.2, 2.1]$ and $z_{\rm phot} \in [0.3, 1.2]$. Thereby, it can significantly mitigate catastrophic failure and improve the estimation accuracy of photometric redshifts.

The experimental results also show that without cross-matching multiband observations from multiple surveys, we can effectively apply a Gaussian kernel function in SVM to identify outliers in Group 1 and Group 2 to protect from catastrophic failure by mapping attributes from a single data source into a high dimensional feature space. The identification helps us correct those outliers and thereby improves estimation accuracy.

In order to compare the performance of photometric redshift estimation by the KNN algorithm with that by the KNN and SVM approach, the photometric redshift estimation with these two methods is shown in Figures 6 and 7, respectively. As indicated by Figures 6 and 7, we can see clearly that the outliers in both Group 1 and Group 2 have been significantly decreased by adopting the new method of KNN+SVM. This intuitively demonstrates that our proposed approach is effective.

5 CONCLUSIONS AND DISCUSSION

Catastrophic failure is an unsolved problem with a long history and it exists in most photometric redshift estimation approaches. In this paper, we first analyze the reasons for catastrophic failure associated with quasars and point out that the outliers result from being non-linearly separable in Euclidean feature space of an input pattern. Next, we propose a new estimation approach by integration of KNN and SVM methods together. By using a Gaussian kernel function in SVM, we map a multiband input pattern from an original Euclidean space into a high dimensional feature space. In this way, many outliers can be identified by a hyperplane and then corrected. The experimental results based on SDSS data for quasars show that the integrated approach can significantly mitigate catastrophic failure and improve the photometric redshift es-

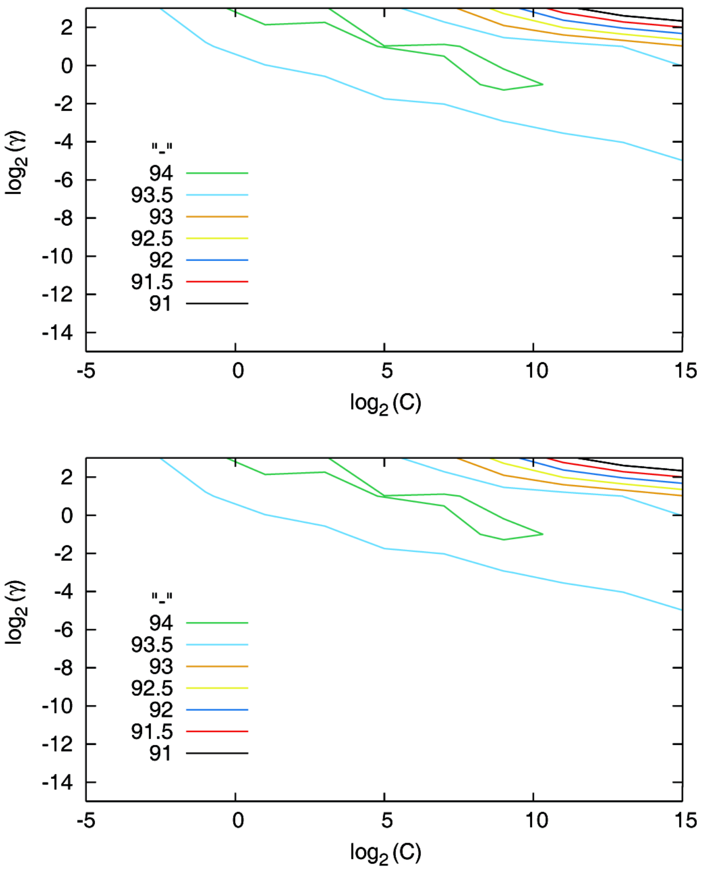

Fig. 5 Top: the best model parameter in SVM1 is obtained by grid search, i.e. $C = 2$, $\gamma = 8$ and the accuracy of classification achieves 94.12%. Bottom: the best model parameter in SVM2 is obtained by grid search, i.e. $C = 128$, $\gamma = 0.5$ and the accuracy of classification is 90.04%.

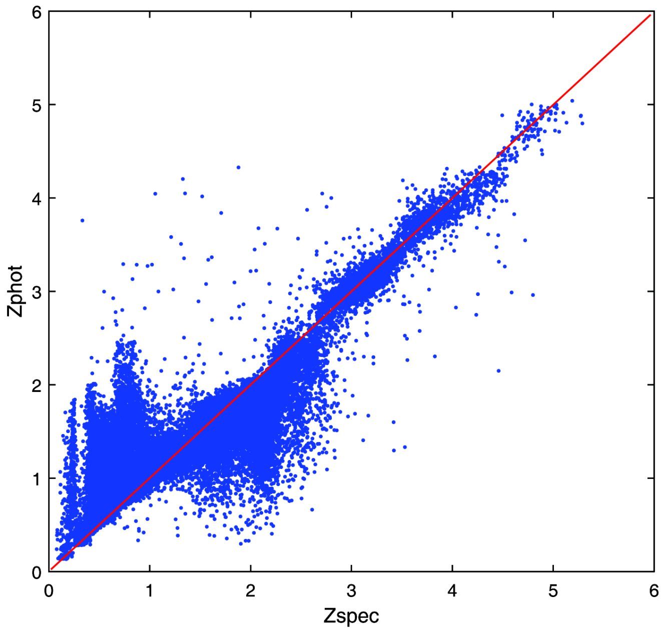

Fig. 6 Photometric redshift estimation by KNN.

Table 1 Comparison of KNN and Our Integrated Approach

Method	$\|\Delta z\|< 0.1(\%)$	$\|\Delta z\|< 0.2(\%)$	$\|\Delta z\|< 0.3(\%)$	rms error
KNN ($k = 17$)	71.96±0.20	83.78±0.18	89.73±0.16	0.204±0.004
SVM+KNN ($\mu = 0.1$)	75.06±3.03	81.43±2.31	85.51±1.69	0.232±0.022
SVM+KNN ($\mu = 0.2$)	80.86±1.19	85.56±1.95	86.57±1.81	0.224±0.013
SVM+KNN ($\mu = 0.3$)	**83.47±0.86**	**89.83±0.51**	**90.90±0.42**	**0.192±0.007**
SVM+KNN ($\mu = 0.4$)	81.63±0.64	89.53±0.32	91.54±0.33	0.193±0.005
SVM+KNN ($\mu = 0.5$)	78.89±0.22	88.30±0.24	91.63±0.21	0.194±0.005
SVM+KNN ($\mu = 0.6$)	75.84±0.14	86.60±0.13	90.58±0.11	0.199±0.003

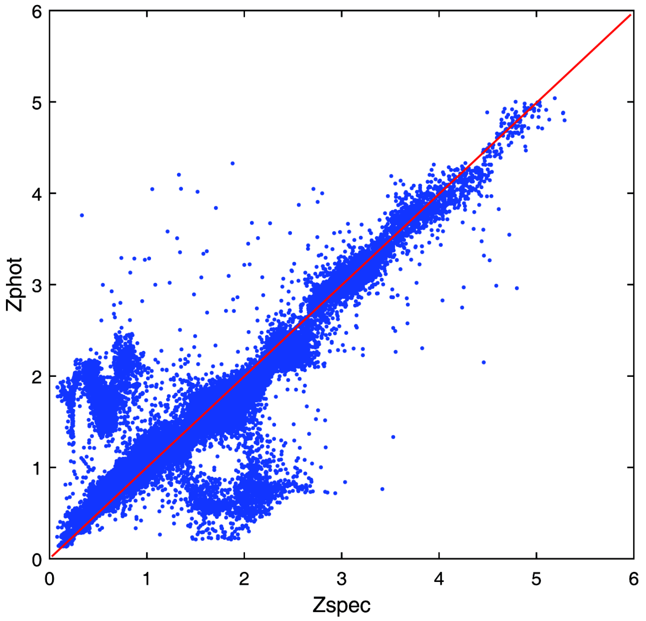

Fig. 7 Photometric redshift estimation by KNN+SVM.

timation accuracy, e.g. the percentages in different $|\Delta z|$ ranges and rms error are 83.47%, 89.83%, 90.90% and 0.192, respectively. Although different previous research has tried to mitigate catastrophic failure by cross-matching the data from several surveys, our approach can achieve a similar objective from only a single survey and it does not need to cross-match among multiple surveys, thus avoiding cross-match efforts especially for the growing amount of large survey data. Moreover, not all sources have observations from different surveys. Therefore this method can be widely applied for a single large photometric data set from a sky survey. In addition, the integrated method with data from more bands may further improve the accuracy of estimating the photometric redshifts of quasars.

Acknowledgements We are very grateful to the referee's important comments and suggestions which helped us improve our paper. This work is supported by the National Natural Science Foundation of China (Grant Nos. 61272272 and U1531122), the Natural Science Foundation of Hubei province (Grant 2015CFA058), the National Key Basic Research Program of China (2014CB845700) and the NSFC-Texas A&M University Joint Research Program (No.11411120219). We acknowledge the SDSS database. 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, University of Cambridge, 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.

References

Abdalla, F. B., Amara, A., Capak, P., et al. 2008, MNRAS, 387, 969

Ball, N. M., Brunner, R. J., Myers, A. D., et al. 2008, ApJ, 683, 12

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

Bovy, J., Myers, A. D., Hennawi, J. F., et al. 2012, ApJ, 749, 41

Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., & Mercurio, A. 2013, ApJ, 772, 140

Carrasco Kind, M., & Brunner, R. J. 2013, MNRAS, 432, 1483

Chang, C.-C., & Lin, C.-J. 2011, ACM Transactions on Intelligent Systems and Technology (TIST), 2, 27

Freeman, P. E., Newman, J. A., Lee, A. B., Richards, J. W., & Schafer, C. M. 2009, MNRAS, 398, 2012

Gerdes, D. W., Sypniewski, A. J., McKay, T. A., et al. 2010, ApJ, 715, 823

Ilbert, O., Capak, P., Salvato, M., et al. 2009, ApJ, 690, 1236

Richards, G. T., Weinstein, M. A., Schneider, D. P., et al. 2001, AJ, 122, 1151

Rowan-Robinson, M., Babbedge, T., Oliver, S., et al. 2008, MNRAS, 386, 697

Schneider, D. P., Richards, G. T., Hall, P. B., et al. 2010, AJ, 139, 2360

Vapnik, V. 1995, The Nature of Statistical Learning Theory (New York: Springer)

Way, M. J., & Klose, C. D. 2012, PASP, 124, 274

Weinstein, M. A., Richards, G. T., Schneider, D. P., et al. 2004, ApJS, 155, 243

Wu, X.-B., Zhang, W., & Zhou, X. 2004, ChJAA (Chin. J. Astron. Astrophys.), 4, 17

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579

Zhang, Y., Ma, H., Peng, N., Zhao, Y., & Wu, X.-b. 2013, AJ, 146, 22