IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 28, 2020, accepted March 23, 2020, date of publication March 27, 2020, date of current version April 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2983745

Rare Object Search From Low-S/N Stellar Spectra in SDSS

MINGLEI WU[1], JINGCHANG PAN[1], ZHENPING YI[1], AND PENG WEI[2]

[1]School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China
[2]Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China

Corresponding author: Jingchang Pan (pjc@sdu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant U1931209, Grant 11603012, Grant 11873037, Grant 11603014 and Grant 11803016.

ABSTRACT Rare objects such as white dwarf+main sequence (WDMS) and cataclysmic variables (CVs) are very important for studying the evolution of the galaxy and the universe. The large amount of spectra obtained by the large sky surveys such as the Sloan Digital Sky Survey (SDSS) are rich sources of these rare objects. However, a considerable fraction of these spectra are low-S/N spectra. These low-S/N spectra contain similar useful information as the high-S/N spectra, and making better use of these spectra can significantly improve the chance of finding rare objects. Nevertheless, little research has been done on them. In this study we propose a novel method based on the combination of PCA (Principal Components Analysis) and CFSFDP (Clustering by Fast Search and Find of Density Peak) to search for rare objects from low-S/N spectra. The PCA first extracts principal components from high-S/N spectra to generate general feature spectra and reconstructs low-S/N stellar spectra with these general feature spectra. Then the CFSFDP calculates the Local Density ρ and the Distance δ of the reconstructed spectra, and select the outliers through the decision graph quickly and accurately. We first apply our method to spectra in SDSS stellar classification template library with adding white gaussian noise to search for rare objects (carbon stars, carbon white dwarfs, carbon_lines, white dwarfs and white dwarfs magnetic). Then we apply our method to observed spectra with different low-S/Ns from SDSS and compared with Lick-index+K-means and Support Vector Machines (SVM). The experimental results show that our method has a higher efficiency compared to other methods.

INDEX TERMS SDSS, stellar spectra, machine learning, rare object search.

I. INTRODUCTION

Detecting and analyzing rare objects are important for studying the Galactic and extra-galactic structure and evolution. Cataclysmic variables (CVs), for example, is critical to understand many astrophysical problems such as black holes and supernovae type Ia [1]. Multi-object spectroscopy has greatly improved the observation efficiency and large data sets have been obtained from wide-field spectroscopic surveys. Such large data sets provide us with more samples to search for rare objects.

Traditional methods of searching for rare objects are mainly colour-colour (CC), colour-magnitude (CM) diagrams and atmospheric parameters, and so on, where various types of objects (like stars and galaxies) appear in separate areas due to differences in observed colours and parameters [2]. CVs are close binary systems that is composed of a white dwarf and a late-type main-sequence companion. They are central to our understanding of compact binary evolution. P. Szkody et al have identified 285 CVs from SDSS Early Data Release (EDR) with using photometric criteria [3]. Since CVs white dwarfs are relatively hot objects (Teff $\leq$10 000 K), A. F. Pala et al used effective temperatures of white dwarfs as a probe of cataclysmic variables evolution [4]. M. R. Kennedy et al presented X-ray emission-line measurements to verify a CVs [5]. R. Ridden-Harper identified CVs candidates from the K2/Kepler Campaign 11 field equivalency by using width (EW) and light curve [6]. A. S. Oliveira et al found 13 previously unreported CVs by exploiting magnitudes from CRTS [7]. J. J. Wallace et al extract light curves for 4554 objects from K2 superstamp observations and found 72 new variables [8]. A white dwarf (WD) primary and a main-sequence (MS) companion star are referred to as

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

white dwarf+main sequence (WDMS) binaries, which is considered to be possible progenitors of Type Ia supernova and important for cosmological studies. J. J. Ren et al presented the data release (DR) 5 catalogue of WDMS from LAMOST. The catalogue contains 357 WDMS that have not been published before [9]. F. M. Jimenez-Esteban presented a catalogue of 73 221 white dwarf candidates extracted from Gaia-DR2 catalogue by using the astrometric and photometric data [10]. M. Perpinya-Valles et al reported the discovery of J1953–1019, the first resolved triple white dwarf system based on Gaia DR2 photometry and astrometry combined with the follow-up spectroscopy [11], and many other authors also detect the WDMS from various astronomical data [12]–[16]. In addition, Lei et al identified 294 hot subdwarf stars from LAMOST DR5 [17] and 182 single-lined hot subdwarf stars by using atmospheric parameter and spectral lines from LAMOST DR6 and DR7 [18]. Wang *et al.* proposed a method based Lick-index for excavation and analysis of outliers from massive stellar spectra [19]. Li *et al.* used spectral lines to find six new Oe stars and four new B0e stars in LAMOST DR5 [20].

Today's various sky survey projects contain the order of a million catalogued sources each, which means that the traditional ways of dealing with the resulting catalogues by direct human inspection are not practicable. Therefore, machine learning methods have emerged for searching for rare objects from these survey projects [21]–[28].

The quality of the spectra obtained by large sky surveys such as SDSS and LAMOST [29]–[32] has been significantly improved. However, the low-quality spectra still occupy a considerable proportion. These spectra show significant quality defects such as large noise, unobvious spectral line features, low signal-to-noise ratio (S/N), continuum anomalies, and splicing anomalies. Among them, low-S/N spectra account for a great part. The processing and analysis of these low-S/N spectra are of great significance for the improvement of spectral utilization, multi-band cross-validation and the discovery of rare celestial bodies.

Due to the difficulty of analyzation, little work has been done on the low-S/N spectra. Difficulties mainly include two aspects: (1) the features of the low-S/N spectra are usually hidden in the background noise, which are difficult to detect, and they are easy to fall into overfitting when training low-S/N samples and the accuracy will be affected; (2) good method to evaluate the results of denoising is absent. To deal with the first problem, we randomly select high-S/N spectra covering varies subclasses of the spectra from SDSS DR14. Then we apply PCA to these spectral samples to build a general feature spectral library, and utilize the general feature spectral library to reconstruct the low-S/N spectra. To deal with the second problem, we collect a set of stars each of which has both high-S/N and low-S/N spectra. We apply the PCA to the low-S/N spectra to reduce the noise, and then compare it with the corresponding high-S/N spectra to evaluate the effect of denoising. In addition, most of existing studies are supervised methods that can only detect some specific types of rare objects at a time, and they require large time and space complexity. As an unsupervised method, our method is faster than other methods due to avoiding the selection of cluster centers.

The structure of the paper is as follows. Section II reviews the related work about rare objects in astronomy. Section III introduce the method and principle used in the paper. In Section IV and Section V, we discuss the implementation of the method. We use this method to search for CVs and compare to lick-index+K-means and Support Vector Machines (SVM) in Section VI. Finally, we give our conclusions and the application prospect of this method in Section VII.

II. RELATED WORK

Rare objects search based on machine learning in astronomy could be summarized in two aspects: supervised learning and unsupervised learning. In this section, we summarize the recent research trend and describe some approaches that are usually used for rare objects search in astronomy.

A. SUPERVISED LEARNING

Supervised learning algorithms that employ existing label data to learn feature efficiently that can be used to detect outliers, like Line Regression, Artificial Neural Networks (ANN), SVM and Random Forest. In supervised learning, most algorithms detect rare objects by measuring uncertainty and classification probability [33], [34].

Linear Regression method is one of the simplest methods, but it can only capture linear features. ANN are a set of algorithms with structures that are vaguely inspired by the human brain. The flexible structure and non-linearity of ANN make it popular in Astronomy [35]–[37]. However, ANN have many hyper-parameters, and they usually require a large amount of data to train on. SVM is one of the most popular supervised learning for outlier detection, especially one-class SVM. However, one-class SVM is suitable for data with a handful of features, so it can be only applied to derived features of astronomical observations such as images, light-curves, or spectra. In addition, the kernel shape and free parameters need to be chosen for the resulting decision function [20], [38], [39]. Random Forest is trained to predict the class of previously unseen objects according to the classification probability [21]. For example, an object that is classified as a star with a probability of 0.65 is probably more anomalous than an object that is classified as a star with a probability of 0.85. Therefore, the outliers are usually defined as ''shout the loudest'' [40]. The Random Forest generalizes well to previously unseen data. The main disadvantage of Random Forest is its inability to take into account feature and label uncertainties.

B. UNSUPERVISED LEARNING

Unsupervised learning of rare object search can be roughly divided into clustering analysis and dimensionality reduction.

In clustering analysis, one can define outliers that have a large distance from other objects. In general, Euclidean distance is used as metric, but, in some cases, it might not result in an optimal performance. Therefore, other metrics should be considered, such as the unsupervised Random Forest based distance [41], cross correlation based distances [42] and cosine distance. K-means method is one of the most widely used clustering methods. It is often used in astronomy to study stellar and galaxy spectra, X-ray spectra, asteroids spectra, and so on [18], [43], [44]. However, K-means is not optimized to detect outliers, which needs to randomly select k objects as the initial centroids and repeat iteratively until reaching convergence. In addition, there are also some other clustering algorithms that have been applied to various astronomical problems, such as Hierarchical clustering and Gaussian Mixture Models [45]–[48].

In astronomy, another way of outlier detection is dimensionality reduction. Principal component analysis (PCA) has been widely used in the field of astronomy. Whitney *et al.* used PCA for spectra classification and error analysis [49], [50]. Singh *et al.* and Qin *et al.* used PCA to perform stellar spectra classification, respectively [51], [52]. Williamson *et al.* used PCA to detect outliers in stripped-envelope SN spectra [53]. While PCA is widely in astronomy, it allows for negative values in the feature spectra. However, in most astronomical applications, negative feature spectra are not physical, because we only have a sum of positive flux contribution from different sources, and there is no negative flux. Therefore, in many cases astronomers use nonnegative matrix factorization (NMF), instead of PCA [40]. However, PCA usually outperforms the NMF by a large margin when applied to outlier detection. In addition, auto-encoder and tSNE are also popular dimensionality reduction methods for detecting outliers [54].

Although many machine learning methods are used to search for rare objects in astronomy, there are little specific approach used to search for rare objects from low-S/N spectra.

III. METHODOLOGY

In this section, we propose a hybrid method of PCA and CFSFDP as our proposed method. This method first introduce the PCA to reconstruct the low-S/N spectra, and then use the clustering method called CFSFDP to detect rare objects from these reconstructed spectra.

A. SPECTRA RECONSTRUCTION BASED ON PCA

PCA uses orthogonal conversion to transform a set of potentially related variables into a set of linearly independent variables (called principal components) [55].

We suppose $Z = (Z_1, Z_2, \cdots, Z_n)$, where $Z_i^T = (z_{i1}, \cdots, z_{im})$. We an compute the matrix

$$Z_S = BZD^{-1}, \tag{1}$$

where

$$B = E - \frac{1}{n}D_0 \tag{2}$$

is the centering matrix, E is an identity matrix, and D_0 is a matrix with all of its elements equal to 1; D is a diagonal matrix

$$D = diag\{||\overline{GZ_{(1)}}||, \cdots, ||\overline{GZ_{d)}}||\} \tag{3}$$

i.e. its diagonal element is $||\overline{GZ_{(i)}}||$ $(i = 1, 2, \cdots, d)$, where

$$\overline{Z}_{(i)} = \begin{pmatrix} z_{1i} \\ z_{2i} \\ \vdots \\ z_{ni} \end{pmatrix} \tag{4}$$

and $\overline{G} = (1, \cdots, 1)$. Then we can renew Z with Z_S. After renewing Z, we need to compute the eigenvalues of covariance matrix Z^TZ, whereZ^T is the transpose of the matrix Z. We let $E_{m\times m}$ be the eigenvector matrix of the convariance matrix Z^TZ. Then the principal component matrix P can be obtained by

$$P = ZE. \tag{5}$$

We now show how to obtain the reconstruction of data by using the first k principal components. From Eq. (5) we have $Z = PE^{-1}$. If we set the last $m - k$ rows of E^{-1} to be zero, we can obtain a matrix E^*. We let $Z^* = PE^*$. Then Z^* is the reconstructed matrix of Z using the first k principal components. If we define p_{ij} to be the element in the ith row and jth column of P, Z_i^* to be the ith row of X^* and the jth row of E^*, and then we have $Z_i^* = \sum_{j=1}^{k} p_{ij}E_j^*$

B. CFSFDP: NEW FAST CLUSTERING METHOD BASED ON DENSITY PEAKS

Rodriguez proposed a method for fast clustering based on density peaks [56]. The method considers that cluster centers should have the following two characteristics:

(i) The density is higher than adjacent points

(ii) The distance from points with higher density is far.

We suppose x_i $(i = 1, 2, \cdots, n)$ are from the sample set S, and define two variables ρ_i and δ_i for these data

Definition 1: (Local Density ρ_i): The Local Density ρ_i is defined as follows

$$\rho_i = \sum_{j\in I_s/\{i\}} \varkappa(d_{ij} - d_c) \tag{6}$$

where the parameter d_c is the cutoff distance, and the function $\varkappa(x)$ is

$$\varkappa(x) = \begin{cases} 1 & x < 0 \\ 0 & x \geq 0 \end{cases} \tag{7}$$

It can be seen from Eq. (6) that ρ_i denote the number of points whose distance from x_i is less than d_c.

Definition 2: (Distance δ_i): Let $\{q_i\}_{i=1}^n$ be the indices of $\{\rho_i\}_{i=1}^n$, such that

$$\rho_{q_1} \geq \rho_{q_2} \geq \cdots \geq \rho_{q_N} \tag{8}$$

Then the distance δ_i is defined by

$$\delta_i = \begin{cases} min\{d_{q_i q_j}\} & i \geq 2 \\ \max\limits_{j \geq 2}\{\delta_{q_j}\} & i = 1 \end{cases} \quad (9)$$

In this way, (ρ_i, δ_i) can be calculated for each data point x_i in S, and then the decision graph is drawn with the horizontal axis ρ and the vertical axis δ. Then the points located at upper right corner of the decision graph are the cluster centers, and the points located as the upper left corner of the decision graph are the outliers [56].

C. THE PROPOSED METHOD

We propose a method to search for rare stellar spectra from low-S/N stellar spectra by synthesizing the above two methods. The steps are as follow:

(i) For all low-S/N stellar spectra, spectra reconstruction is performed using the PCA. In this paper, the general feature spectra are extracted from various types of high-S/N spectra by PCA, and then used to reconstruct various types of low-S/N spectra.

(ii) The local density and distance of these reconstructed spectra are calculated according to the CFSFDP method. Then we draw a decision graph based on density and local distance, and the points on the upper left will be considered as rare objects. As described in the original literature [56], one can choose distance δ so that the average local density ρ is around 1% to 2% of the total number of points in the data set. However, for the distance that is used to select outliers, there is no exact discussion in the original literature, so it is a hyperparameter. In this paper, we take 1/30 of the average value of ρ to select outliers through experiment.

IV. SPECTRAL DENOISING EXPERIMENT

A. THE DATA

All the experimental spectra are taken from SDSS DR14. The SDSS is currently one of the largest spectroscopic surveys, which started observations in 1998 and has completed three different phases. The fourth phase, named SDSS-IV, is now underway [57]. SDSS DR14 is the second release of the SDSS-IV. It has provided more than 2.54 million spectra, including 928 859 stellar spectra. Among them, the spectra with S/N≤10 account for about 20%. There have been a number of research around them [58]–[62].

Our data consists of two sets. The first set comprise the template spectra, which are taken from SDSS stellar classification template library. It contains a total of 36 template spectra built on the SDSS observed spectra (see Table 1). Each template spectrum covers a logarithmic wavelength range from 3.5781 to 3.9672 with a sampling resolution of 0.0001. The second set consists of the observed spectra selected from SDSS DR14. The detailed information of these spectra are listed in Table 2.

The observed spectra are from the SDSS DR14. The detailed information of these data sets are listed in Table 2 and Table 3. Among them, the data in Table 2 is used for extracting general feature spectra and building the general feature spectral library, and the data in Table 3 is used for search experiments. Based on PCA, the following steps are used to build the general feature spectral library and denoise low-S/N stellar spectra:

TABLE 1. The spectral type of the SDSS template library.

Spectral Type	Subclass	Number
Main sequence	O, B,A,F,G,K,M	31
WD	WD	1
WD magnetic	WD magnetic	1
Carbon	Carbon	1
Carbon_lines	Carbon_lines	1
Carbon WD	Carbon WD	1

TABLE 2. The spectral type of observed spectra library for extracting feature spectra.

Subclass	Number	Proportion
O (S/N≥30)	215	10.02%
B (S/N≥40)	231	10.77%
A (S/N≥95)	205	9.56%
F (S/N≥105	291	13.57%
G (S/N≥95)	267	12.45%
K (S/N≥105)	154	7.18%
M (S/N≥60)	223	10.40%
Carbon (S/N≥30)	157	7.32%
CV (S/N≥45)	144	6.71%
WD (S/N≥45)	176	8.21%
WD magnetic (S/N≥20)	34	1.59%
L (S/N≥20)	48	2.24%

(i) All the spectra are selected.

(ii) The wavelength is unified to 3800-9000 $\AA$ with a step size of 1 $\AA$ (the total number of sampling points is 5201) [63] and the flux after interpolation is obtained.

(iii) PCA is performed on all fluxes, and the first k feature spectra with a cumulative variance contribution rate (*CVCR*) exceeding 99.99% are selected. The *CVCR* is defined as follows:

$$CVCR = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{n} \lambda_j} \quad (10)$$

where $\sum_{i=1}^{k} \lambda_i$ represents the first k feature spectra and $\sum_{j=1}^{n} \lambda_j$ represents all feature spectra. (iv) The new spectra is denoiseded using the first k feature spectra obtained in (iii).

1) THE GENERAL FEATURE SPECTRAL EXTRACTION AND RECONSTRUCTION OF TEMPLATE SPECTRA

According to the above steps, 36 spectra in the SDSS stellar classification template library are taken as input. After PCA, the first 18 feature spectra with variance contribution ratio over 99.99% are selected (see Fig. 1). Owning to the space limitation, only the first 6 feature spectra are shown in Fig. 2. Various spectra contained in the template library can be reconstructed by using the general feature spectra obtained from the above steps. In order to compare the reconstruction effect, the first 1, 3, 9 and 18 feature spectra are selected to reconstruct two randomly selected template spectra(A0, F5).

TABLE 3. The SDSS observed spectra used for searching for rare objects.

Data Set Name	S/N	Number of normal spectra[a]	Number of rare spectra[b]
Data Set 1	1-2	2220	12
	2-3	3374	12
	3-4	2039	12
	4-5	3788	12
Data Set 2	1-5	9815	8
	11-15	12064	6
	31-35	12108	7
Data Set 3	1-2	7167	8
	2-3	14671	6
	3-4	15057	6
	4-5	11936	8

[a] The normal spectra refer to O, B, A, F, G, K, M-type.
[b] In Data Set 1, the rare spectra refer to Carbon, Carbon WD and CVs, but it refers to CVs in Data Set 2 and Data Set 3.

The results are shown in Fig. 3. It can be seen that the residual spectrum hardly contains any feature, so the effect of reconstruction is excellent.

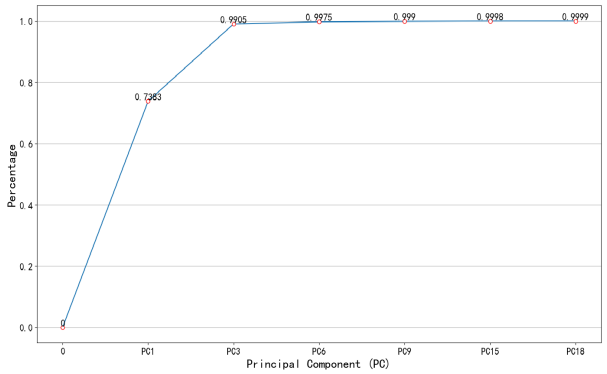


FIGURE 1. The first 18 principal components of SDSS template spectra and their variance contribution ratio.

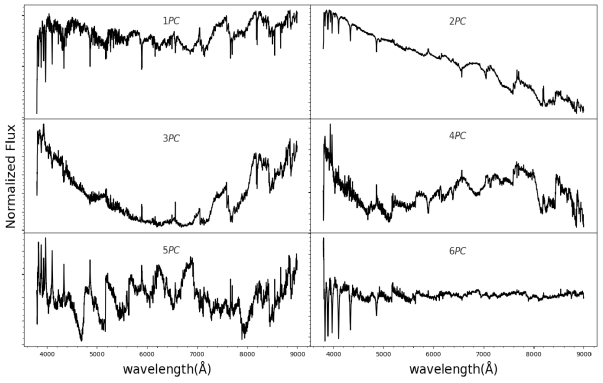


FIGURE 2. The first 6 feature spectra of 36 SDSS template spectra. The first row represents the first 2 feature spectra, the second row represents the 3rd and 4th and the third row represents the 5th and 6th feature spectra.

In order to verify the denoising effectiveness on low-S/N spectra, the above two template spectra (as shown in Fig. 3) are selected, and added Gaussian white noise such that S/N=1, 3, 5, 7, 9 and 11. As shown in Fig. 4 and Fig. 5, the red line is the original spectra, the grey background is the spectra after adding noise, and the blue line is the denoised spectra. The comparison of red line and the blue line shows that this method is very effective to acquire high-S/N spectra through denoising the corresponding low-S/N ones.

2) THE GENERAL FEATURE EXTRACTION AND RECONSTRUCTION OF OBSERVED SPECTRA

Traditional PCA performs feature extraction on only a specific type of spectrum, and only reconstructs this type of spectrum using corresponding features. However, our method performs feature extraction on randomly selected various types of high-S/N spectra from SDSS DR14, so our feature spectra can be used as the general feature spectra library for denoising low-S/N spectra. Since we extract feature spectra from the high-S/N spectra, we can retain the spectral information as much as possible.

We randomly select 2 144 stellar spectra of various types with S/N≥20 from SDSS DR14 as the observed spectra. Table 2 shows their subclass distribution and Fig. 6 shows their S/N distribution. The first 221 feature spectra, whose cumulative variance contribution rate exceeds 99.99%, are selected to build the general spectral feature library (see Fig. 7). Due to the space limitation, only the first 6 feature spectra are shown in Fig. 8. In order to verify the effectiveness of PCA for reconstructing real low-S/N spectra, we apply it to the following two sets of spectra randomly selected from SDSS DR14.

One set contain five different low-S/N spectra with S/N=1.80, 3.71, 5.53, 7.60, 9.59, and 11.69 (see Table 4). After denoising, they are compared with the corresponding template spectra. Due to the inaccuracy of subclass in the low-S/N spectra, we select the corresponding template spectra by SSE (sum of squared errors):

$$SSE = \sum_{i=1}^{N} (F_{Ri} - F_{Ti})^2 \tag{11}$$

where FR is the fluxes of the reconstructed spectra, FT is the fluxes of the template spectra, and N is their wavelength with

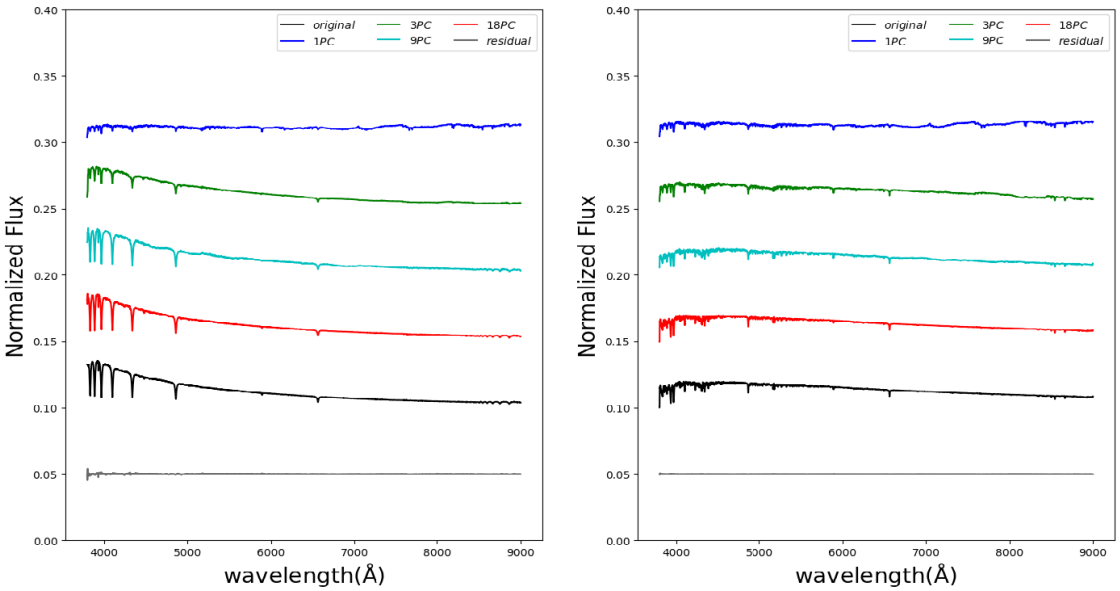


FIGURE 3. Reconstruction comparison. In both left and right panels, the reconstruction with the first one (solid blue curves), three (solid green curves), nine (solid cyan curves), eighteen (solid red curves) principal components are plotted, the solid black lines are the original spectra of A0-type (left panel) and F5-type (right panel), and the soild dimgray lines are the residual spectra i.e. the resconstruction with the first eighteen principal components minus original one.

TABLE 4. Six low-S/N spectra from SDSS DR14.

Plate	MJD	FiberID	Class	Subclass	R.A.(J2000)	Decl.(J2000)	S/N
1343	52790	13	STAR	K7	16:56:23.2	30:26:20	1.80
1072	52643	268	STAR	K7	02:19:51.7	-00:18:35	3.71
1509	52942	632	STAR	K7	02:35:42.6	00:27:54	5.53
1663	52973	491	STAR	K7	23:23:37.7	53:06:10	7.60
1123	52882	305	STAR	K7	00:30:17.1	-00:25:31	9.59
1130	52669	484	STAR	K7	00:52:16.4	00:14:29	11.69

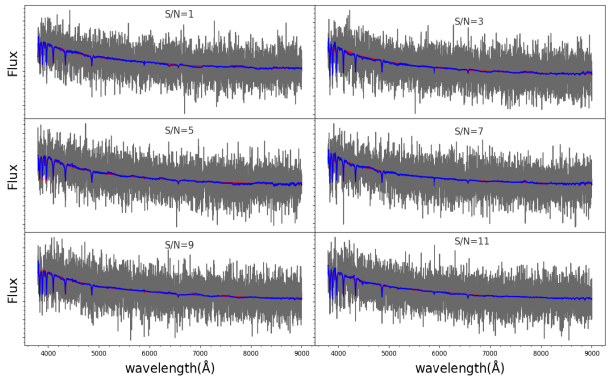


FIGURE 4. Denoising effect of A0-type spectra with different S/Ns. The first row represents the denoising effect of spectra with S/N=1 and 3, the second row represents the denoising effect of spectra with S/N=5 and 7, and the third row represents the denoising effect of spectra with S/N=9 and 11.

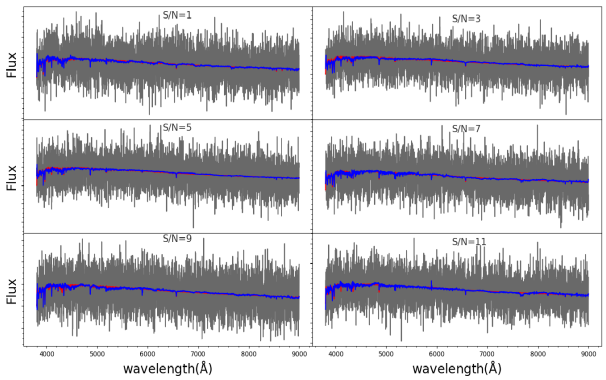


FIGURE 5. Denoising effect of F5-type spectra with different S/Ns. The first row represents the denoising effect of spectra with S/N=1 and 3, the second row represents the denoising effect of spectra with S/N=5 and 7, and the third row represents the denoising effect of spectra with S/N=9 and 11.

a fixed step of 1 Å. We choose the template spectra with the smallest SSE for comparison. Fig. 9 shows the comparison of the denoised spectra and the template spectra. The second set contains six pairs of spectra, each of which with the same right ascension (R.A.) and declination (Decl.) consists of a high-S/N spectrum and a low-S/N one (see Table 5). Many of the spectra from the Catalog Archive Server (CAS) database have been observed many times, and many spectra with different S/Ns are generated due to the variable conditions at the time of shooting. So we could compare the denoised

TABLE 5. Six groups spectra with the same R.A. and Decl. from SDSS DR14.

R.A.(J2000)	Decl.(J2000)	Plate	MJD	FiberID	Class	Subclass	S/N
03:20:59.9	00:45:06	1180	52995	549	STAR	A0p	1.99
		413	51821	545	STAR	A0	29.34
11:11:17.4	18:53:57	2872	54468	312	STAR	K7	2.49
		2872	54533	314	STAR	K7	21.38
11:21:40.2	18:36:14	2872	54468	15	STAR	B6	3.19
		2495	54533	4	STAR	B6	27.58
07:54:50.0	26:44:31	3227	54864	57	STAR	G0	4.96
		3227	54893	58	STAR	F2	21.27
00:21:20.9	00:04:03	1119	52562	497	STAR	K7	5.92
		1119	52581	486	STAR	K7	20.38
00:15:52.4	00:46:13	1119	52562	333	STAR	F5	6.47
		1542	53734	333	STAR	F5	28.67

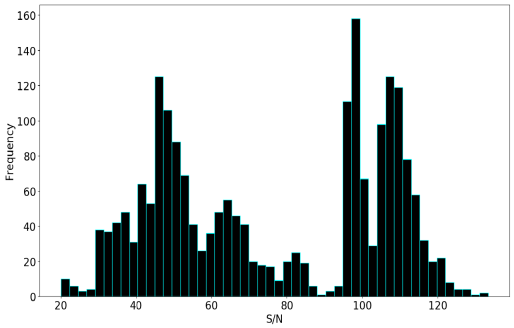


FIGURE 6. S/N distribution of 2144 stellar spectra.

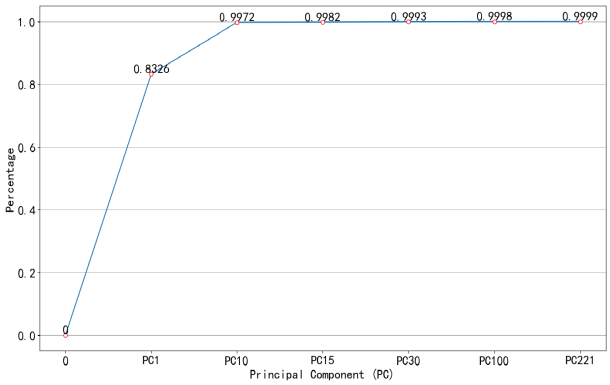


FIGURE 7. Cumulative variance contribution rate.

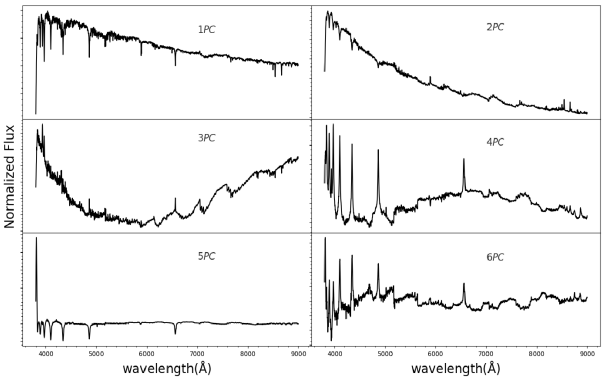


FIGURE 8. The first 6 feature spectra of 2144 SDSS stellar spectra. The first row represents the first 2 feature spectra, and the second row represents the 3rd and 4th feature spectra and the third represents the 5th and 6th feature spectra.

low-S/N spectra with the corresponding high-S/N ones to verify denoising effect. The Fig. 10 shows the comparison results. It can be seen from the results of Fig. 9 and Fig. 10 the denoised spectra are considerably less noisy and the residual spectrum contains no major features, so PCA can effectively denoise the low-S/N spectra.

V. RARE OBJECT SEARCH EXPERIMENT

In this section, we use the CFSFDP method to search for rare objects from the above two sets of spectra (template spectra and observed spectra).

A. SEARCHING FOR RARE OBJECTS FROM THE TEMPLATE SPECTRA

We randomly select O, B, A, F, G, K and M-type spectra from the above template library as normal spectra (the normal spectra refer to O, B, A, F, G, K, M-type in our experiments, except where indicated), and randomly add Gaussian white noise to simulate 1000 low-S/N spectra. Then we randomly select carbon, carbon WD, carbon_lines, WD and WD magnetic-type spectra from the above template library as rare ones, and also add Gaussian white noise to simulate 5 low-S/N spectra.

These simulated low-S/N spectra are denoised with PCA, and then these denoised ones are clustered by the CFSFDP to search for rare objects. The cutoff distance in CFSFDP is data-dependent, which is usually given by experience. So different spectra are processed with different cutoff distances. The Fig. 11 shows search results with different cutoff distances in the simulated spectra with S/N=1, and Fig. 12 shows the search results in the simulated spectra with S/N=1, 2, 3, 4. We use two metrics to measure the experiment results, that is, the recall rate (*RR*) and the candidate ratio (*CR*):

$$RR = \frac{TP}{TP + FN} \tag{12}$$

where TP represents the number of true rare objects, FN represents the number of false normal objects and $TP + FN$ represents the number of all rare objects.

$$CR = \frac{RS}{TS} \tag{13}$$

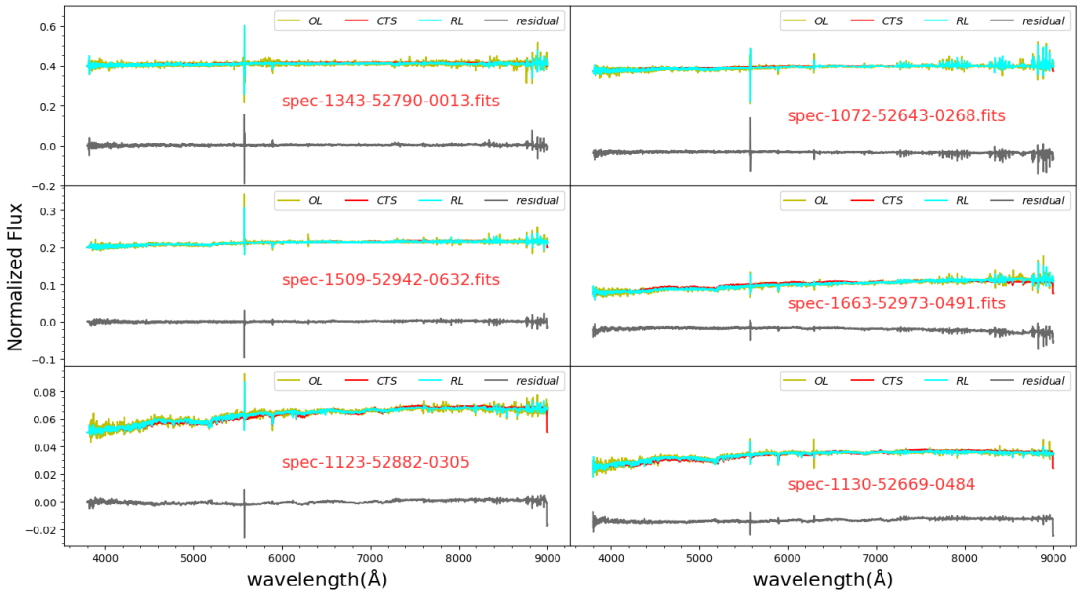


FIGURE 9. Comparison of various low-S/Ns denoised spectra and the corresponding template ones. The yellow, red, cyan and dimgray spectra are the original spectra, the template spectra, the denoised spectra and the residual spectra i.e. resconstructed minus original, respectively.

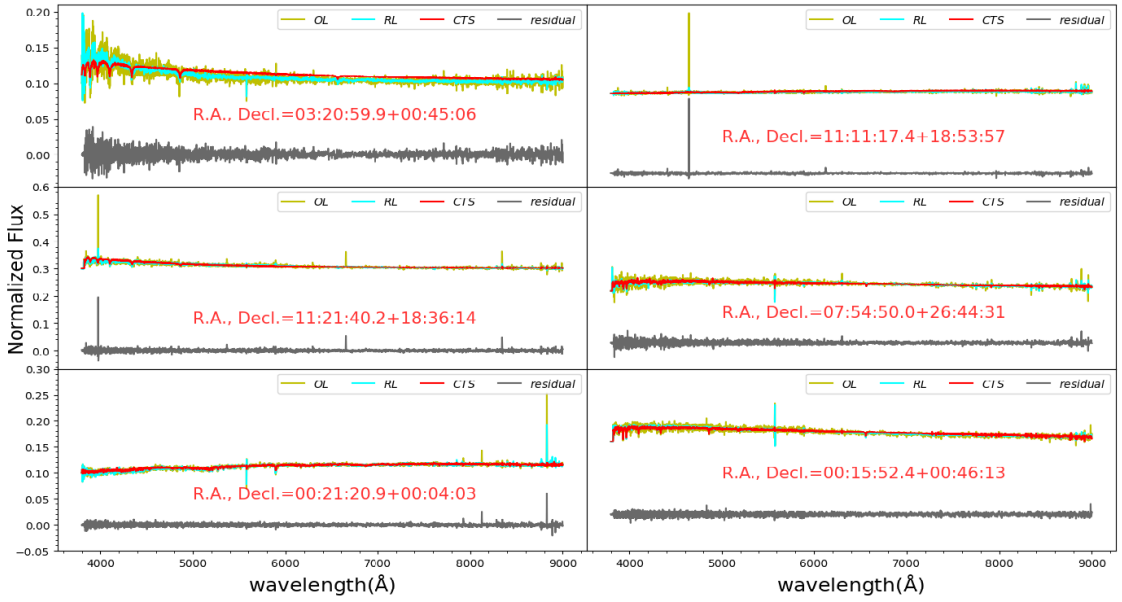


FIGURE 10. Comparison of various low-S/Ns reconstruction spectra and the corresponding high-S/Ns original ones. The legend is the same as in Fig. 9.

where *RS* represents the number of candidate spectra after retrieval and *TS* represents the number of original spectra. Therefore, the higher the *RR*, the better the experimental effect, and the smaller the *CR*, the better the experimental effect. From Fig. 12, we find that our method can perfectly separate the rare objects from the normal spectra, which demonstrates the effectiveness of our method in finding the rare objects.

B. SEARCHING FOR RARE OBJECTS FROM THE OBSERVED SPECTRA

In this subsection, we apply our method to the observed spectra. We randomly select normal spectra, and randomly select Carbon, Carbon WD, and CVs spectra as the rare spectra from SDSS DR14. According to the above criterion, 11421 spectra with $1 \leq S/N \leq 5$ are selected (see Data Set 1 in Table 3), and then are denoised with the first 221 general feature spectra.

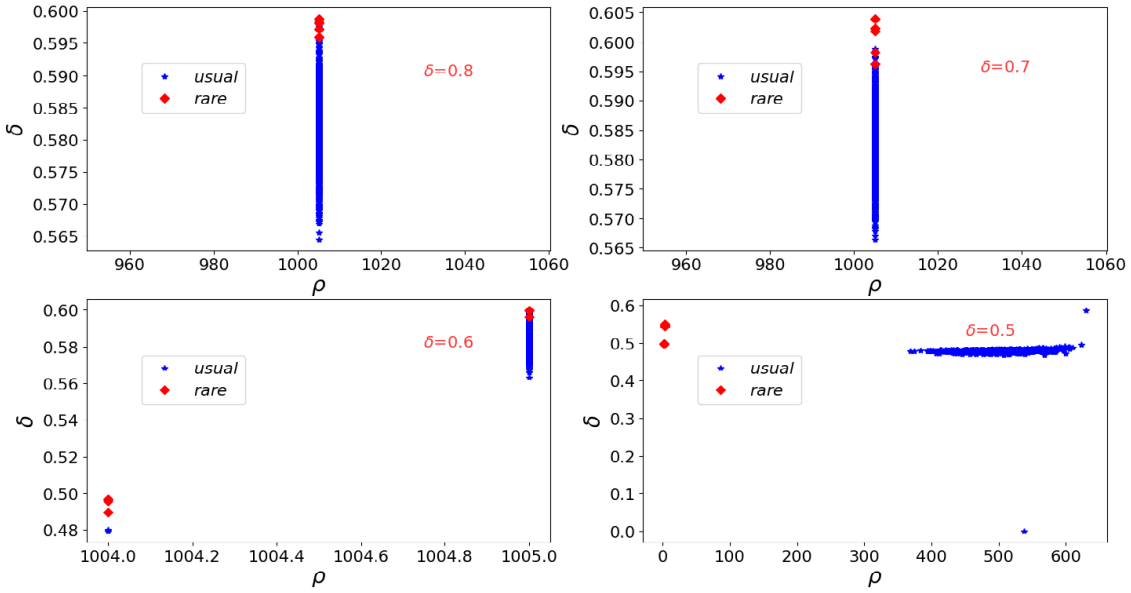


FIGURE 11. We select these spectra with S/N=1 from the template spectra with added noise, and the blue and red markers represents the normal and rare stellar spectra, respectively.

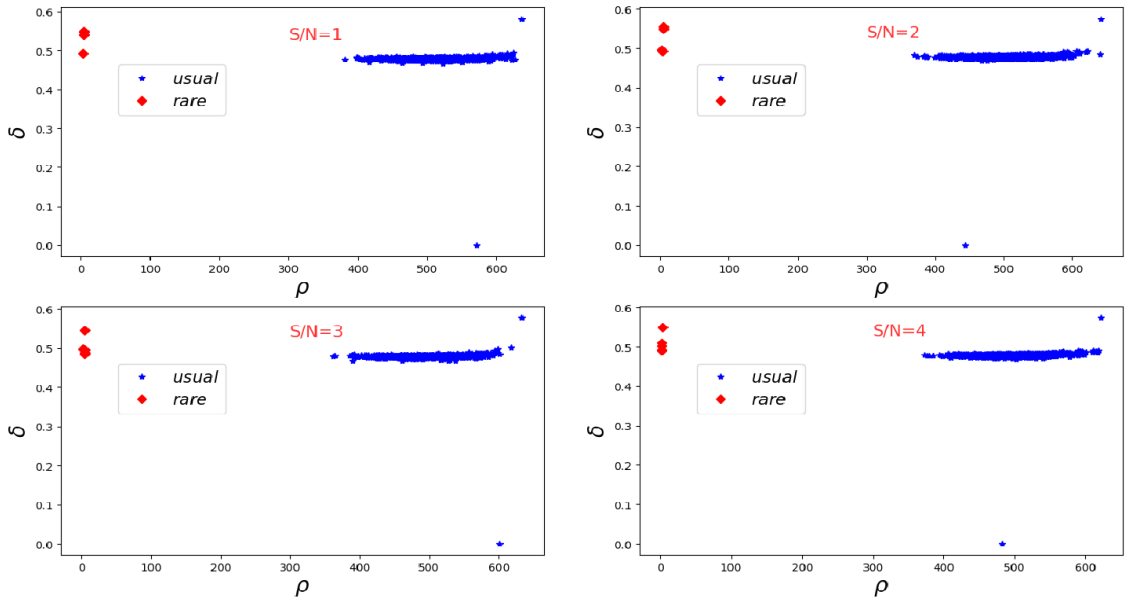


FIGURE 12. Search results of noise-added template spectra with different low-S/Ns. The cutoff distance in the top left panel, top right panel, bottom left panel, and bottom right panel is 0.5. The blue represents the normal spectrum; the red represents the rare stellar spectrum.

As shown in the Fig. 13, the red points are the rare stellar spectra, and the blue ones are considered as the normal spectra. For the spectra with 1≤S/N≤5, most rare objects can gather on the upper left corner of the decision graph. However, some of rare objects are mixed with the normal ones, so the results are not as good as the ones of the template spectra. The reasons may be: (1) the real noise may differ from simulated Gaussian white noise; (2) the low-S/N stellar classification in SDSS may not be very accurate (eg. RA=07:54:50.0, Decl.= +26:44:31 in Table 5), so the normal spectra and rare spectra may have errors. However, it can be seen from the data (see Table 6) that the proposed method can greatly reduce the number of rare object candidates, which can greatly improve the subsequent search efficiency of rare objects.

Due to the classification errors of low-S/N spectra in SDSS, the credibility of the experiments is likely to be reduced. So we used CVs to further explore the effectiveness of our method in the following section.

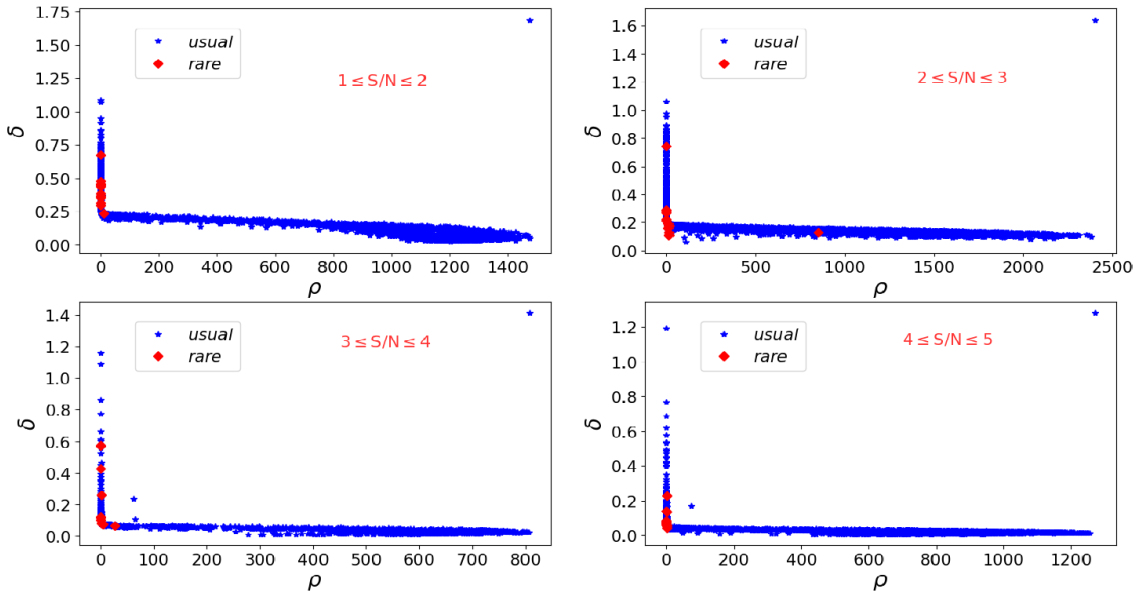


FIGURE 13. Search results of observed spectra with different low-S/Ns. The cutoff distances of the top left, top right, bottom left, and bottom right panels are 0.25, 0.2, 0.08, 0.05, respectively. The blue points and red points represent the normal and rare stellar spectra, respectively.

TABLE 6. Search result of the observed spectra with 1≤S/N≤5.

S/N	Total Number[a]	Number of Candidates[b]	*RR*	*CR*
1-2	2220	684	100%	30.8%
2-3	3374	691	91.67%	20.5%
3-4	2039	1091	91.67%	53.5%
4-5	3788	877	100%	23.2%

[a] Number of total spectra for each S/N.
[b] Number of candidate spectra for each S/N.

VI. CV S SEARCH VERIFICATION

In this section, we use the method to search for the CVs, and analyze from different aspects. These CVs are chosen from 285 identified CVs [3]

A. SEARCH UNDER DIFFERENT S/N S

We search for CVs from spectra with different S/Ns (see Data Set 2 in Table 3). The search results are plotted in Fig. 14. From Fig. 14, we can see that almost all the CVs (red diamond) can be gathered on the upper left corner of the decision graph. The specific informations are shown in Table 7.

TABLE 7. Search results of three groups of spectra with different S/Ns.

S/N	Total Number[a]	Number of Candidates[b]	*RR*	*CR*
31-35	12115	30	100%	0.16%
11-15	12070	275	100%	2.28%
1-5	9823	3300	75%	33.59%

[a] Number of total spectra for each S/N.
[b] Number of candidate spectra for each S/N.

Nevertheless, it can be noticed from the bottom panel in Fig. 14 that there are two CVs that are mixed into the normal spectra. That may be because the S/N of these spectra is too low, so the denoising effect is not very well. However, beside from the two spectra, the *CR* can reach 33.59%, and the *RR* can also reach 75%. But this *CR* is still worse than the previous two groups (31≤S/N≤35 and 11≤S/N≤15). In order to further study this discrepancy, the spectra with 1≤S/N≤5 are further analyzed in detail.

B. DETAILED SEARCH UNDER 1≤S/N≤5

In this section, we divide the 1≤S/N≤5 into 4 groups, and reselect the corresponding spectra from SDSS DR14. The specific spectra informations are shown in Data Set 3 of Table 3. The results are shown in Fig. 15.

From Fig. 15, we could see that almost all the CVs (red diamond) are gathered on the upper left corner of the decision graph. The specific informations are shown in Table 8.

TABLE 8. Search results of four groups segment spectra with 1≤S/N≤5.

S/N	Total Number[a]	Number of Candidates[b]	*RR*	*CR*
1-2	7167	1007	75%	14.03%
2-3	14671	4250	83.33%	28.9%
3-4	15057	3300	66.67%	26.8%
4-5	11936	1149	87.5%	9.62%

[a] Number of total spectra for each S/N.
[b] Number of candidate spectra for each S/N.

As can be seen from the above analysis, for the spectra with 1≤S/N≤5, the *RR* and *CR* obtained by the disassembled analysis are basically much better than the mixed analysis in Section VI-A. The problem may be caused by the wide range of data distribution. From this perspective, the analysis results could be greatly improved by further subdividing the S/N for all spectra, especially very low-S/N ones.

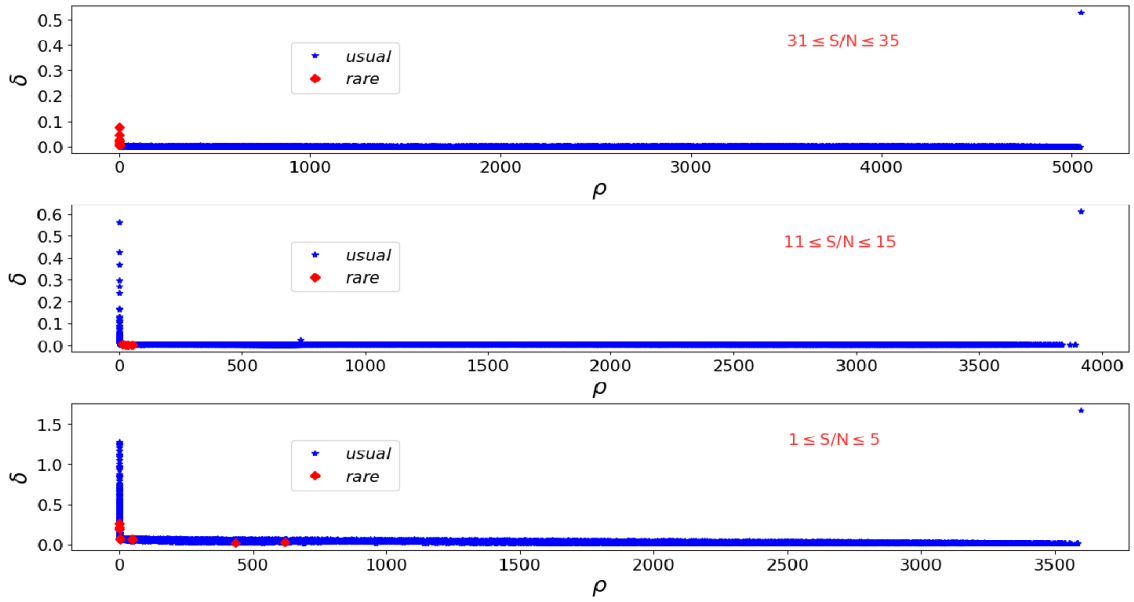


FIGURE 14. Search results of CVs with three different S/Ns. The cutoff distances from top panel to bottom panel are 0.005, 0.003, 0.05, respectively. The blue and red markers represent the normal and rare stellar spectra, separately.

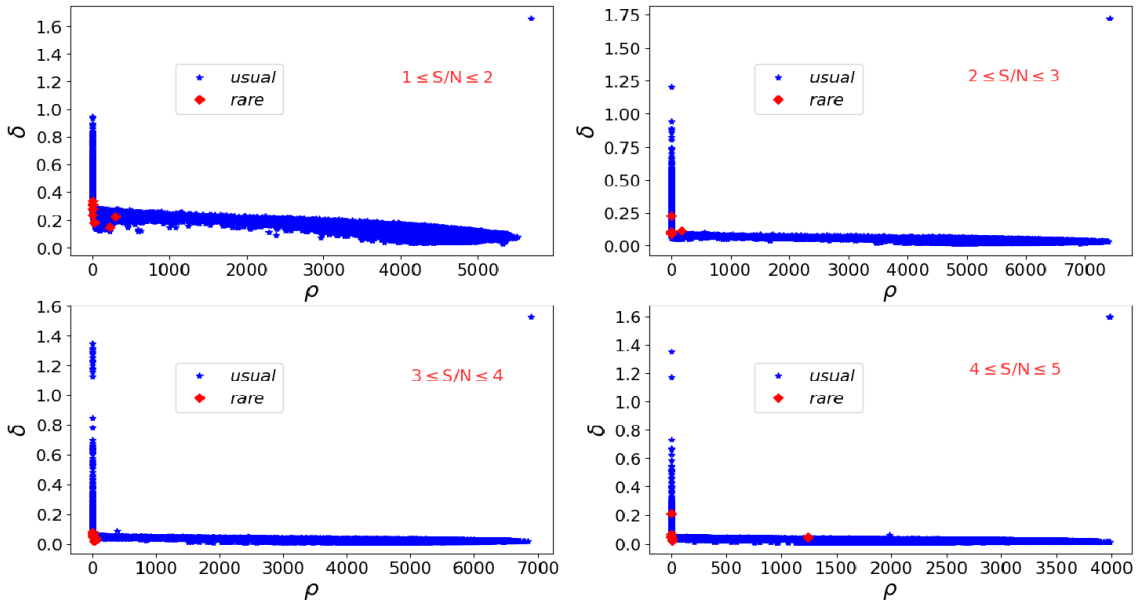


FIGURE 15. Search results of CVs of four segment spectra with 1≤S/N≤5. The cutoff distances of the top left, top right, bottom left and bottom right panels are 0.3, 0.1, 0.08, 0.05, respectively. The blue and red markers represent the normal and rare stellar spectra, separatively.

C. COMPARISON OF METHOD PERFORMANCE

In order to verify the advantages of the proposed method in processing the low-S/N spectra, we compare our method with the commonly used search method in this subsection. Because there are few methods to specifically handle low-S/N spectra, we only select the Lick-index based method and SVM for comparison. The Lick-index is not sensitive to S/N [64], so it is often used to process low-S/N spectra. As described in [24], for Lick-index+K-means, we set the number of clusters K to 100. With regard to the SVM, we set the parameters similar to those described in [65].

The data is same as Section VI-A (see Data Set 2 in Table 3). Table 9 shows the search result of the proposed method, Lick-index+K-means, and SVM from the low-S/N spectra in Data Set 2, respectively. Among these, the proposed method and Lick-index+K-means are unsupervised method,

TABLE 9. Comparision of methods.

S/N	Total Number	Methods	*CR*	*RR*
31-35	12115	SVM	0.96%	100%
		Lick-index+K-means	5.03%	100%
		Proposed method	**0.25**%	100%
11-15	12070	SVM	2.36%	75%
		Lick-index+K-means	8.35%	75%
		Proposed method	**2.28**%	**100**%
1-5	9823	SVM	45.55%	50%
		Lick-index+K-means	53.65%	25%
		Proposed method	**33.59**%	**75**%

while SVM is supervised method. These results indicate that our method obtains better search results compared to the other methods. That is, our method is able to not only get the maximum recall rate (*RR*), but also get the minimum candidate rate (*CR*), especially in extremely low-S/N. In addition, as an unsupervised method, our method is faster than other methods due to avoiding the selection of cluster centers.

VII. CONCLUSIONS

In this study, we propose a new method of denoising the low-S/N spectra and discovering the rare objects from low-S/N spectra based on PCA and CFSFDP. Different from traditional denoising methods, we apply PCA to a spectral data set composed of various high-S/N spectra to extract the general spectral features, and denoise various low-S/N spectra accurately based on these general features. Based on the accurately denoised spectra, we use the fast clustering method CFSFDP to discover the rare objects from the low-S/N spectra. The main conclusions of this study are as follows:

(1) The main difference between our method and traditional method is that we extract the features from the high-S/N spectra and build the general spectral feature library. This method avoid the possibility that we extract some unuseful features. The result shows that this method can accurately denoise various low-S/N spectra.

(2) Based on the accurately denoised spectra and the fast clustering method CFSFDP, we can efficiently discover the rare objects such as the WD and CVs from the low-S/N spectra.

(3)The proposed method is based on the idea that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities. This idea forms the basis of a clustering procedure in which the number of clusters arises intuitively, outliers are automatically spotted and excluded from the analysis. It avoids multiple iterations and greatly improves the operation speed, especially when processing large amounts of data.

However, the method proposed in this paper also shows the following limitations, which will be further studied in the follow-up study.

(1)The main idea for finding low-S/N outliers is to fit a PCA on high-S/N stellar spectra, and then apply these feature spectra to reconstruct low-S/N spectra. We assume that the underlying population of low and high-S/N spectra are the same. This assumption is not always true, because there could be objects in the low-S/N sample which are not represented at all in the high-S/N sample. So these unknown low-S/N could not be well reconstructed, and they may not be identified as outliers. Some delicate outliers such as O or B type stars with weird absorption line ratios could not be found. Our method can only detect ''outliers that shout the loudest'' as discussion in [40], such CVs.

(2) As described in the original literature [56], one can choose δ so that the average ρ is around 1% to 2% of the total number of points in the data set. However, for the distance that is used to select outliers, there is no exact discussion in the original literature, so the value is needed to set according to different situations.

(3) It is worth noting that after reconstructing, those reconstruction spectra with large error are excluded. This problem may be caused by data used for principal component extraction or other factors, but there are also many important informations in these excluded spectra. Therefore, this suggests that the spectra pre-processing approaches need further improvement.

In future, we hope to solve the above limitations and apply our method to other svrveys such as LAMOST, 2MASS and LSST, and so on.

ACKNOWLEDGEMENTS

This research has made use of data products from SDSS. Funding for the Sloan Digital Sky Survey 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 http://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.

REFERENCES

[1] H. Xiong, X. Chen, P. Podsiadlowski, Y. Li, and Z. Han, "Subdwarf b stars from the common envelope ejection channel," *Astron. Astrophys.*, vol. 599, p. A54, Mar. 2017.

[2] T. H. Jarrett, M. E. Cluver, C. Magoulas, M. Bilicki, M. Alpaslan, J. Bland-Hawthorn, S. Brough, M. J. I. Brown, S. Croom, S. Driver, B. W. Holwerda, A. M. Hopkins, J. Loveday, P. Norberg, J. A. Peacock, C. C. Popescu, E. M. Sadler, E. N. Taylor, R. J. Tuffs, and L. Wang, "Galaxy and mass assembly (GAMA): Exploring the WISE Web in G12," *Astrophys. J.*, vol. 836, no. 2, p. 182, Feb. 2017.

[3] P. Szkody, S. F. Anderson, K. Brooks, B. T. Gänsicke, M. Kronberg, T. Riecken, N. P. Ross, G. D. Schmidt, D. P. Schneider, M. A. Agüeros, A. N. Gomez-Moran, G. R. Knapp, M. R. Schreiber, and A. D. Schwope, "Cataclysmic variables from the Sloan digital sky survey. VIII. The final year (2007–2008)," *Astronomical J.*, vol. 142, no. 6, p. 181, Dec. 2011.

[4] A. F. Pala, B. T. Gänsicke, D. Townsley, D. Boyd, M. J. Cook, D. De Martino, P. Godon, J. B. Haislip, A. A. Henden, I. Hubeny, and K. M. Ivarsen, "Effective temperatures of cataclysmic-variable white dwarfs as a probe of their evolution," *Monthly Notices Roy. Astronomical Soc.*, vol. 466, no. 3, pp. 2855–2878, Apr. 2017.

[5] M. R. Kennedy, P. Callanan, P. M. Garnavich, M. Fausnaugh, and J. C. Zinn, "*XMM–Newton* observations of the peculiar cataclysmic variable lanning 386: X-ray evidence for a magnetic primary," *Monthly Notices Roy. Astronomical Soc.*, vol. 466, no. 2, pp. 2202–2211, Apr. 2017.

[6] R. Ridden-Harper, B. E. Tucker, P. Garnavich, A. Rest, S. Margheim, E. J. Shaya, C. Littlefield, G. Barensten, C. Hedges, and M. Gully-Santiago, "Discovery of a new WZ Sagittae-type cataclysmic variable in the *Kepler/K2* data," *Monthly Notices Roy. Astronomical Soc.*, vol. 490, no. 4, pp. 5551–5559, Dec. 2019.

[7] A. S. Oliveira, C. V. Rodrigues, D. Cieslinski, F. J. Jablonski, K. M. G. Silva, L. A. Almeida, A. Rodríguez-Ardila, and M. S. Palhares, "Exploratory spectroscopy of magnetic cataclysmic variables candidates and other variable objects," *Astronomical J.*, vol. 153, no. 4, p. 144, Apr. 2017.

[8] J. J. Wallace, J. D. Hartman, G. Á. Bakos, and W. Bhatti, "A search for variable stars in the globular cluster M4 with K2," *Astrophysical J. Suppl. Ser.*, vol. 244, no. 1, p. 12, Sep. 2019.

[9] J.-J. Ren, A. Rebassa-Mansergas, S. G. Parsons, X.-W. Liu, A.-L. Luo, X. Kong, and H.-T. Zhang, "White dwarf–main sequence binaries from LAMOST: The DR5 catalogue," *Monthly Notices Roy. Astronomical Soc.*, vol. 477, no. 4, pp. 4641–4654, Jul. 2018.

[10] F. M. Jiménez-Esteban, S. Torres, A. Rebassa-Mansergas, G. Skorobogatov, E. Solano, C. Cantero, and C. Rodrigo, "A white dwarf catalogue from *Gaia*-DR2 and the virtual observatory," *Monthly Notices Roy. Astronomical Soc.*, vol. 480, no. 4, pp. 4505–4518, Nov. 2018.

[11] M. Perpinyà-Vallès, A. Rebassa-Mansergas, B. T. Gänsicke, S. Toonen, J. J. Hermes, N. P. Gentile Fusillo, and P.-E. Tremblay, "Discovery of the first resolved triple white dwarf," *Monthly Notices Roy. Astronomical Soc.*, vol. 483, no. 1, pp. 901–907, Feb. 2019.

[12] A. Rebassa-Mansergas, J. J. Ren, S. G. Parsons, B. T. Gänsicke, M. R. Schreiber, E. García-Berro, X.-W. Liu, and D. Koester, "The SDSS spectroscopic catalogue of white dwarf-main-sequence binaries: New identifications from DR 9–12," *Monthly Notices Roy. Astronomical Soc.*, vol. 458, no. 4, pp. 3808–3819, Jun. 2016.

[13] R. Cojocaru, A. Rebassa-Mansergas, S. Torres, and E. García-Berro, "The population of white dwarf–main sequence binaries in the SDSS DR 12," *Monthly Notices Roy. Astronomical Soc.*, vol. 470, no. 2, pp. 1442–1452, Sep. 2017.

[14] A. Rebassa-Mansergas, E. Solano, S. Y. Xu, C. Rodrigo, F. M. Jiménez-Esteban, and S. Torres, "Infrared-excess white dwarfs in the *Gaia* 100 pc sample," *Monthly Notices Roy. Astronomical Soc.*, pp. 3990–4000, Sep. 2019.

[15] T. G. Wilson, J. Farihi, B. T. Gänsicke, and A. Swan, "The unbiased frequency of planetary signatures around single and binary white dwarfs using *Spitzer* and *Hubble*," *Monthly Notices Roy. Astronomical Soc.*, vol. 487, no. 1, pp. 133–146, Jul. 2019.

[16] Y. Bai, J.-F. Liu, and S. Wang, "*Gaia* calibrated UV luminous stars in LAMOST," *Res. Astron. Astrophys.*, vol. 18, no. 12, p. 156, Dec. 2018.

[17] Z. Lei, J. Zhao, P. Németh, and G. Zhao, "New hot subdwarf stars identified in *Gaia* DR2 with LAMOST DR5 spectra," *Astrophys. J.*, vol. 868, no. 1, p. 70, Nov. 2018.

[18] Z. Lei, J. Zhao, P. Németh, and G. Zhao, "Hot subdwarf stars identified in Gaia DR2 with spectra of LAMOST DR6 and DR7. I. Single-lined spectra," *Astrophys. J.*, vol. 889, no. 2, p. 117, Feb. 2020.

[19] G. P. Wang, J. C. Pan, Z. P. Yi, and P. Wei, "Research on the clustering of massive Stellar Spectra Based on Line index," *Spectrosc. Spectral Anal.*, vol. 36, no. 8, pp. 2646–2650, Aug. 2016.

[20] G.-W. Li, J.-R. Shi, B. Yanny, Z.-R. Bai, S.-C. Yu, Y.-Q. Dong, Y.-J. Lei, H.-L. Yuan, W. Zhang, and Y.-H. Zhao, "New Oe stars in LAMOST DR5," *Astrophys. J.*, vol. 863, no. 1, p. 70, Aug. 2018.

[21] B. Jiang, A. L. Luo, and Y. H. Zhao, "Data mining for cataclysmic variables candidates in massive spectra," *Spectrosc. Spectral Anal.*, vol. 31, no. 8, pp. 2278–2282, Aug. 2011.

[22] B. Jiang, A. L. Luo, and Y. H. Zhao, "Data mining approach to cataclysmic variables candidates based on random forest algorithm," *Spectrosc. Spectral Anal.*, vol. 32, no. 2, pp. 510–513, Feb. 2012.

[23] S. Akras, M. L. Leal-Ferreira, L. Guzman-Ramirez, and G. Ramos-Larios, "A machine learning approach for identification and classification of symbiotic stars using 2MASS and WISE," *Monthly Notices Roy. Astronomical Soc.*, vol. 483, no. 4, pp. 5077–5104, Mar. 2019.

[24] G. P. Wang, J. C. Pan, Z. P. Yi, and P. Wei, "Outlier data mining and analysis of LAMOST stellar spectra in line index feature space," *Spectrosc. Spectral Anal.*, vol. 36, no. 10, pp. 3364–3368, Oct. 2016.

[25] Y.-B. Li *et al.*, "Carbon stars identified from LAMOST DR4 using machine learning," *Astrophys. J. Suppl. Ser.*, vol. 234, no. 2, p. 31, Feb. 2018.

[26] M. Abbas, E. K. Grebel, N. F. Martin, N. Kaiser, W. S. Burgett, M. E. Huber, and C. Waters, "An optimized method to identify RR Lyrae stars in the SDSS×Pan-STARRS1 overlapping area using a Bayesian generative technique," *Astronomical J.*, vol. 148, no. 1, p. 8, Mar. 2014.

[27] S. Kheirdastan and M. Bazarghan, "SDSS-DR12 bulk stellar spectral classification: Artificial neural networks approach," *Astrophys. Space Sci.*, vol. 361, no. 9, p. 304, Sep. 2016.

[28] D.-W. Kim and C. A. L. Bailer-Jones, "A package for the automated classification of periodic variable stars," *Astron. Astrophys.*, vol. 587, p. A18, Mar. 2016.

[29] D. G. York *et al.*, "The Sloan digital sky survey: Technical summary," *Astronomical J.*, vol. 120, no. 3, pp. 1579–1587, May 2000.

[30] X. Q. Cui *et al.*, "The large sky area multi-object fiber spectroscopic telescope (LAMOST)," *Res. Astron. Astrophys.*, vol. 12, no. 9, pp. 1197–1242, Sep. 2012.

[31] G. Zhao, Y. H. Zhao, Y. Q. Chu, Y. P. Jing, and L. C. Deng, "LAMOST spectral survey—An overview," *Res. Astron. Astrophys.*, vol. 12, no. 7, pp. 723–734, Jul. 2012.

[32] A.-L. Luo *et al.*, "Data release of the LAMOST pilot survey," *Res. Astron. Astrophys.*, vol. 12, no. 9, pp. 1243–1246, Sep. 2012.

[33] Y. Yang, J. Cai, H. Yang, J. Zhang, and X. Zhao, "TAD: A trajectory clustering algorithm based on spatial-temporal density analysis," *Expert Syst. Appl.*, vol. 139, Jan. 2020, Art. no. 112846.

[34] Y. Li, J. Cai, H. Yang, J. Zhang, and X. Zhao, "A novel algorithm for initial cluster center selection," *IEEE Access*, vol. 7, pp. 74683–74693, 2019.

[35] A. Mahabal, K. Sheth, F. Gieseke, A. Pai, S. G. Djorgovski, A. Drake, and M. Graham, "Deep-learnt classification of light curves," 2017, *arXiv:1709.06257*. [Online]. Available: http://arxiv.org/abs/1709.06257

[36] B. Arsioli and Y.-L. Chang, "The γ-ray emitting region in low synchrotron peak blazars: Testing self-synchrotron Compton and external Compton scenarios," *Astron. Astrophys.*, vol. 616, p. A63, Aug. 2018.

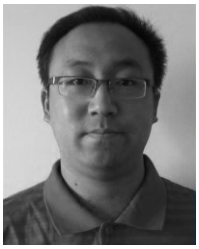
[37] M. Huertas-Company, J. R. Primack, A. Dekel, D. C. Koo, S. Lapiner, D. Ceverino, R. C. Simons, G. F. Snyder, M. Bernardi, Z. Chen, H. Domínguez-Sánchez, C. T. Lee, B. Margalef-Bentabol, and D. Tuccillo, "Deep learning identifies High-z galaxies in a central blue nugget phase in a characteristic mass range," *Astrophys. J.*, vol. 858, no. 2, p. 114, May 2018.

[38] C. Qu, H. Yang, J. Cai, J. Zhang, and Y. Zhou, "DoPS: A double-peaked profiles search method based on the RS and SVM," *IEEE Access*, vol. 7, pp. 106139–106154, 2019.

[39] A. Solarz, M. Bilicki, M. Gromadzki, A. Pollo, A. Durkalec, and M. Wypych, "Automated novelty detection in the WISE survey with one-class support vector machines," *Astron. Astrophys.*, vol. 606, p. A39, Oct. 2017.

[40] D. Baron, "Machine learning in astronomy: A practical overview," 2019, *arXiv:1904.07248*. [Online]. Available: http://arxiv.org/abs/1904.07248

[41] I. Reis, D. Poznanski, and P. B. Hall, "Redshifted broad absorption line quasars found via machine-learned spectral similarity," *Monthly Notices Roy. Astronomical Soc.*, vol. 480, no. 3, pp. 3889–3897, Nov. 2018.
[42] I. Nun, P. Protopapas, B. Sim, and W. Chen, "Ensemble learning method for outlier detection and its application to astronomical light curves," *Astronomical J.*, vol. 152, no. 3, p. 71, Sep. 2016.
[43] J. D. Simpson, P. L. Cottrell, and C. C. Worley, "Spectral matching for abundances and clustering analysis of stars on the giant branches of ω centauri," *Monthly Notices Roy. Astronomical Soc.*, vol. 427, no. 2, pp. 1153–1167, Dec. 2012.
[44] J. Sánchez Almeida and C. Allende Prieto, "Automated unsupervised classification of the Sloan digital sky survey stellar spectra using k-means clustering," *Astrophys. J.*, vol. 763, no. 1, p. 50, Jan. 2013.
[45] R. S. de Souza, M. L. L. Dantas, M. V. Costa-Duarte, E. D. Feigelson, M. Killedar, P.-Y. Lablanche, R. Vilalta, A. Krone-Martins, R. Beck, and F. Gieseke, "A probabilistic approach to emission-line galaxy classification," *Monthly Notices Roy. Astronomical Soc.*, vol. 472, no. 3, pp. 2808–2822, Dec. 2017.
[46] D. Baron, D. Poznanski, D. Watson, Y. Yao, N. L. J. Cox, and J. X. Prochaska, "Using machine learning to classify the diffuse interstellar bands," *Monthly Notices Roy. Astronomical Soc.*, vol. 451, no. 1, pp. 332–352, Jul. 2015.
[47] M. A. Peth, J. M. Lotz, P. E. Freeman, C. McPartland, S. A. Mortazavi, G. F. Snyder, G. Barro, N. A. Grogin, Y. Guo, S. Hemmati, J. S. Kartaltepe, D. D. Kocevski, A. M. Koekemoer, D. H. McIntosh, H. Nayyeri, C. Papovich, J. R. Primack, and R. C. Simons, "Beyond spheroids and discs: Classifications of CANDELS galaxy structure at $1.4 <z< 2$ via principal component analysis," *Monthly Notices Roy. Astronomical Soc.*, vol. 458, no. 1, pp. 963–987, May 2016.
[48] R. Ma, R. A. Angryk, P. Riley, and S. F. Boubrahimi, "Coronal mass ejection data clustering and visualization of decision trees," *Astrophys. J. Suppl. Ser.*, vol. 236, no. 1, p. 14, May 2018.
[49] C. A. Whitney, "Principal components analysis of spectral data. I—Methodology for spectral classification," *Astron. Astrophys. Suppl. Ser.*, vol. 51, pp. 443–461, Mar. 1983.
[50] C. A. Whitney, "Principal components analysis of spectral data. II—Error analysis and applications to interstellar reddening, luminosity classification of M supergiants, and the analysis of VV Cephei stars," *Astron. Astrophys. Suppl. Ser.*, vol. 51, pp. 463–478, Mar. 1983.
[51] H. P. Singh, R. K. Gulati, and R. Gupta, "Stellar spectral classification using principal component analysis and artificial neural networks," *Monthly Notices Roy. Astronomical Soc.*, vol. 295, no. 2, pp. 312–318, Apr. 1998.
[52] D. M. Qin, Z. Y. Hu, and Y. H. Zhao, "A PCA based efficient stellar spectra classification method," *Spectrosc. Spectral Anal.*, vol. 23, no. 1, pp. 182–186, Feb. 2003.
[53] M. Williamson, M. Modjaz, and F. Bianco, "Optimal classification and outlier detection for stripped-envelope core-collapse supernovae," 2019, *arXiv:1903.06815*. [Online]. Available: http://arxiv.org/abs/1903.06815
[54] D. Giles and L. Walkowicz, "Systematic serendipity: A test of unsupervised machine learning as a method for anomaly detection," *Monthly Notices Roy. Astronomical Soc.*, vol. 484, no. 1, pp. 834–849, Mar. 2019.
[55] I. T. Jolliffe, *Principal Component Analysis* Berlin, Germany: Springer, 1986.
[56] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.
[57] M. R. Blanton *et al.*, "Sloan digital sky survey IV: Mapping the Milky Way, nearby galaxies, and the Distant Universe," *Astronomical J.*, vol. 154, no. 1, p. 28, Jun. 2017.
[58] Y. Wang, L. Xu, and G.-B. Zhao, "A measurement of the hubble constant using galaxy redshift surveys," *Astrophys. J.*, vol. 849, no. 2, p. 84, No. 2017.
[59] J. Hou *et al.*, "The clustering of the SDSS-IV extended baryon oscillation spectroscopic survey DR14 quasar sample: Anisotropic clustering analysis in configuration space," *Monthly Notices Roy. Astronomical Soc.*, vol. 480, no. 2, pp. 2521–2534, Oct. 2018.
[60] M. Blomqvist, M. M. Pieri, H. du Mas des Bourboux, N. G. Busca, A. Slosar, J. E. Bautista, J. Brinkmann, J. R. Brownstein, K. Dawson, V. D. S. Agathe, J. Guy, W. J. Percival, I. Pérez-Ràfols, J. Rich, and D. P. Schneider, "The triply-ionized carbon forest from eBOSS: Cosmological correlations with quasars in SDSS-IV DR14," *J. Cosmol. Astropart. Phys.*, vol. 2018, no. 5, p. 029, May 2018.
[61] H. Gil-Marín *et al.*, "The clustering of the SDSS-IV extended baryon oscillation spectroscopic survey DR14 quasar sample: Structure growth rate measurement from the anisotropic quasar power spectrum in the redshift range $0.8 <z< 2.2$," *Monthly Notices Roy. Astronomical Soc.*, vol. 477, no. 2, pp. 1604–1638, Jun. 2018.
[62] B. Ansarinejad and T. Shanks, "Detection significance of baryon acoustic oscillations peaks in galaxy and quasar clustering," *Monthly Notices Roy. Astronomical Soc.*, vol. 479, no. 3, pp. 4091–4107, Sep. 2018.
[63] A. S. Bolton *et al.*, "Spectral classification and redshift measurement for the SDSS-III baryon oscillation spectroscopic survey," *Astronomical J.*, vol. 144, no. 5, p. 144, Nov. 2012.
[64] G. Worthey, S. M. Faber, J. J. Gonzalez, and D. Burstein, "Old stellar populations. 5: Absorption feature indices for the complete LICK/IDS sample of stars," *Astrophys. J. Suppl. Ser.*, vol. 94, pp. 687–722, Oct. 1994.
[65] I. N. Pashchenko, K. V. Sokolovsky, and P. Gavras, "Machine learning search for variable stars," *Monthly Notices Roy. Astronomical Soc.*, vol. 475, no. 2, pp. 2326–2343, Apr. 2018.



MINGLEI WU was born in January 1986. He received the B.S. degree from Shandong Normal University, in 2010, and the M.S. degree from the Harbin University of Science and Technology, in 2013. He is currently pursuing the Ph.D. degree with Shandong University, China. His current research interests include big data and data mining.

JINGCHANG PAN received the B.S. degree from Peking University, China, in 1986, the M.S. degree from the Harbin Institute of Technology, China, in 2006, and the Ph.D. degree from Shandong University, China, in 2011. He is currently a Professor with the School of Mechanical, Electrical and Information Engineering, Shandong University. His research interests include data mining, parallel computing, and stellar spectra data analysis.

ZHENPING YI received the B.S. and M.S. degrees from Shandong University, China, in 2002 and 2005, respectively, and the Ph.D. degree from the University of Chinese Academy of Sciences, China, in 2015. She is currently an Associate Professor with the School of Mechanical, Electrical and Information Engineering, Shandong University. Her research interests include big data and stellar spectra data analysis.

PENG WEI received the B.S. and M.S. degrees from Shandong University, China, in 2008 and 2011, respectively, and the Ph.D. degree from the University of Chinese Academy of Sciences, China, in 2014. His research interests include artificial intelligence and data mining.