Research in Astron. Astrophys. **2009** Vol. **9** No. **3**, 341–348
http://www.raa-journal.org *http://www.iop.org/journals/raa*

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

Automated spectral classification using template matching *

Fu-Qing Duan[1], Rong Liu[2], Ping Guo[1], Ming-Quan Zhou[1] and Fu-Chao Wu[3]

[1] College of Information Science and Technology, Beijing Normal University, Beijing 100875, China; *fqduan@gmail.com*
[2] Base Department, Beijing Institute of Clothing Technology, Beijing 100029, China
[3] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China

Received 2008 March 18; accepted 2008 June 13

Abstract An automated spectral classification technique for large sky surveys is proposed. We firstly perform spectral line matching to determine redshift candidates for an observed spectrum, and then estimate the spectral class by measuring the similarity between the observed spectrum and the shifted templates for each redshift candidate. As a byproduct of this approach, the spectral redshift can also be obtained with high accuracy. Compared with some approaches based on computerized learning methods in the literature, the proposed approach needs no training, which is time-consuming and sensitive to selection of the training set. Both simulated data and observed spectra are used to test the approach; the results show that the proposed method is efficient, and it can achieve a correct classification rate as high as 92.9%, 97.9% and 98.8% for stars, galaxies and quasars, respectively.

Key words: methods: data analysis — techniques: spectroscopic — stars: general — galaxies: stellar content

1 INTRODUCTION

The rapid development of astronomical observations has led to many large sky surveys such as SDSS (Sloan Digital Sky Survey), 2dF (2 degree Fields) and LAMOST (Large Sky Area Multi-Object Spectroscopic Telescope). Since these surveys have produced very large numbers of spectra, automated spectral recognition becomes desirable and necessary for efficiency. Spectral classification by astronomers mainly focuses on stellar classification (von Hippel et al. 1994; Bailer-Jones et al. 1998; Bai et al. 2005) and galaxy classification (Connolly et al. 1995; Galaz et al. 1998; Zaritsky et al. 1995), where the latter usually needs to know redshifts of the spectra. How to automatically separate spectra with unknown redshifts seems a more difficult and challenging task. In this study, the spectra with unknown redshifts will be roughly classified into three types: star, galaxy and quasar (QSO). Qin et al. (2003) and Zhang & Zhao (2003, 2004) have done some research on the rough classification of spectra using support vector machines (SVM) and radial basis function neural networks (RBF). However, both methods need long training times and are sensitive to the distribution of the training samples.

In this work, we present a new template matching method that combines cross-correlation and spectral line matching. We firstly determine the redshift candidates for the observed spectrum by spectral line matching, and then classify the observed spectrum by measuring the similarity between the spectrum

* Supported by the National Natural Science Foundation of China.

and the templates shifted to match each candidate. Principal components analysis (PCA) is a popular technique for data compression and feature extraction. It has been widely used in automated spectral analysis (Bailer-Jones et al. 1998; Connolly et al. 1995). We use PCA to construct the template spectra of the star and galaxy. The similarity measure adopted, similar to evidence accumulation, is the weighted sum of several similarity evidences. It can measure the similarity between two spectra more reasonably. Compared with some spectral classification methods based on learning, such as neural networks, SVM etc, the proposed approach needs no training, which is time-consuming and sensitive to the selection of the training set. As a byproduct of this approach, the spectral redshift can also be estimated with high accuracy.

The organization of this paper is as follows. Section 2 shows the data set used in our study. Section 3 discusses the spectral line extraction. Section 4 describes the spectral classification and redshift determination. Section 5 shows and discusses the experimental results. This paper concludes in Section 6.

2 THE DATA SET

PCA is a linear feature extraction method based on the variance of the data distribution. It finds the principle components that can most effectively characterize the inputs. Here, we use PCA to build the rest templates of star and galaxy spectra, and the composite quasar spectrum constructed by Vanden Berk et al. (2001) is used as the quasar template, which is shown in Figure 1(e).

The 161 standard stellar spectra, contributed by Jacoby et al. (1984), are selected to build the stellar templates. Two eigen-spectra, as shown in Figure 1(a) and Figure 1(b), are obtained by PCA with the variance contribution rate of 99%. The first spectrum mainly shows the spectral feature of early-type stars, while the second one mainly captures the spectral feature of late-type stars. These two eigen-spectra are the stellar templates.

The eleven rest galaxy spectra presented by Kinney & Calzetti et al. (1996), which are E, S0, Sa, Sb, Sc, Sb1, Sb2, Sb3, Sb4, Sb5 and Sb6, are used to build the rest galaxy templates. From the four normal galaxy spectra, E, S0, Sa and Sb, we get one eigen-spectrum by PCA, as shown in Figure 1(c). From the seven starburst galaxy spectra, Sc, Sb1, Sb2, Sb3, Sb4, Sb5 and Sb6, we also get an eigen-spectrum, as shown in Figure 1(d). These two eigen-spectra are the galaxy templates. It can be seen that these two templates capture the features of the rest absorption lines and emission lines of galaxy spectra respectively.

3 SPECTRAL LINE EXTRACTION

3.1 Preprocessing

The spectral preprocessing includes sky subtraction, continuum subtraction and de-noising. Usually, the observed spectra show a number of residual sky features in the regions of strong atmospheric emission and absorption lines. While these sky residuals are the strongest features in the spectra, they will result in false peaks in cross-correlation between the templates and the observed spectra, which can lead to incorrect spectral classification. We remove the sky residuals around 557 nm, 630 nm, etc by median filtering.

Continuum subtraction reduces the smoothly varying background to zero. It has the same effect as filtering out the long-period Fourier components of spectra. Without continuum subtraction, the cross-correlation function mainly represents the cross-correlation of the two continua, with a small spectral cross-correlation superimposed. The continuum is fitted from the observed spectrum by running a 60-nm median filter. Since the spectral line widths are generally less than 60-nm, nearly all absorption or emission features are removed by this processing. Although it fits the continuum badly in areas where the spectral shape changes rapidly, this will have little effect on the following spectral line extraction, as long as the continuum fits are right for those main spectral lines.

Since high noise in spectra will make the spectral line extraction very difficult, noise reduction is necessary. The wavelet soft-thresholding (Donoho 1995) is used for the noise reduction.

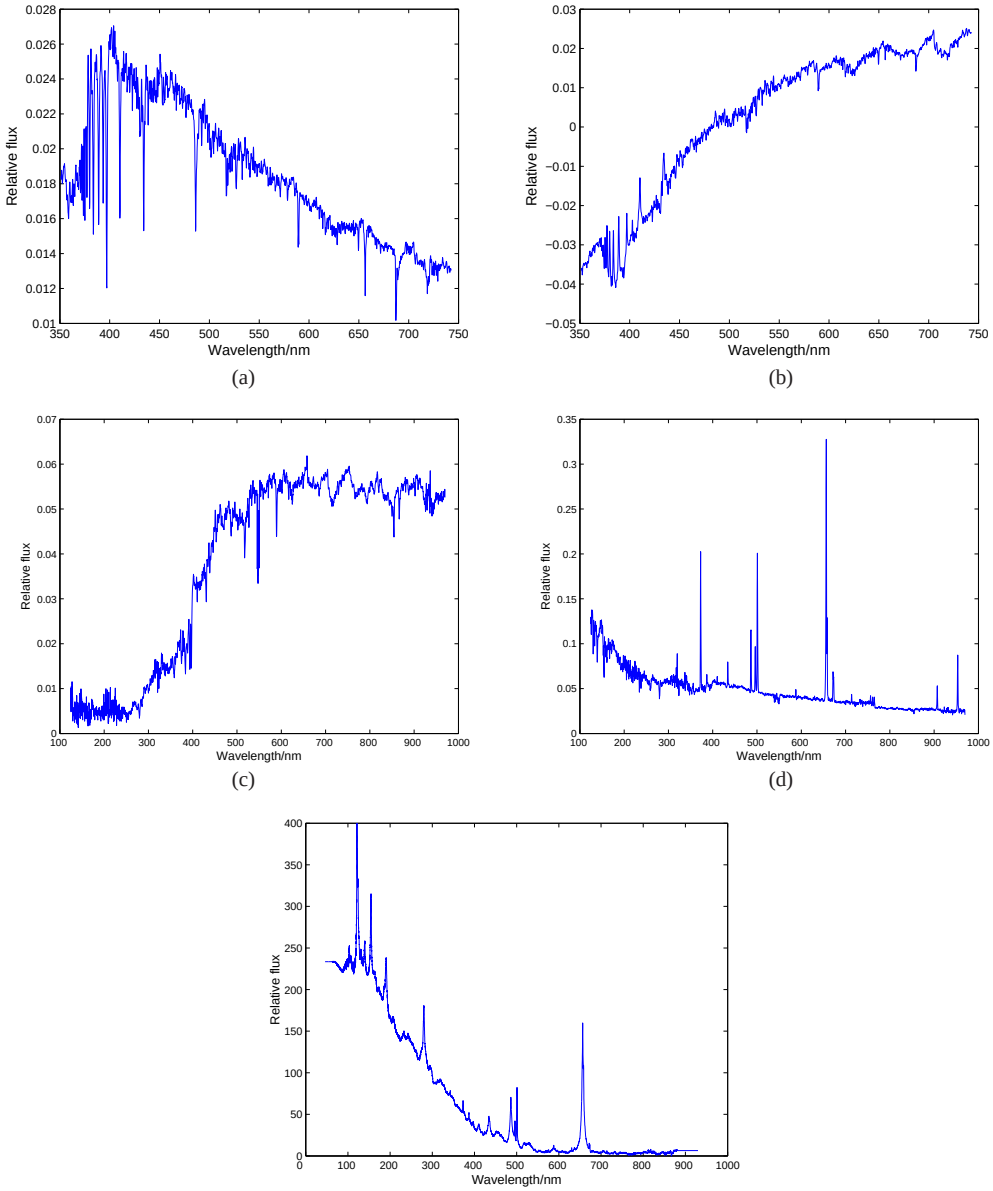


Fig. 1 Template spectra: (a): stellar template 1; (b): stellar template 2; (c): galaxy template 1; (d): galaxy template 2; (e): quasar template.

We call the spectrum after preprocessing the line spectrum. The line spectrum associated with the spectrum in Figure 2(a) are shown in Figure 2(b).

3.2 Spectral Line Extraction

Spectral lines are key features used in spectral analysis. Because of the rough continuum fitting or the high noise of the celestial spectra, usually there are many spurious spectral lines in the line spectra. It

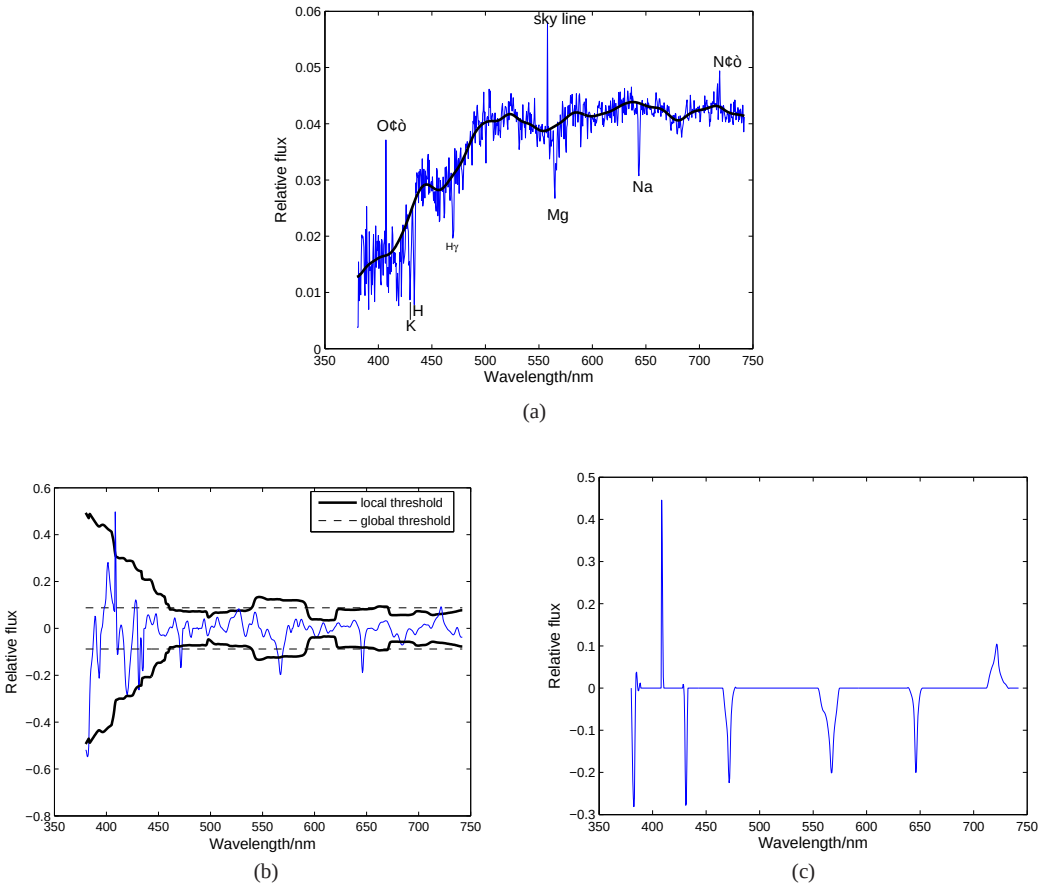


Fig. 2 Preprocessing and feature extraction: (a): original spectra and the continuum; (b): spectral line extraction; (c): spectral line candidates.

is not easy to extract all spectral lines from line spectra. Spectral line extraction in this study includes two steps: Firstly, the feature wavelength of the spectral line candidate is acquired via a one by one search following the thresholding of the line spectrum; Secondly, the start and end wavelengths of the line candidates are located by a local search. In the following, two constraints, which are threshold constraint and shape constraint, are employed for spectral line extraction.

Since the intensity at the feature wavelength of each spectral line is different, it is difficult to choose an appropriate universal threshold. If the threshold is too high, it is possible that no spectral line can be obtained. Conversely, if the threshold is too low, many spurious spectral lines will be obtained. According to the process which forms spectral lines, we know the intensity at the feature wavelength of a spectral line is locally maximal. Therefore, the threshold constraint, which is the conjunction of the local thresholding and universal thresholding defined by Equation (1), is adopted here.

$$s(i) = \begin{cases} s(i), & |s(i)| > T(i) \& |s(i)| > T_0, \\ 0, & \text{else}, \end{cases} \qquad i = 1, 2, \cdots, N, \tag{1}$$

where $s(i)$ is the spectral intensity at the ith point, $T(i)$ is the local threshold at that point, and T_0 is the universal threshold that is the lower limit for intensity at the feature wavelength of a spectral line candidate. For each point in the line spectrum, we take a fixed-width window centered on the point and

set $T(i) = c*$ RMS, where RMS denotes the root mean square of the intensities of all the points in the window and c is a constant. The first step is only for obtaining the feature wavelength of the spectral line candidates, so the setting of the local thresholds is not strict due to the sparse distribution of spectral lines. We found in experiments that the feature extraction was stable with the window width varying from 25 nm to 100 nm and the coefficient c varying from two to three. The universal thresholding tries to set a lower limit so as to exclude some weak spurious lines which appeared in the local thresholding, hence, the value of the universal threshold is also not rigid. In all the experiments below, the window width is 50 nm, $c = 2.5$ and $T_0 = 1*$ rms, where rms denotes the RMS of the array$\{s(i), i = 1, 2, \cdots\}$.

Since the continua catch the low frequency nature of the spectra, the intensities at the two boundaries of a spectral line on the line spectra should be lower than the intensity at the center of it. This shape constraint is employed in the second step to remove some spurious lines.

The thresholds are shown in Figure 2(b) and the spectral line candidates are shown in Figure 2(c).

4 SPECTRAL RECOGNITION

Let $L = (\lambda, t, \ldots)$ denote a spectral line candidate of the observed spectrum, and $L' = (\lambda', t', \ldots)$ denote a rest spectral line of the astronomical spectra, where λ and λ' denote wavelengths, and t and t' denote line types (1 for emission line and –1 for absorption line). Set

$$z = t\lambda/t'\lambda' - 1. \tag{2}$$

We use Equation (2) to determine the redshift candidates of the observed spectrum. According to the meaning of the redshift, $t = t', \lambda >= \lambda'$ must be true. Therefore, a redshift candidate must be nonnegative.

Let $d-$dimensional vector $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote spectra A and B respectively. We evenly divide the two spectra into K parts, that is, $\boldsymbol{X} = [X_1, \ldots, X_K]$ and $\boldsymbol{Y} = [Y_1, \ldots, Y_K]$. Let

$$r_{AB} = \sum_{i=1}^{K} w_i \alpha_i, \tag{3}$$

where $\alpha_i = (X_i, Y_i^T)/(\|X_i\|\|Y_i\|)$, w_i denotes the weight and is subject to $\sum_{i=1}^{K} w_i = 1$.

In the following, r_{AB} is used to measure the similarity between spectra A and B. The procedure of spectral classification using template matching is as follows:

Step 1: Perform the continuum subtraction for all the template spectra.
Step 2: Perform preprocessing and spectral line extraction for the observed spectrum.
Step 3: Determine redshift candidates of the observed spectrum using the spectral line candidates and the rest spectral lines of the astronomical spectra.
Step 4: For each redshift candidate, the galaxy and quasar templates are shifted according to the candidate. Measure the similarity between the shifted templates and the observed spectrum after sky and continuum subtraction.
Step 5: Measure the similarity between the stellar templates and the observed spectrum after sky and continuum subtraction, and set the corresponding redshift candidate to be zero.
Step 6: Choose the redshift candidate with the highest similarity as the redshift, and the type of the corresponding template as the class of the observed spectrum.

Generally, galaxies have smaller redshifts, so in step4, we only need to compare redshift candidates larger than one with quasar templates. Moreover, the redshift of stars is so small that it can be ignored. In measuring the similarity between the templates and the observed spectrum, the template spectra must be linearly interpolated to the same wavelength range and have the same bin size as those of the observed spectrum.

If K is equal to one in Equation (3), the proposed similarity measure becomes the traditional one. Due to the high local correlation caused by high noise or by two strong lines corresponding to some

redshift candidates, it is possible that the similarity corresponding to the redshift will be lower than those of other candidates if the similarity is measured traditionally. Similar to evidence accumulation, the proposed similarity measure is the weighted sum of several pieces of similarity evidence. The principle of setting the weights is that the larger the value of α_i is, the higher the corresponding weight w_i. Hereafter, we will call the proposed method LMCC.

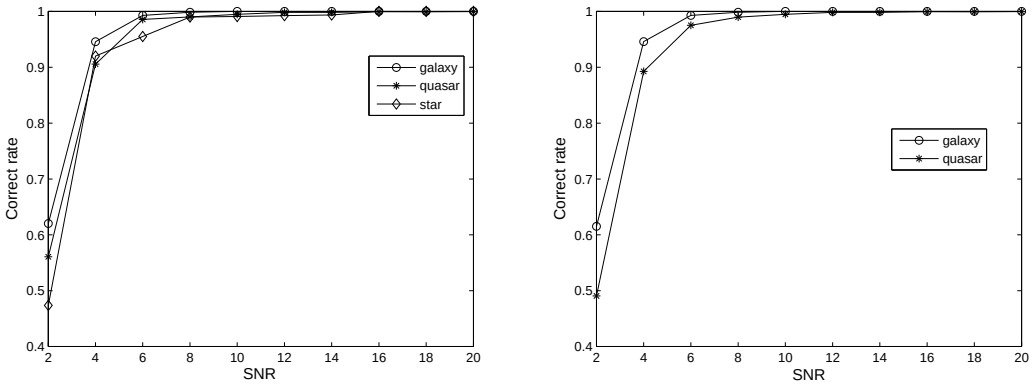
5 EXPERIMENTS AND RESULTS

In this section, both simulated spectra and observed spectra are used to verify the effectiveness of LMCC. In Section 5.1 and 5.2, the spectrum is divided into four segments for the similarity measure, and the weight is set to be $0.1 : 0.2 : 0.3 : 0.4$.

5.1 Simulated Spectra

The simulated spectra are generated as follows: shift the eleven galaxy templates presented by Kinney & Calzetti et al. (1996), respectively with redshifts ranging from 0 to 0.5 by a step of 0.01; shift the quasar template with redshifts ranging from 0 to 5 by a step of 0.01. All simulated spectra are linearly interpolated between the wavelength range of $380 - 742$ nm by a bin size of 0.5 nm. The 161 stellar spectra mentioned in Section 2 are chosen as the test data of the stars. Therefore, the overall simulated data include 561 galaxy spectra, 501 quasar spectra and 161 stellar spectra.

A Gaussian noise with zero mean and standard deviation (noise level σ) is added to these spectra. The correct classification rate vs. SNR (SNR=$1/\sigma$) is plotted in Figure 3(a). We also show the correct rates for redshift estimation in Figure 3(b), where the error accuracies are 0.01 and 0.001 for quasars and galaxies respectively. All the results in Figure 3 are the average of one hundred independent trials. It can be seen from the figure that the correct classification rates increase with the increase of SNR and all the correct classification rates reach 95% or above, at the SNR of six. This indicates that the proposed technique is robust and effective.



(a) (b)

Fig. 3 Correct rate vs. SNR. (a) classification; (b) redshift determination.

5.2 Observed Spectra

The data set from SDSS DR2 includes 2509 observed galaxy spectra from sky regions 0271–0275, 1878 stellar spectra from sky regions 0266–0295 and 3026 QSO spectra from about fifty sky regions. The redshifts of these spectra vary from zero to six.

Since SVM is often used in spectral classification, we compare the proposed method with SVM in this section. The program which we use is the SVM$^{\rm light}$ (Joachims 1999), and the kernel function is Gaussian. Two classifiers are constructed for the classification by SVM. The first classifier is used to recognize the stellar spectra, and the second one is used to classify the galaxy and QSO spectra. The data are divided into two parts evenly, where one part is the training set and the other is the test set. We extract 20-dimensional features from the spectra by PCA as the recognition features of SVM. The five-fold cross-validation (Stone 1974) is used to determine the kernel width of SVM. The resultant kernel widths in the two classifiers are 0.1 and 0.2 respectively.

Correct recognition rates for the three classes of spectra are shown in Table 1. It can be seen from Table 1 that the recognition rate of the stellar spectra by LMCC is lower than that by SVM, while the recognition rates of the galaxy and QSO spectra by LMCC are higher than those by SVM.

Table 1 Correct Rates of Classification (%)

Method	Star	Galaxy	QSO
SVM	96	94.3	95.06
LMCC	92.9	97.9	98.8

From analysis of the misclassified spectra in LMCC, we find that the misclassification mainly includes two cases: one is that the stellar spectra are misclassified as galaxy or QSO, the other is that the galaxy spectra are misclassified as QSO or the inverse. Because of noise contamination, many misclassified spectra deviate far from their templates. In particular, many stellar spectra with strong noise are very close to the galaxy spectra. It is possible that these spectra can be misclassified by LMCC. However, different from template matching, SVM is a supervised learning algorithm, which exploits the features of the training samples to learn the classifier. The galaxy and QSO share many emission spectral lines, so many galaxy and QSO spectra are difficult to separate by SVM. Moreover, the selection of the training set is crucial for SVM. Overall, the proposed method is superior to SVM in spectral classification.

5.3 Different Similarity Measure Strategies

In order to study the influence of the parameter K (segment number), we perform the redshift estimation of QSOs. The data are the 3026 spectra described in Section 5.2, whose redshifts are from SDSS. The spectra are evenly divided into $K = 1-7$ segments, and the weights are chosen as $1/K$ in each similarity measure respectively. The correct rates for different matching accuracies ε are shown in Table 2. From Table 2, we can see that for each accuracy ε, the correct rate is increased with K varying from 1 to 3, while it is decreased with K varying from 4 to 7. The result of using the traditional similarity measure ($K = 1$) is the weakest. This indicates that the proposed similarity measure is more effective than the traditional one.

Table 2 Correct Rates (%) by Different Similarity Measure Strategies

Accuracy	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
$\epsilon = 0.01$	83.5	86.1	87.4	87.1	86.7	85.9	85.1
$\epsilon = 0.02$	89.5	91.8	93.1	93.2	92.5	92.1	91.7
$\epsilon = 0.03$	90	92.4	94	94	93.5	93.1	92.5

6 CONCLUSIONS

Automated spectral recognition is very important in large redshift surveys. In this study, a novel template matching approach for spectral classification is presented. We firstly use spectral line matching to determine the redshift candidates, and then estimate the spectral class by measuring the similarity

between the observed spectrum and the templates shifted by each redshift candidate. The similarity is measured by the weighted sum of several similarity evidences. Compared with some spectral classification methods based on learning such as neural networks, SVM, etc, the proposed technique needs no training. It is well known that the training process is time-consuming and susceptible to the selection of training set. As a byproduct of this approach, the spectral redshift can also be estimated. Both the simulated data and observed spectra are used to test this approach, and the results show that the proposed method is efficient, and it can achieve a correct classification rate as high as 92.9%, 97.9% and 98.8% for stars, galaxies and QSOs, respectively.

Acknowledgements This work was partially funded by the Natural Science Foundation of China (NSFC) (grant Nos. 60773040 and 60872127) and Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality PHR (IHLB).

References

Bai, L., Guo, P., & Hu, Z. Y. 2005, ChJAA (Chin. J. Astron. Astrophys.), 5, 203
Bailer-Jones, C., Irwin, M., & von Hippel, T. 1998, MNRAS, 298, 361
Connolly, A. J., Szalay, A. S., Bershady, M. A., et al. 1995, AJ,110, 1071
Donoho, D. L. 1995, IEEE Trans.on IT, 41, 613
Jacoby, G. H., Hunter, D. A., Christian, C. A.,1984, ApJS, 56, 257
Joachims, T. 1999, Advances in Kernel Methods - Support Vector Learning, MIT-Press
Kinney, A. L., Calzetti, D., Bohlin, R. C., et al. 1996, ApJ, 467, 38
Galaz, G., & Lapparent, V. 1998, A&A, 332, 459
Qin, D. M., Guo, P., Hu, Z. Y., et al. 2003, ChJAA (Chin. J. Astron. Astrophys.), 3, 277
Stone, M. 1974, Journal of the Royal Statistical Society, 36, 111
Zaritsky, D., Zabludoff, A. I., & Willick, J. A. 1995, AJ, 110, 1602
Zhang, Y., & Zhao, Y. 2004, A&A, 422, 1113
Zhang, Y., & Zhao, Y. 2003, PASP, 115, 1006
Vanden Berk, D. E., et al. 2001, AJ, 122, 549
von Hippel, T., Storrie-Lombardi, L. J., Storrie-Lombardi, M. C., et al. 1994, MNRAS, 269, 97