RAA 2019 Vol. **19** No. **6**, 88(20pp) doi: 10.1088/1674–4527/19/6/88

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

Research in Astronomy and Astrophysics

A catalog of DB white dwarfs from the LAMOST DR5 and construction of templates

Xiao Kong[1,2], A-Li Luo[1,2] and Xiang-Ru Li[3]

[1] Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China; *lal@bao.ac.cn*
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Mathematical Sciences, South China Normal University, Guangzhou 510631, China

Received 2018 August 29; accepted 2018 December 20

Abstract In this study, we employ machine learning to build a catalog of DB white dwarfs (DBWDs) from the LAMOST Data Release (DR) 5. Using known DBs from SDSS DR14, we selected samples of high-quality DB spectra from the LAMOST database and applied them to train the machine learning process. Following the recognition procedure, we chose 351 DB spectra of 287 objects, 53 of which were new identifications. We then utilized all the DBWD spectra from both SDSS DR14 and LAMOST DR5 to construct DB templates for LAMOST 1D pipeline reductions. Finally, by applying DB parameter models provided by D. Koester and the distance from *Gaia* DR2, we calculated the effective temperatures, surface gravities and distributions of the 3D locations and velocities of all DBWDs.

Key words: (stars:) white dwarfs — catalogs — surveys — methods: data analysis

1 INTRODUCTION

With initial masses of up to $\sim 9\,M_{\odot}$ (Woosley & Heger 2015), white dwarfs (WDs) are the final state of stellar evolution for stars on the main sequence. Owing to the absence of nuclear reactions, energy from most WDs is generated by radiation from residual gravitational contraction, which can lead to relatively low brightness. Although the mean mass of a majority of WDs is $\sim 0.593 \pm 0.016\,M_{\odot}$ (Kepler et al. 2007), the common radii of these stars are often of the same order as that of the Earth, which implies that an extremely long cooling time is required.

Among WDs, those with atmospheres that are mostly rich in hydrogen account for $\sim$80%, which are assigned to the DA spectral type. The other 20% fall into the DO (He II) or DB (He I) categories, whose atmospheres are dominated by helium with, occasionally, minute traces of hydrogen and heavy elements. In most cases, hot DO stars can be observed at temperatures of $\sim$45 000 K. DB WDs (DBWDs) have effective temperatures ($T_{\rm eff}$) averaging $<$30 000 K, with only He I in their spectra. When the temperature drops to 10 000 K, helium becomes spectroscopically invisible, such as for featureless DC, carbon-present DQ and metal-rich DZ spectra (Voss et al. 2007).

Because the helium atom prevails in its ground state in the atmosphere, DBWDs represent the best sample of hydrogen-deficient stars in the universe. Many hydrogen-dominated DA WDs transform into DBs with a helium atmosphere, and the ratio of DA to non-DA WDs varies as a function of $T_{\rm eff}$ along the cooling sequence (Fontaine et al. 2001). An expanded sample of DBWDs is helpful to better understand the evolution of WDs.

At the end of the 20th century, only approximately 80 optical spectra and 25 ultraviolet spectrophotometric spectra had been investigated (Beauchamp et al. 1996). Since the Sloan Digital Sky Survey (SDSS) commenced data releases (DRs), systematic research has been conducted to gather larger samples of DBWDs (Eisenstein et al. 2006). Kleinman et al. (2013) identified 923 DB stars from the SDSS DR7 (Abazajian et al. 2009), and Kepler et al. (2015) added another 450 in DR10 (Eisenstein et al. 2011). DR12 (Alam et al. 2015) featured 121 more stars (Kepler et al. 2016), from which Koester & Kepler (2015) selected 1107 objects with signal-to-noise (S/N) ratios of $>$10 and analyzed their atmospheric parameters. Using a machine learning (ML) approach to extract DBWD features and recognize their spectra, we sought in past work to identify DBWDs from the entire spectral data of SDSS DR12 and DR14, and presented their particular features us-

ing a line list (Kong et al. 2018, KONG2018 hereinafter), thus increasing the total number of identified DB stars to 1999.

In addition to SDSS, the Large Sky Area Multi-object Fiber Spectroscopic Telescope (LAMOST, also called the Guo Shou Jing Telescope, Cui et al. (2012)) released 9 017 844 spectra from a spectroscopic survey, including 9574 WDs, in DR5. In spite of its massive number of released spectra, research on DBWDs using LAMOST DRs has been scarce compared with that from SDSS DRs. A few years ago, Guo et al. (2015) presented 34 DB and 1056 DA WDs from LAMOST DR2, and estimated their $T_{\rm eff}$, surface gravity (log g) and distance. Ren et al. (2013, 2014, 2018) mined the WD-main sequence binaries from the LAMOST pilot DR, DR1 and DR5. All of them were DA main sequence binaries.

In general, spectra from both SDSS and LAMOST are more or less the same: They share a similar resolution ($R \approx 1800$[1,2]) and waveband ($\sim$3900–$\sim$9000 Å). Given these similarities, we intend to search for DBWDs in LAMOST DR5 and provide their particular program-sensitive features by using the ML method applied to KONG2018. It should be noted that there is a "WD" class in both SDSS DR12 and LAMOST DR5 that contains DA WDs with a few other subtypes of WDs mixed in.

In this paper, we build a DB catalog from LAMOST DR5 using ML. Some known DB spectra from KONG2018 first served as positive samples. We carried out a data-mining procedure and obtained some DB spectra from the LAMOST DR5. The positive samples that we selected must show at least two obvious He I lines (4471.5 or 5875.6 Å). Otherwise, they can hardly be spectroscopically identified. According to our experience at examining spectra, the spectra with S/N in the g band (S/N$_g$) above 15 would generally meet these conditions. Following manual inspection, we discarded those without clear He I lines (4471.5 and 5875.6 Å), and arranged the remaining spectra in descending order S/N$_g$ and selected the top 100 as positive samples. The S/N$_g$ of the 100th spectrum was 14.6 (specid is 20150413HD140137N164527M0116152_v2.9.7). These were used to extract features and recognize DBs by using the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1996) and a support vector machine (SVM) (Cortes & Vapnik 1995), respectively. "Feature" here has the same meaning as flux at some particular wavelengths or specific location in the spectra.

Moreover, there was a DB template in the SDSS one-dimensional (1D) pipeline before SDSS DR7, which was later removed probably because of its low accuracy when using the DB templates for classification. In this experiment, however, we constructed DB templates for the LAMOST 1D pipeline, and propose criteria to improve the accuracy of recognition by analyzing the results from classification of several control groups of data.

[1] http://classic.sdss.org/dr1/

[2] http://www.lamost.org/

Table 1 Class/Subclass from LAMOST DR5 catalog (except for "Unknown" class), also types of spectra we applied in our experiment.

Class[a]	Subclass[a]
Star	O, B, A, F, G, K, M, WD, carbon, CV, doublestar
Galaxy	null
QSO	null

[a] "Class" and "subclass" are adopted from the data archive of LAMOST DR5.

Wegg & Phinney (2012) investigated the relationship between the kinematics and mass of young DA WDs using SDSS DR4 and the Palomar-Green (PG) WD survey (Liebert et al. 2005), and found a strong connection. We calculated the three-dimensional (3D) velocities of DBWDs and obtained a similar conclusion. With respect to the kinematics of old stars, low-mass WDs (between $0.45\,M_\odot$ and $0.75\,M_\odot$) have a higher velocity dispersion and asymmetric drift whereas the ones with greater mass are the opposite.

The remainder of this paper is organized as follows: The datasets are defined and constructed in Section 2. Section 3 describes the training procedure; i.e., feature extraction by implementing LASSO and DB recognition by using an SVM. We then select all DBWD spectra from LAMOST DR5 in Section 4 employing the ML method. Application of this experiment to the LAMOST 1D pipeline is then introduced in Section 5; i.e., for DB template construction. Section 6 provides $T_{\rm eff}$, log g, and 3D distributions of the locations and velocities of the DBWDs, and we analyze the similarities and differences between DBWDs from LAMOST DR5 and SDSS DR14. Finally, we summarize the conclusions of this study in Section 7.

2 DATASETS

In total, there were >2000 DB spectra in SDSS DR12, based on catalogs from Kleinman et al. (2013), Kepler et al. (2015, 2016) and KONG2018. These spectra were originally classified as O, QSO, B, galaxy or some other type by the SDSS 1D pipeline. Moreover, the core algorithm of the LAMOST Pipeline is the same as that of the SDSS Pipeline: full-spectral template matching. Similar to the analysis in our previous work, DB spectra might be found in all types (Table 1) of spectra in LAMOST DR5.

For the data archive of LAMOST DR5, each released spectrum was assigned a specific title, namely, a class and a subclass, as displayed in Table 1. Therefore, we needed

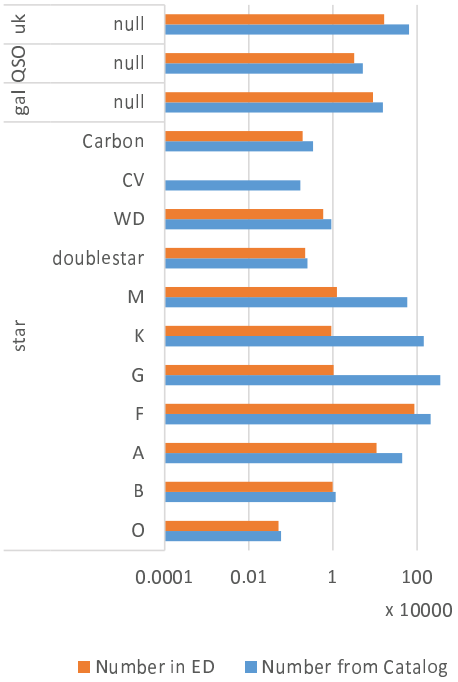

Fig. 1 Data usage of our experiment. The number of spectra from the LAMOST DR5 catalog in each CPS are shown by the *blue lines* while those from ED are in *red*. "uk" stands for "Unknown" (*Color version is online*).

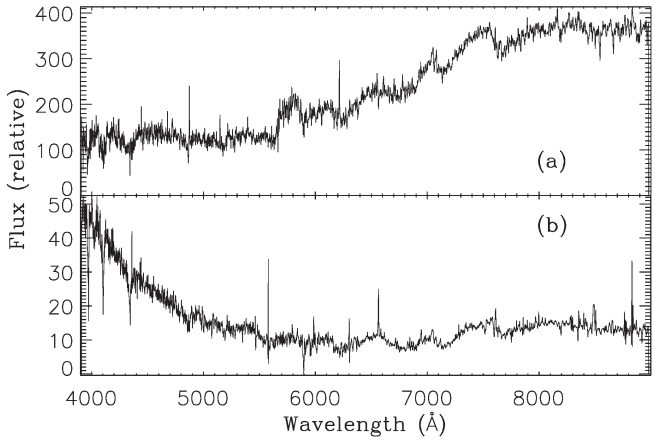

Fig. 2 Two examples of "DoubleStar" from LAMOST DR5: (a) DA+M and (b) A+M.

to classify between DB and each "subclass" in parallel. For convenience, we abbreviate "class+subclass" (CPS) as the identification of each subclass in this experiment, such as "star+O" or "QSO+null."

There was, however, a notable difference in class from SDSS and LAMOST in that there was an "Unknown" class in the catalog of LAMOST, whereas the spectra from SDSS only had three classes; star, galaxy and QSO; in which the term "class" was employed from the data archive of these two sky surveys. Therefore, the spectral data classified as "Unknown" should have been preprocessed before recognition (Sect. 2.3).

As listed in Table 2, in this experiment, we constructed three subsets from LAMOST DR5 for training, testing and searching. The training set was used for learning; i.e., to fit the parameters (features and hyperplanes) of the classifier. The testing set was used for adjusting the parameters of the classifier: to choose the best features and the most suitable kernel function of the SVM. A 10-fold cross-validation was built into both the LASSO and SVM packages, and was executed automatically by using the training set. For application, we selected candidates, referred to as "experimental data" (ED), from all spectral data, and explain the selection procedure in Section 2.3.

Table 2 Roles of the Three Datasets

Dataset	Role
Training Data	To be used in the training process, i.e. Detecting features by LASSO (Sect. 3.2); Estimating the parameterizing model by LIBSVM (Sect. 3.4).
Testing Data	To be used in the training process, i.e. Determining the parameters in LASSO (Sect. 3.2); Determining the hyper-planes in LIBSVM (Sect. 3.4).
Experimental Data	Application of Sect. 3, to be used in Searching for DB spectra from Experimental Data (Sect. 4).

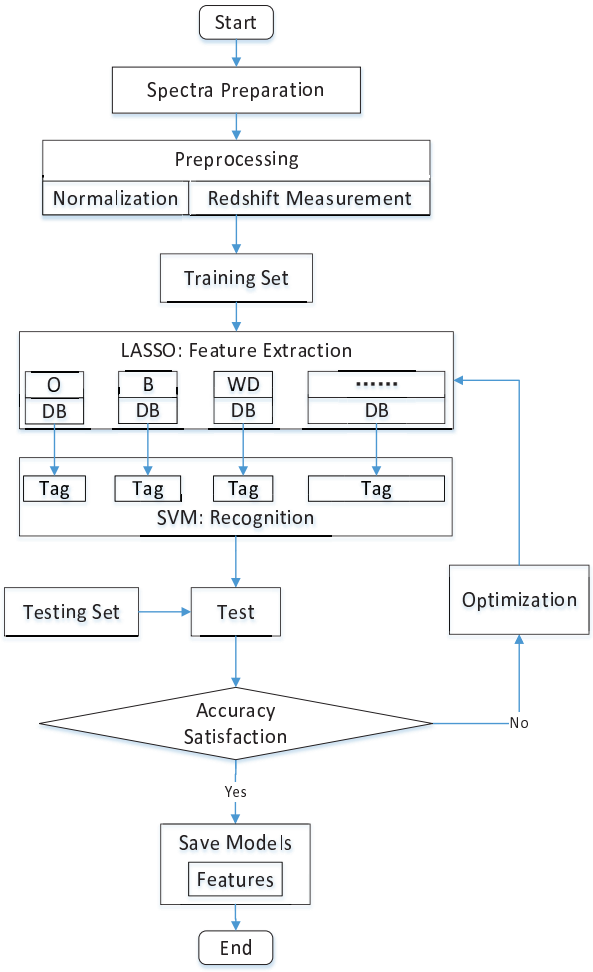

Fig. 3 Flowchart for the training process of the DB-mining procedure in our experiment, which mainly consists of preprocessing, feature extraction, classification and optimization.

2.1 Preparation of Positive Samples

Consider the similarity of spectra from both LAMOST and SDSS. Only 34 DBWDs were identified from LAMOST DR2 (Guo et al. 2015) and six DB candidates from LAMOST DR3 (Gentile Fusillo et al. 2015). Because few DB spectra were available in LAMOST DR5, DB spectra from SDSS DR14 were required to complete this experiment. A total of 300 known DBWDs and DB feature spaces obtained from KONG2018 first served as the positive samples. With spectra from LAMOST DR5 as negative samples, we conducted the recognition process using an SVM program.

Following the classification procedure, which is explained in the following sections, we checked all data marked as positive by the program and obtained 278 DB spectra from LAMOST DR5. However, more DB spectra may have been overlooked by the program owing to the positive samples from the SDSS. We then chose 100 DB spectra, including DB, DBA, DBZ and DB binaries, with

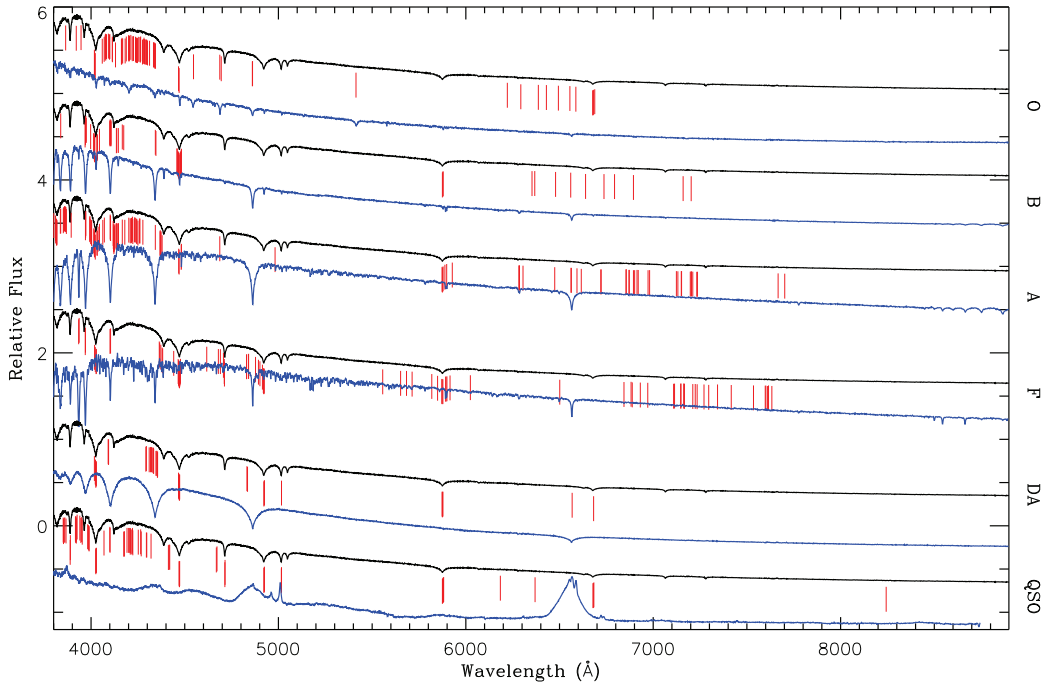

Fig. 4 Features of DB versus O, B, A, F, WD(DA) and QSO, from top to bottom, respectively. To make them more explicit, the wavelengths of features are marked with *red lines* above each type of spectrum (*blue*); the DB spectra are plotted in *black*.

relatively high S/N$_g$ as positive samples for the training set. Each group of negative samples was then compared with this set of positive samples for rarity.

2.2 Data for the Training Process

The redshifts of these positive samples and DB spectra in the negative samples from the LAMOST catalog were not correct because they had been measured by using non-DB templates. Hence, we needed to recalculate the z of all data by using DB templates and move the spectra to the rest frame. To acquire more comprehensive and accurate values, full-spectral template matching (the core algorithm in the pipeline of SDSS (Lee et al. 2008) and LAMOST (Luo et al. 2015) was used to compute the z values of all spectra.

However, we needed to guarantee that the negative samples were as pure as possible; i.e., we needed to ensure that every spectrum in the negative sets had a correct classification and retained the characteristics typical for it. For each CPS, spectra were selected from all data, and ranked by S/N$_g$ in descending order, as negative samples.

In implementing the algorithm, the SVMs proposed in the past experienced many limitations, such as low recognition accuracy when applied to binary classification from unbalanced datasets, in which negative instances heavily outnumbered positive ones. Remedies have since been developed to solve this problem (Akbani et al. 2004). However, we maintained the balance between positive and negative samples within the group of CPS cases, which means that the number of non-DB spectra in each group of CPS was exactly 100. To obtain more comprehensive results, we set the number of groups within each CPS to five; i.e., there were five groups of positive and negative samples for every spectral type or 1000 spectra in a CPS. The "CPS" in this subsection means the CPS in the training set. More details are available in Table 1.

Table 3 Classification of "Unknown" Spectra in LAMOST DR5

Class[a]	Subclass[a]	N[b]	N in ED[c]
Star	O	7004	131
Star	B	354	62
Star	A	1516	711
Star	F	4580	409
Star	G	3088	263
Star	K	2798	119
Star	M	18 898	331
Star	WD	1783	725
Star	carbon	5799	174
Star	CV	19	0
Galaxy	null	206 287	1084
QSO	null	262 358	1837

[a] "Class" and "Subclass" are adopted from the data archive of LAMOST DR5.
[b] Number of spectral types from "Unknown" portion of LAMOST DR5, classified by full spectral template matching.
[c] These numbers are a small part from column 3, listing the number of spectra.

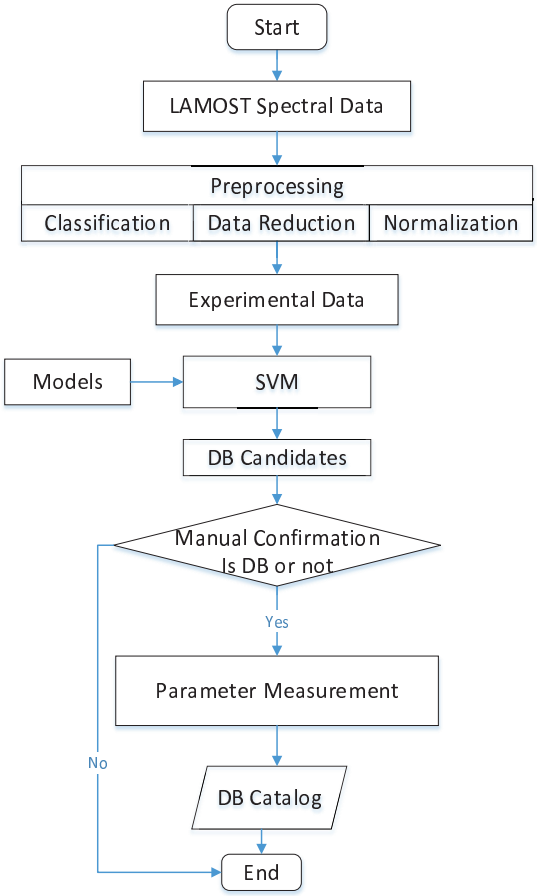

Fig. 5 Flowchart of the recognition procedure. This is the application stage of Fig. 3 and Sect. 3. The final catalog is also generated in this part.

2.3 Data for Recognition

LAMOST DR5 contained >9 million spectral data items, including 153 090 galaxies, 51 133 QSOs, 8 171 443 stars and 642 178 unknown items, among which there were 5 807 771 spectra with $S/N_g > 10$.

Compared with stellar templates from the SDSS Pipeline, templates in the LAMOST Pipeline contain more subclasses, and can provide a more accurate stellar classification for about 95% of spectra (Wei et al. 2013). Theoretically, the majority of data with high S/N should be correctly classified by the LAMOST 1D pipeline. As a result, obvious non-DB spectra needed to be excluded in advance; otherwise, it would be inefficient if all spectra were considered in the recognition process.

We used full-spectral template matching, which is described briefly in Section 3.1, to classify all spectra from LAMOST DR5. The only distinction, however, in our application of this process was that the DB templates were replaced by all templates from the LAMOST 1D pipeline (Luo et al. 2015). Moreover, the χ^2 value between the best and second-best fit was also considered to confirm the final type. Then, both the class and the redshift of each spectrum could be obtained from the best-fitting template. After this procedure was executed, some spectra were discarded from the ED while others were maintained, as is shown with the red and blue bars, respectively, in Figure 1.

It is worth noting that there was a subclass of the DR5 category called "DoubleStar," which is not a typical stellar type. The spectra with "DoubleStar" classification mostly represented the binary companion of a WD, or an early-type object, and an M star, such as DA+M (M-P_S-F is 56264-HD090427N432630M01_07-094; Fig. 2a) (Ren et al. 2014; Guo et al. 2015; Kleinman et al. 2013; Girven et al. 2011; Silvestri et al. 2007) or A+M (M-P_S-F is 55858-F5909_04-104; Fig. 2b). Both spectra plotted in this figure were smoothed by convolving them with a Gaussian function, where $\sigma = 1$ and $\mu = 10$. The components in panel (b) are A2 type and M3 giant stars. These kinds of spectra exhibit characteristics of both types of stars, features of which were more explicit in the blue and red wavebands, respectively. Therefore, we placed all spec-

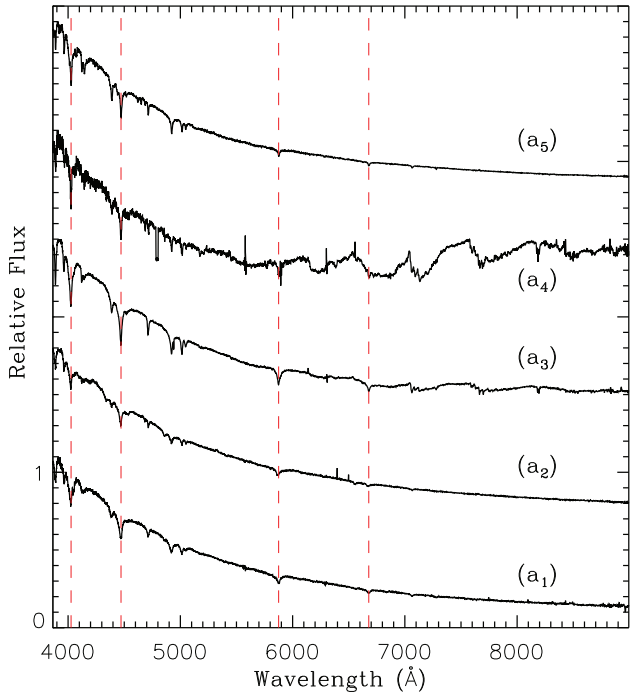

Fig. 6 Five clustering centers of DBWDs from LAMOST DR5. Some He I lines (4026, 4472, 5876 and 6678 Å) are marked in *dashed lines*. Centers (a_3) and (a_4) are both binaries: DB and M type stars.

Table 4 Results of Experiment and Evaluation of the Algorithm Model

Class[a]	Subclass[a]	ED[b]	Candidate[c]	DB[d]	Ratio[e]
Star	O	508	265	262	99.5%
Star	B	9757	129	2	98.7%
Star	A	108 367	2278	2	97.9%
Star	F	863 200	16 404	3	98.1%
Star	G	10 401	138	3	98.7%
Star	K	9197	157	1	98.3%
Star	M	12 423	296	10	97.7%
Star	WD	5834	72	2	98.8%
Star	doublestar	2198	98	1	95.6%
Star	carbon	1910	93	1	95.2%
Galaxy	null	89 120	2231	3	97.5%
QSO	null	32 091	544	31	98.4%
Unknown	null	163 977	10 020	30	93.9%
Total		1 608 983	32 725	351	96.7%

[a] "Class" and "Subclass" are adopted from the data archive of LAMOST DR5.
[b] Number of spectra for every CPS in ED.
[c] DB candidate, number of positive samples in every CPS directly derived from SVM.
[d] Number of positive samples in every CPS after visual inspection.
[e] The approximate identification precision when the predication negative samples are all right, i.e. the correct proportion of classification $(1 - \frac{\text{Candidate}-\text{DB}}{\text{ED}})$.

tra specified as "DoubleStar" into the CPS of M, WD, B and A.

Compared with all LAMOST DR5 data, the ED eventually comprised ∼25% on average of the data.

3 TRAINING PROCESS METHOD

In general, the fundamental idea of classification here is to use an SVM as a classifier to sort DBWDs from all spectral data based on features extracted by using LASSO. An SVM is a binary classification-based algorithm (Duan & Keerthi 2005), which means that it focuses on building a model that assigns new examples to one category or another. LASSO is a method of regression analysis in statistics and ML that conducts both variable selection and regularization. It can extract distinctions between datasets.

In the initial step of our experiment, all data were normalized in preprocessing and a redshift measurement was

Table 5 Number of Stellar Templates for Each "Subclass" from the LAMOST 1D Pipeline

Subclass[a]	O	B	A	F	G	K	M	Carbon	CV	DoubleStar	WD	Total
Number	2	2	49	25	24	36	38	3	1	1	2	183

[a] "Subclass" is adopted from the data archive of LAMOST DR5.

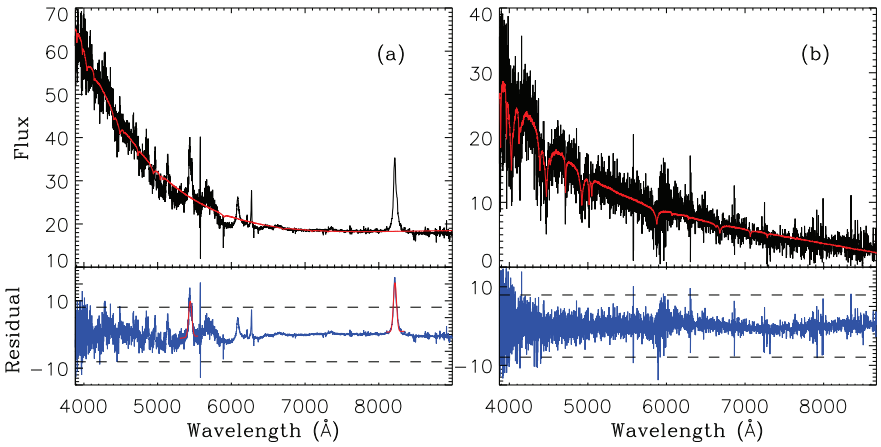

Fig. 7 Two examples of the classification procedure. M-P_S-Fs of panels (a) and (b) are 57460-HD143837N545111M01_04-157 and 57844-HD144325N263140M02_11-038, respectively. The *red lines* in the upper panels are the best fit of the spectra, both of which are assigned to DB type by the software. However, the spectrum in panel (a) is a galaxy, while that in panel (b) is a DBWD. The lower panels display the residuals (*blue*) obtained by flux – best-fit, with the *black dashed lines* indicating the $\pm 3\sigma$ range of the residuals. The *red lines* on the residuals are the results of a Gaussian function fitting at a width of 50–100 Å.

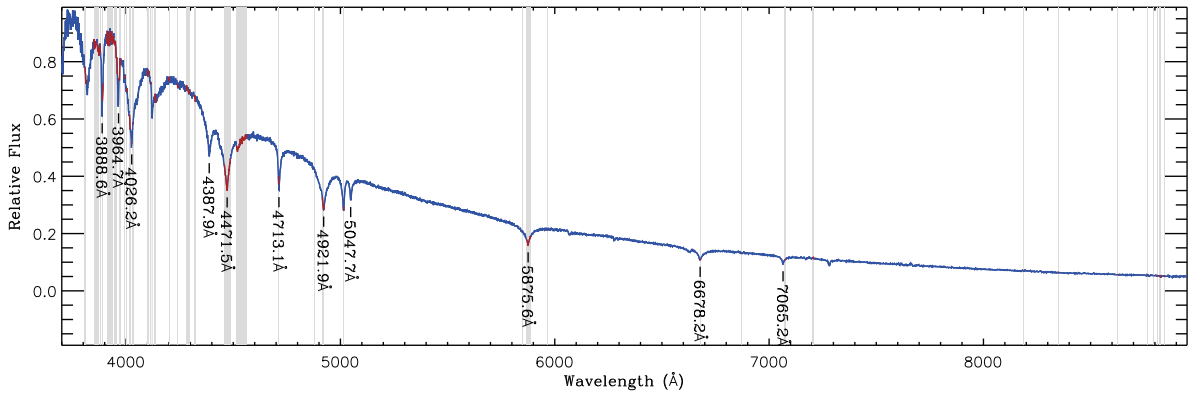

Fig. 8 DB features combined from all CPSs. A DB spectrum is employed to indicate the wavelengths of features. All features are shown in red in the spectrum over a light gray background.

made for the positive samples. The main body of the experiment consisted of separating the DBs from other types using a binary classification module of the SVM, followed by feature extraction with LASSO. If the accuracy (the ratio expressing how many samples are recognized correctly) was not sufficiently high, then optimization was needed; i.e., removing contamination of the DBs from negative sample sets and restarting of this loop. Following the completion of the training process, unique features of the DBs and hyperplanes from each group were derived by using LASSO and an SVM, respectively. The entire procedure conducted in KONG2018 was similar to that in this paper, and is illustrated in Figure 3. Therefore, most of the following parts are only briefly introduced in this section.

3.1 Data Preprocessing

1. The normalization of the positive and negative samples is given by

$$\hat{\mathbf{x}} = \frac{\mathbf{x} - \bar{\mathbf{x}}}{\sigma_{\mathbf{x}}},$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ represents a spectrum, n ($n > 0$) is the number of points, and $\bar{\mathbf{x}}$ and $\sigma_{\mathbf{x}}$ are the mean value and standard deviation of $\mathbf{x}$, respectively.

Table 6 Results of Classification Using Templates Including DB for the Three Groups

	G0[a]	G1[b]	G2[c]
TP	305	303	303
FP	11	13	13
TN	35 987	36 785	36 518
FN	538	40	7
Accuracy	98.5%	99.8%	99.9%
Ratio[d]	1.5%	0.1%	0.0%

[a] Include DB templates, no feature space, no other criterion. [b] Include DB templates, in the feature space, no other criterion. [c] Include DB templates, in the feature space, with some additional criteria. [d] The ratios refer to FN/(TN+FN), which means the percent of how many non-DB spectra are mistaken for DBWDs.

Table 7 Classification results grouped by S/N$_g$. The total number of column 2 (G2) corresponds to the sum of TP and FP from the fourth column in Table 6.

S/N$_g$	G2[a]	DB[b]	Ratio (%)[c]
2–5	41	34	17.0
5–10	96	92	4.2
10–20	84	82	2.4
> 20	95	95	0.0

[a] The number of DB spectra.

[b] The number of DB spectra classified properly by the LAMOST 1D Pipeline (TP).

[c] Proportion of program classification errors ($1 - \frac{\mathrm{DB}}{\mathrm{G2}}$).

The component x_i represents the flux of the spectrum $\mathbf{x}$, $i \in \{1, 2, \ldots, n\}$.

2. A redshift measurement for positive sample groups was made using full-spectral template matching: Known DB spectra with high quality were used as templates to calculate the redshift of the positive samples. Simply put, this was approached as a χ^2 minimization problem. The pseudo continua of templates were reshaped at the beginning to be consistent with the spectrum. Afterwards, the distance between a template and the spectrum at each step within a specific redshift range was calculated. Finally, z was derived from the best fit — the template that reached the minimum χ^2.

3.2 Feature Extraction

In general, features include the continuum and some typical spectral lines when a spectrum is recognized. We believe that the pseudo continua of templates should be reshaped to be consistent with that of the spectrum before the distance (χ^2) between a template and the spectrum is calculated. However, for classification using an algorithm, all data points are not equally important and the continuum may not have much of an effect. Some positions of the line wings rather than centers may be more sensitive in distinguishing a DB from other types of spectra (O-/B-/A-type star, galaxy, QSO, etc.).

Table 8 DB Features near He I Lines

He I Line (Å)	Wavelength (Å)	Width[a] (Å)
3819.6	3807.0 – 3813.3	6.3
3888.6	3849.3 – 3879.7	30.4
	3889.4 – 3892.2	2.8
	3911.0 – 3921.0	10
	3931.8 – 3940.9	9.1
3964.7	3944.5 – 3959.1	14.6
	3967.2 – 3972.8	5.6
4026.2	3992.0 – 3993.0	1.0
	4001.2 – 4003.1	1.9
	4014.1 – 4019.7	5.6
	4027.0 – 4028.1	1.1
	4031.7 – 4032.7	1.0
4120.8	4100.0 – 4105.8	5.8
	4134.2 – 4139.0	4.8
	4199.4 – 4201.4	2.0
4387.9	4239.2 – 4242.3	3.1
	4280.4 – 4286.4	6.0
	4291.3 – 4297.3	6.0
	4318.0 – 4324.1	6.1
4471.5	4456.7 – 4487.1	30.4
	4513.3 – 4564.1	50.8
4713.1	4710.7 – 4711.9	1.2
4921.9	4877.4 – 4878.6	1.2
	4915.7 – 4924.9	9.2
5015.7	5015.2 – 5016.4	1.2
5875.6	5849.1 – 5850.5	1.4
	5868.0 – 5888.4	20.4
6678.2	6677.1 – 6681.8	4.7
	6872.1 – 6873.8	1.7
7065.2	7071.3 – 7078.8	7.5
	7202.6 – 7211.0	8.4
	8188.2 – 8190.2	2.0
8361.7	8350.1 – 8354.0	3.9

[a] The lengths of features. Some features, $<\sim$1.5 Å wide, are only one data point.

During this experiment, LASSO was employed to extract particular features of DB and other kinds of spectra from LAMOST DR5.

We built five sets of negative samples for each CPS as control groups, for the features might not exactly have been identical when the quality or parameters were changed even slightly. We combined features from each group of a CPS into one as final output. Based on our previous research, the full wavelength range (3900–8900 Å) is employed in this experiment.

Following the training process, we obtained conclusions similar to those in section 3 of KONG2018. The features varied from one CPS to another; features from each group under the same CPS had subtle differences at certain wavelengths, owing to variations in line strength and width caused by changing parameters. Some features were also not symmetrical with respect to a spectral line.

Table 9 Classification of DBWDs in LAMOST DR5

Type	Number of stars	Number of spectra
DB	207	249
DB+M[a]	15	16
DBA[b]	41	49
DBO	13	14
DBZ	9	10

[a] One can find the subtype and RV of the M companion in the online table.

[b] A few DBA cases are actually DBAZ or DBAO, but here they are all counted in "DBA."

3.3 Feature Collection

We extracted the features from all groups and combined those within each CPS into an array as output.

Figure 4 shows the features of some CPSs: O, B, A, F, DAWD and QSO, from top to bottom respectively, marked using short red lines above each spectrum. In general, many features on both sides of all spectral lines were imperfectly symmetrical, probably due to the asymmetry of the spectral line and the continuum. Moreover, the range of wavelength for features on the right side of a spectral line, in some cases, was wider than that on the left. The number of metal elements increased with the order of stellar type from early to late, which corresponded to the rise in the number of features in the red band.

3.4 SVM

Given a set of training samples divided into positive and negative categories in a dual clustering system, an SVM algorithm builds a robust binary linear classifier model that assigns new data to either type. We adopted the LIBSVM (Chang & Lin 2011) software to select DB spectra from all spectra based on the DB features.

Parameter selection is crucial for any ML algorithm, and LIBSVM is no exception. LIBSVM provides four basic kernels: linear, polynomial, sigmoid and radial basis kernel function. We randomly selected thousands of spectra with various types and S/N$_g$ in LAMOST DR5, together with all the known DB spectra, as the test dataset. Then all the four kernel functions were employed to execute the recognition process. Afterwards, we inspected all the DB candidates recognized by LIBSVM. Many non-DB spectra were mis-classified as DB when using polynomial or sigmoid kernel functions, and the precision ratio could be less than 80%. However, when the linear or radial basis function kernels were adopted, the precision ratio would reach above 95%. These tests have shown that linear and radial-basis function kernels provide better results in terms of discriminating spectral data. The built-in 10-fold cross-validation was utilized to determine all other parameters automatically by using the LIBSVM software.

Table 10 Newly Spectroscopically Identified DB+M Binaries

Designation	Type	RV$_{DB}$ (km s^{-1})	RV$_{M}$ (km s^{-1})
J225336.81+081608.1	DB+M3	-17.7 ± 3.4	90.4 ± 3.2
J130716.45+170220.9	DB+M1	103.0 ± 12.0	10.6 ± 13.3
J025521.23+210444.1	DB+M0	47.3 ± 3.1	-18.7 ± 2.4
J233741.62+454318.0	DB+M3	-26.9 ± 15.2	19.9 ± 32.9

Notes: An explanation of the fields within this table can be found in Table 11.

Given that all data in the training set were assigned to the correct type, some measures of information retrieval and statistical classification were employed to evaluate the performance of the algorithm in terms of accuracy, precision and recall. We applied the true positive (TP) standing for the correct prediction of the positive category, false positive (FP) for that of the incorrect positive category, and false negative (FN) and true negative (TN) for incorrect and true negative classifications, respectively. Because almost all positive samples were recognized properly by LIBSVM, the recall, TP/(TP+FN), was $\sim$99.9%, which demonstrates the percentage of positive samples predicted correctly. The accuracy – (TP+TN)/(TP+FN+FP+TN) – reflects the program's ability to determine the entire sample, which means identifying positive samples as DB spectra and negative samples as non-DB ones. The percentages of mean accuracy and precision – TP/(TP+FP) – were 99.7% and 99.1%, respectively, which indicate high stability and reliability of the algorithm.

3.5 Verification of Method Validity

In our previous work (KONG2018), we applied this ML method to extract features and search DB spectra in the SDSS DR12 (Alam et al. 2015) and DR14 (Abolfathi et al. 2018). In all the spectra of 2700 DBs from SDSS DR12, we spectroscopically identified 704 cases that were not in the catalogs in Kepler et al. (2015, 2016); Kleinman et al. (2013), which verifies the validity of our method. In general, this ML method could be applied to search for DB spectra more effectively.

4 RECOGNITION

4.1 SVM Input

Before the recognition process, each "Unknown" spectrum needed to be assigned a specific type and become a member of a certain CPS. The associated values are presented in Table 3.

Only those with relatively high uncertainty from template matching were added to the ED. Feature planes derived from the training process were also employed as in-

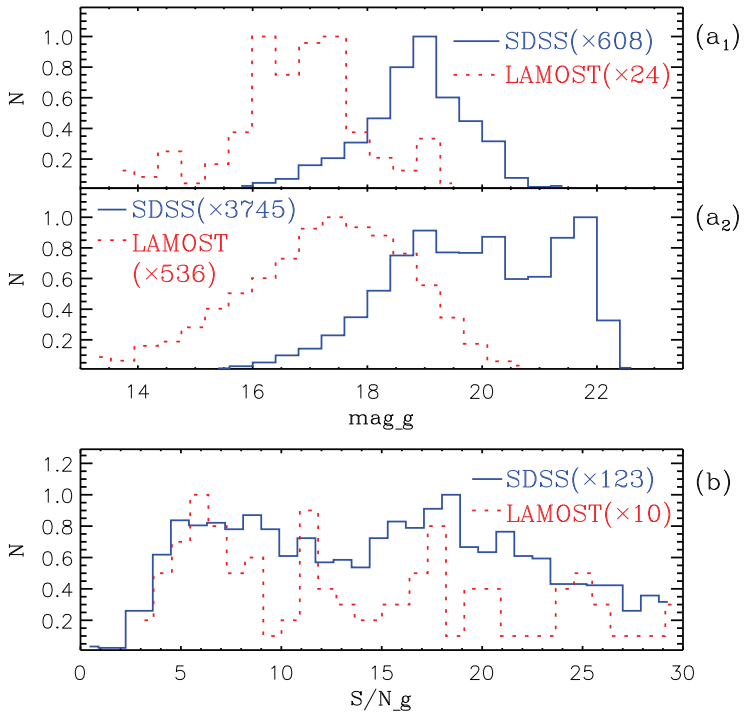

Fig. 9 Distributions of g-band magnitude for (a_1) DBWDs and (a_2) WDs. (b) Histograms of S/N$_g$ for DBWDs. *Red dotted lines* and *blue solid lines*, respectively, represent the amount of spectral data from LAMOST DR5 and SDSS DR12, all of which are normalized to a maximum of 1 in each panel. In the three panels, the actual quantities are multiplied by factors of 608, 3745 and 123 for SDSS and 23, 536 and 10 for LAMOST, all of which are marked in parentheses after their sky survey names.

put to LIBSVM. A flowchart is shown in Figure 5 which demonstrates the entire process described in this section.

4.2 Recognition and Results

A total of 13 CPSs, 1 608 983 spectra, were involved in the recognition procedure. Similar to the preprocessing procedure, we normalized them and moved them to the rest frame. Some hyperplanes in the feature space were used in LIBSVM to distinguish the DBs from all data in the ED.

After inspection, we selected 351 DB spectra from 287 objects in LAMOST DR5, among which 53 stars were newly identified. We provide the results of our experiment in Table 4, which indicates that the mean percent correct from the algorithm was ∼96.7% if all labeled negative samples were real non-DBs.

Clearly, most DBs were identified as O, QSO and Unknown in DR5. Among all 30 DBWDs from the "Unknown" group, the best fits consisted of 21 O stars and four QSOs, which indicate that these types of spectra and those of the DBs had much in common. In a different way, it would often be difficult to distinguish DBs from other types when using the full spectrum to match them, instead of particular wavelengths (features).

5 CONSTRUCTION OF DB TEMPLATES

We provide a solution to DBWD recognition for the LAMOST 1D pipeline.

5.1 Spectral Data

All DBWD spectra with $\mathrm{S/N}_g > 10$ from both SDSS DR14 and LAMOST DR5 were used to build the DB templates by using k-means clustering (MacQueen 1967). We set 12 clustering centers for the clustering process. The centers containing <50 spectra with He I lines that were too weak to be observed were abandoned. The abandoned centers were noise, and helped little when the pipeline identified the spectra. Finally, five clustering centers that corresponded to the most numerous spectra were selected, as illustrated in Figure 6. In general, a majority of DB spectra was recognized by these five clustering centers.

Table 5 shows that the stellar templates of the LAMOST 1D pipeline consisted of 183 spectra (Wei et al. 2013) ranging from O- to M-type stars, together with some particular spectral types; e.g., CV, carbon, WD and DoubleStar. These two "WD" templates were DA WD spectra. To obtain a more comprehensive result, we randomly selected 4000 spectra from each subclass of the

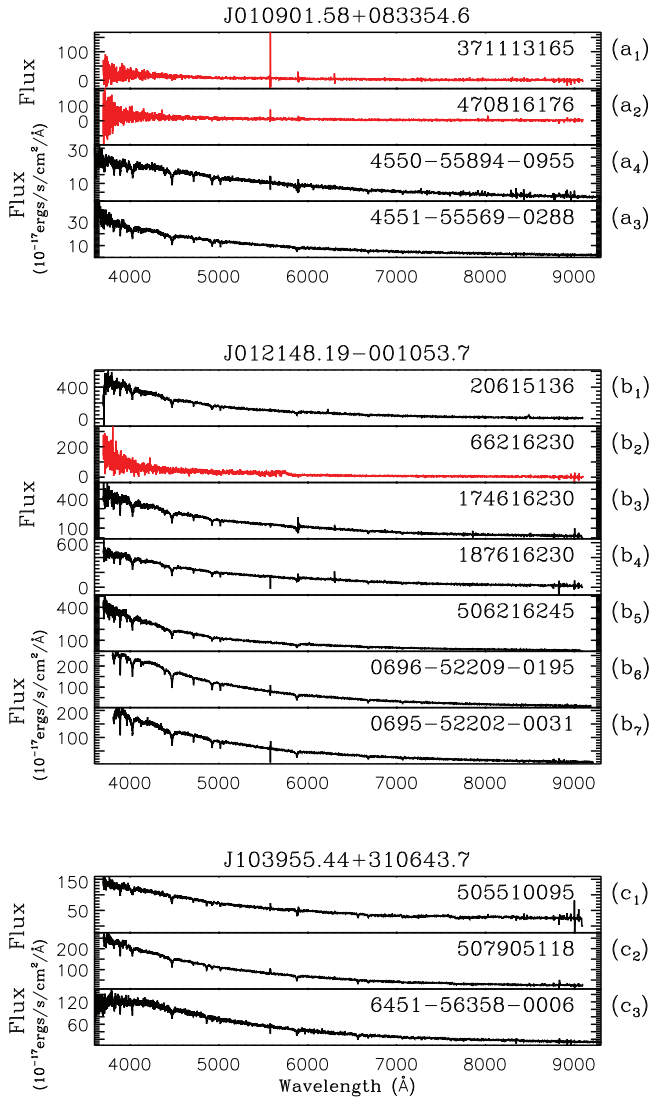

Fig. 10 Three DBWD stars released by both LAMOST and SDSS. (a) J010901.58+083354.6 was recorded twice on either side (a_1 and a_2 are from LAMOST DR5 and a_3 and a_4 are from SDSS DR12). (b) For J012148.19-001053.7, the counts are five (b_1–b_5) and two (b_6 and b_7). (c) For J103955.44+310643.7, the counts are two (c_1 and c_2) and one (c_3). Obsid (LAMOST) and plate-mjd-fiber (SDSS) from catalogs are labeled above their spectra. From top to bottom, their S/N$_g$ values are 2.76, 1.75, 22.45, 23.07 (a_1–a_4), 22.10, 3.48, 26.2, 34.25, 29.47, 50.64, 30.96 (b_1–b_7), 29.92, 44.15 and 22.24 (c_1–c_3), respectively. Because of a feature caused by noise, the three spectra drawn in *red* were not selected by our algorithm.

LAMOST data archive, together with known DBs, as the dataset.

5.2 Classification and Results

The five DB clustering centers were first added to the stellar templates. Considering the features obtained in this paper, we built three control groups and compared the recognition results.

- G0: We directly used the LAMOST 1D pipeline to classify all spectra from the dataset, using templates including DBWDs.
- G1: The difference from G0 was that we performed classification in the feature space and recognized using the DB templates.
- G2: We added criteria to the final stage of classification to improve accuracy (with G2 being the upgraded version of G1). These criteria are explained below.

After classification, we inspected samples identified as DBWDs and compared the results with types from the LAMOST DR5 catalog. Table 6 shows the comparison results. For all groups, a majority of DB spectra was identified from the datasets in general. However, the program mistook many other types of spectra for DBWD spectra if

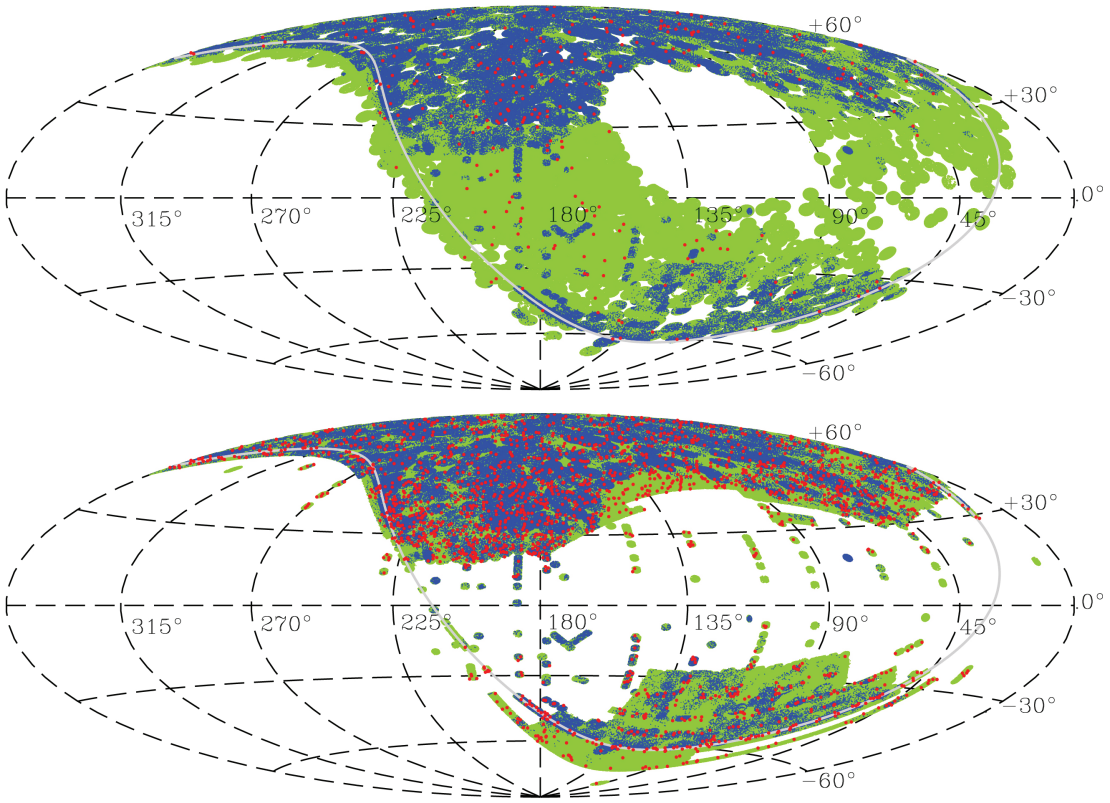

Fig. 11 Footprints of LAMOST DR5 (*upper panel*) and SDSS DR14 (*lower panel*). A band with declination equal to 0 is indicated by a gray line. These targets observed by both surveys are colored blue, while all DBWDs are in red circles. In the sky area where the Galactic latitude is between $-30°$ and $+30°$, LAMOST DR5 added many observations.

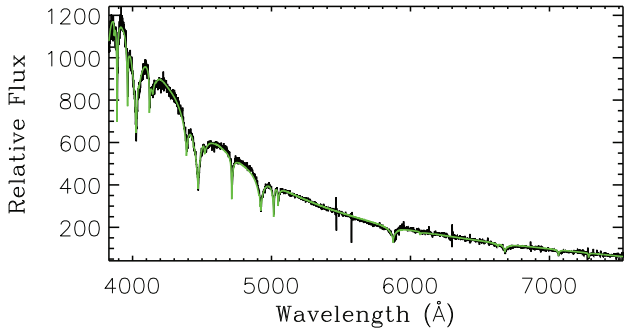

Fig. 12 One example of template matching for parameter measurement. The *black spectrum* is a DBWD (specid is 20150112GAC054N40B101161_v2.9.7) and the *green one* illustrates the best-fitting template.

all of the wavelengths were considered. The use of features offered help in reducing instances of misclassification.

Most DBWD spectra were recognized by the DB templates using the LAMOST 1D pipeline. The majority of DBWDs (FP) discarded by the software was dropped because of strong hydrogen lines (DBA), metal lines (DBZ) or He II lines (DBO), whereas others were misclassified due to noise. Table 7 signifies that noise could have a negative impact on classification results.

From the TN and FN rows in Table 6, we can conclude that some non-DB spectra were mistaken as DBWD spectra using templates containing the DB spectra. Group G0 indicates that >500 non-DB spectra were assigned to the DB type, mainly owing to low S/N. After applying template matching in the feature space, there was a significant reduction in the error (the last row of Table 6).

We then added criteria to the classification process. Because this process is only preformed in the feature space, some typical lines in other types of spectra might not

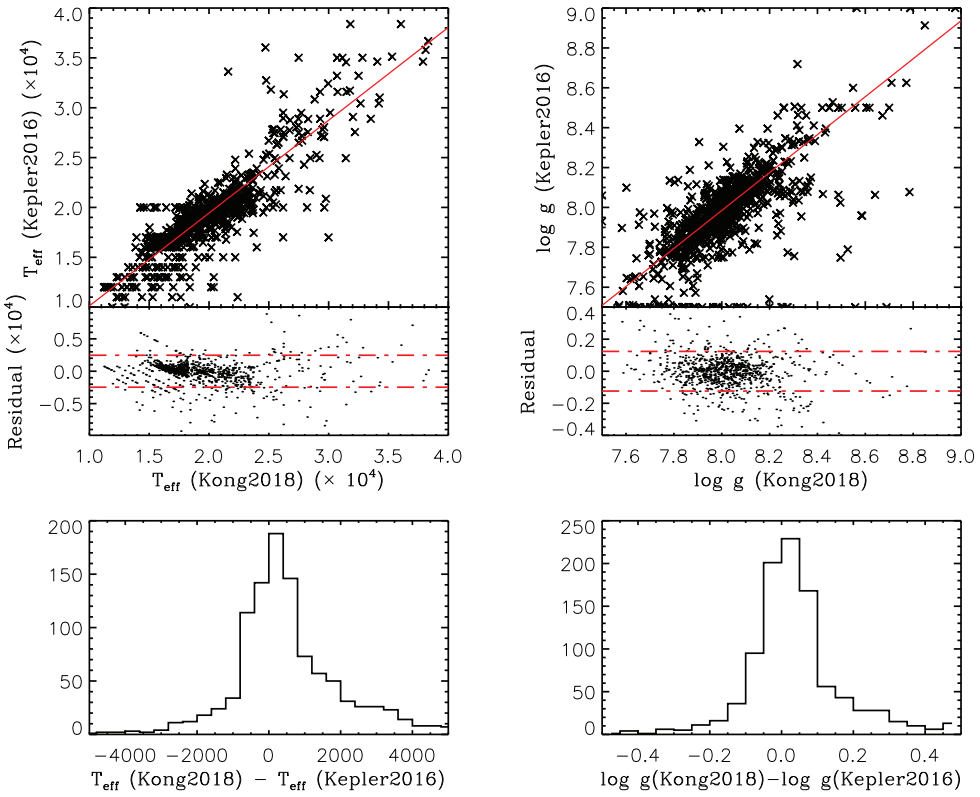

Fig. 13 Parameters of DBWDs calculated by both Kepler et al. (2016) and us. The *red solid lines* in the top two panels are the straight-line fitting results, $y = 0.92x + 927$ for $T_{\rm eff}$ and $y = 0.95x + 0.38$ for $\log g$ (x for parameters from KONG2018 and y for those from Kepler et al. (2016)). The two panels below are the residuals of the fits with $\sigma = 2482$ K and 0.12 cgs for $T_{\rm eff}$ and $\log g$, respectively.

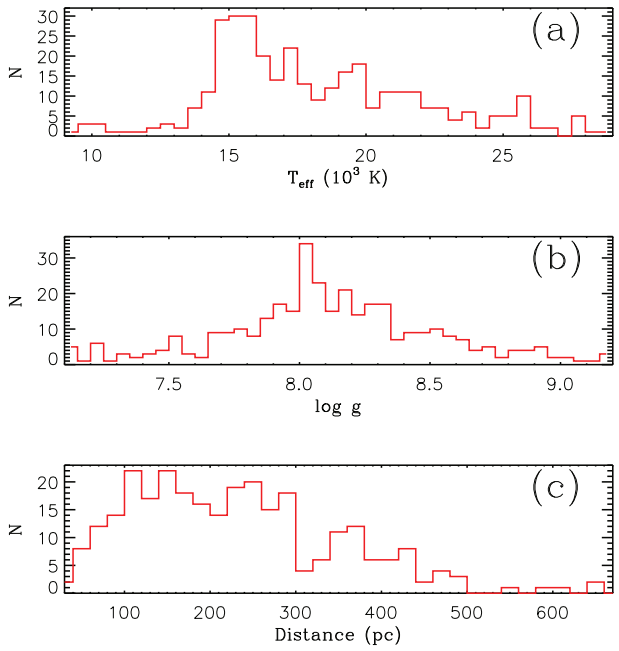

Fig. 14 Distributions of $T_{\rm eff}$, $\log g$ and distance from the Sun for DBWDs from LAMOST DR5.

Table 11 Descriptions of Columns in the Entire Online Table of DBWDs

Column No.	Heading	Description
1	Designation	LAMOST object name (LAMOST 2000J+)
2	M-P_S-F	LAMOST Modified Julian Date-Planid_SPid-Fiberid
3	Type	Classification of objects derived from ML method
4	RV_{DB}	Radial velocity and uncertainty of each spectrum ($km\,s^{-1}$)
5	RV_{M}	Radial velocity and uncertainty of M companions ($km\,s^{-1}$)
6	T_{eff}	Effective temperature (K)
7	$\log g$	Surface gravity (cgs)
8	FUV	Magnitude of FUV from *GALEX* (mag)
9	NUV	Magnitude of NUV from *GALEX* (mag)
10	S/N_g	g band of S/N from catalog of LAMOST DR5
11	PmRA	Proper motion at J2015.5 in the direction of RA ($km\,s^{-1}$)
12	PmDEC	Proper motion at J2015.5 in the direction of DEC ($km\,s^{-1}$)
13	Parallax	Absolute stellar parallax of the source at the reference epoch J2015.5 ($mas\,yr^{-1}$)
14	Mass	Obtained from `http://www.astro.umontreal.ca/~bergeron/CoolingModels/` ($M_{\odot}$)
15	Age	Obtained from `http://www.astro.umontreal.ca/~bergeron/CoolingModels/` (Myr)

Notes: –9999: there is no corresponding value.
The full table is online at `http://www.raa-journal.org/docs/Supp/ms4311etable.rar`.

have been considered. In panel (a) of Figure 7, three emission lines in the spectrum are not included in the feature space, and were omitted by the program. We checked the residuals between the spectrum and the best fit, and found that many rejected spectral lines could extend beyond the $\pm 3\sigma$ region of the residuals. For comparison, we illustrate an example of successful recognition of the DBWDs in panel (b) of Figure 7. Finally, we fit these extended points with a Gaussian function if the spectrum was considered a DBWD by the LAMOST 1D pipeline. If more than one line was fitted successfully, then the pipeline got rid of the DB templates and redid the classification.

6 ANALYSIS

6.1 Features of DBWDs

In Section 3.2, we obtained the differences (features) between the DBWD spectra and others types of spectra. We then combined features from all CPSs and discarded those existing only in fewer than three groups. As Figure 8 and Table 8 show, almost every He I line center was recognized as a DBWD feature. Moreover, some locations within the line wings were also typical features of DBWDs, as discussed in Sections 3.2 and KONG2018.

In the previous section, we tested the effectiveness of these features for spectral classification using the LAMOST 1D pipeline. We think that they can also help when using algorithms to select and analyze the DB spectra.

6.2 Comparison with the Literature

Over 1500 pure DB objects have been identified in the literature, including some DB+M double stars. In this paper, we present 351 DBWD spectra in LAMOST DR5 that corresponded to 287 stars, among which 53 objects were newly spectroscopically confirmed. Table 9 provides the subtypes of DBWDs in our catalog.

We present these newly identified DBWDs online at `http://www.raa-journal.org/docs/Supp/ms4311etable.rar` and the DB+M binaries in Table 10. All information concerning the DBWDs from LAMOST DR5 can be found in the online table, the descriptions of which are given in Table 11.

6.3 Comparison with SDSS Spectra

6.3.1 Distinction

Thus far, the largest datasets of DBs, even of WDs, are from SDSS, in which the number of released WD spectra was 36 093 and 38 575 in DR12 and DR14, respectively. By contrast, the size of LAMOST DR5 is only 9211. The number of helium-dominated WD spectra, more explicitly, from SDSS DR14 was $\sim$2800 in the literature. However, we have discovered 300 more DB spectra from LAMOST DR5, despite its release of $\sim$9 million data items.

These two ratios of WDs with respect to the total number of data items in SDSS and LAMOST are so different mainly owing to the limiting magnitude of the telescope, source selection and data quality. These are discussed in turn.

1. Limiting magnitude.

   With its ability to capture 4000 spectra in a single exposure, LAMOST can reach a limiting magnitude of 16–17 mag (Luo et al. 2015), whereas the observable spectrum of the SDSS photometric camera is brighter than 23.2 for g' (Gunn et al. 1998). Considering the low brightness of WDs, compared with LAMOST, SDSS can capture a much higher value.

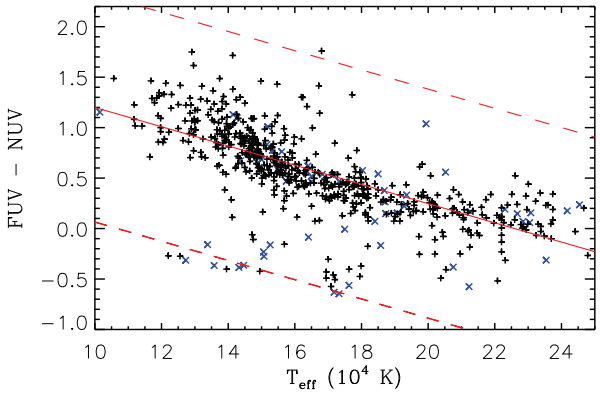

Fig. 15 FUV–NUV as a function of $T_{\rm eff}$. The continuous (*red*) line is the fitting results for DBWDs, with the two *dashed lines* representing the 3σ range.

Table 12 Source Selection of SDSS DR12 and LAMOST DR5

Source[a]	SDSS		LAMOST	
	Number	Ratio[b]	Number	Ratio[b]
QSO[c]	718 810	17.8%	3082	0.03%
galaxy[d]	2 345 709	58.0%	112 031	1.2%
star[e]	260 198	0.6%	8 623 151	95.6%
dwarf[f]	27 508	0.7%	282	0.003%
white dwarf[g]	11 182	0.3%	282	0.003%

[a] Refers to "SourceType" and "Objtype" fields from database of SDSS DR12 and LAMOST DR5, respectively.
[b] The ratio represents a certain type of target accounting for the proportion of all objects, which is derived by dividing column 2 by the number of all spectra, in addition to "NA"/"null"/"SKY" (4 047 254 from SDSS DR12; 9 017 844 from LAMOST DR5).
[c] Sum of numbers from all the fields that include "QSO" (except for "GAL_NEAR_QSO" in SDSS).
[d] Sum of numbers from all the fields that include "GAL" (together with "LRG" in SDSS).
[e] Sum of numbers from all the fields that include "STAR" and "STD" (except for "QSO_STD" in SDSS).
[f] Sum of numbers from all the fields that include "DWARF".
[g] Sum of numbers from "STAR_WHITE_DWARF", "WHITE DWARF_NEW" and "WHITEDWARF_SDSS".

The distributions of g-band magnitude for WDs from SDSS DR12 and LAMOST DR5 are shown in panels (a_1) and (a_2) of Figure 9. Panel (a_1) represents the number of DBWDs from both the literature (Kleinman et al. 2013; Kepler et al. 2015, 2016; Guo et al. 2015; KONG2018) and this experiment, while panel (a_2) accounts for all WDs from the SDSS DR12 and LAMOST DR5 catalogs. It is clear why DBWDs in the LAMOST data archive were so small in number: The majority of DBWDs (WDs) from LAMOST was brighter than those from SDSS by a magnitude of ∼2–4 in the g band.

2. Source selection.

The object selection strategy is crucial to obtaining different kinds of spectral data, and is based on both instrument capability and survey plans. There are always differences between the stars selected from LAMOST and those from SDSS which can lead to a disproportionate number of DBWDs.
We checked the object selection in detail from SDSS[3] and LAMOST[4] databases, and demonstrate the distributions of major sources (QSO, galaxy, star and WD) in Table 12. Tens of thousands of WDs were preserved in the observing strategy of SDSS, whereas only ∼300 were kept in that of LAMOST. In other words, the SDSS project had greater interest in extragalactic objects, and one of the major goals of LAMOST was to collect the spectra of stars in the main sequence.

3. Data quality.

In spite of their similarity in terms of resolution, the qualities of many spectral data from the SDSS and LAMOST DRs were not of the same level. The S/N$_g$ distributions of DBWDs from SDSS DR12 seemed structurally superior to those from LAMOST DR5, as panel (b) of Figure 9 illustrates.
We cross-matched all released spectra from SDSS DR12 and LAMOST DR5 in a circular area with a radius of 3″. There were ∼268 DBWD stars captured by both LAMOST (DR5) and SDSS (DR12), from which we selected multi-observed DBWD spectra. We display them in Figure 10. As panels (b_1)–(b_5) illustrate, most spectra clearly displayed typical spectral lines but some did not. We failed to recognize the three spectra in red, which exhibit few spectral lines and were assigned to "Unknown" by the LAMOST 1D pipeline. However, many DBWD spectra from LAMOST DR5 were similar to, or even better than, those from SDSS DR12. For example, for the

[3] http://skyserver.sdss.org/CasJobs/
[4] http://dr5.lamost.org/sql/s

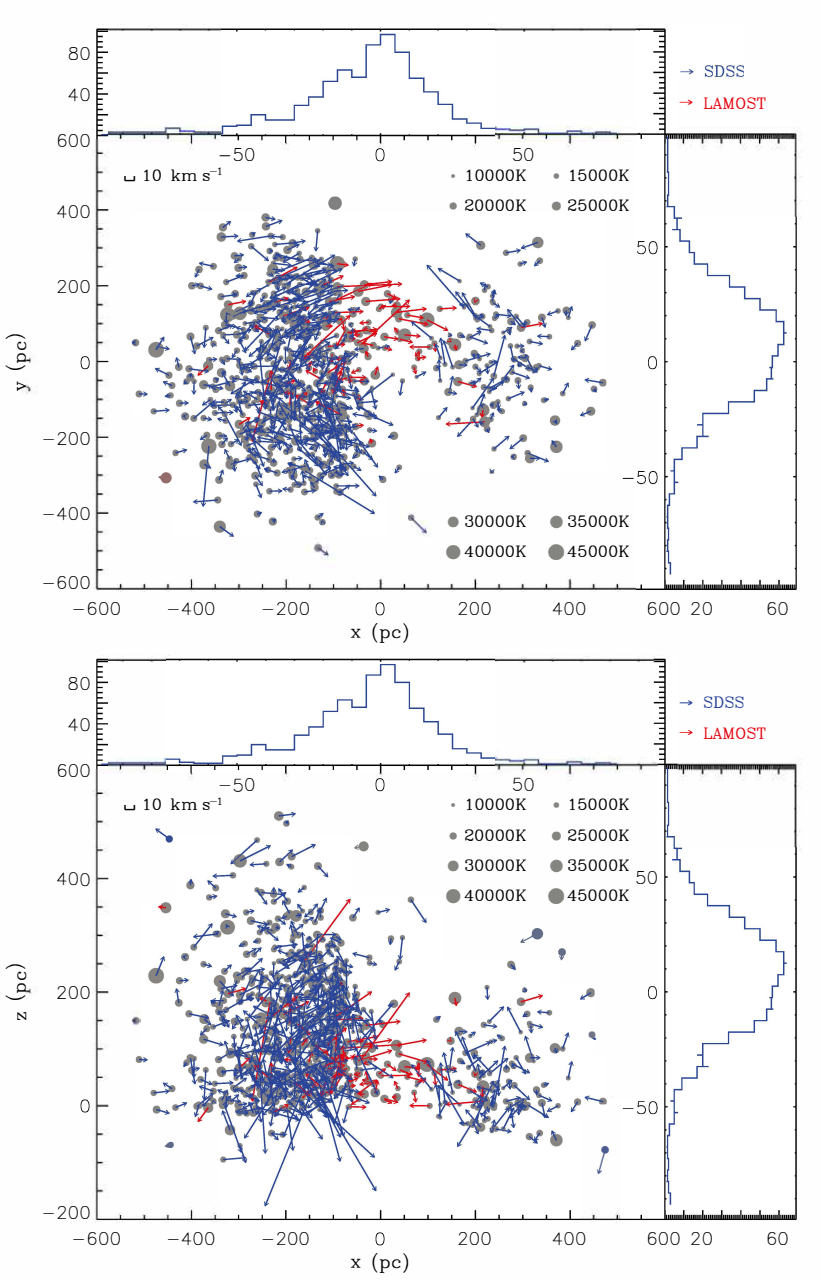

Fig. 16 3D velocities, in the Galactic coordinate system, of high-S/N DBWDs from SDSS DR14 (*blue*) and LAMOST DR5 (*red*). The axes represent the spatial position in parsecs. "x" is the distance from the Sun along the direction of the Galactic center, with that toward the Galactic center being positive. "y" is perpendicular to "x" and follows the right-hand rule. "z" is perpendicular to the xy plane and positive to the north. The top panel is the top view while the bottom one is the side view. All DBWDs are signified by gray circles, the radii of which represent $T_{\rm eff}$. The lines with arrows represent the velocities and their directions on this plane. The two side subplots are histograms of the velocity distribution along the x and y directions, respectively.

three spectra in panel (c), the S/N$_g$ values of LAMOST data ((c$_1$) and (c$_2$)) were 29.92 and 44.15, respectively, while that of the third (c$_3$) from SDSS DR12 was 22.24.

Possibly more spectra from the LAMOST data belonged to the DBWD category, but have not yet been sought by astronomers because of low S/N ratio.

6.3.2 *Connection*

Data from both SDSS DR14 and LAMOST DR5 were low-resolution spectra, with wavelength ranging from ∼3900 to ∼9000 Å. Furthermore, both of the two sky surveys covered most of the northern celestial hemisphere. The footprints of SDSS DR14 and LAMOST DR5 are shown in the Galactic coordinate system in Figure 11.

Table 13 Dispersion of 3D Velocities for DBWDs with $\mathrm{S/N}_g > 10$

M ($M_\odot$)	N[a]	u (km s^{-1})	v (km s^{-1})	w (km s^{-1})	TV[b] (km s^{-1})
0.20 – 0.50	81	9.9 ± 1.6	-39.2 ± 1.2	-5.2 ± 0.9	106.3 ± 1.8
0.50 – 0.80	872	29.5 ± 0.6	11.6 ± 0.8	-13.7 ± 0.3	95.0 ± 1.0
0.80 – 1.20	430	15.0 ± 0.2	-21.2 ± 0.1	-18.7 ± 0.2	97.3 ± 0.2

[a] N is the number of DBWDs in each mass bin.
[b] Total velocity, obtained from $\sqrt{u^2 + v^2 + w^2}$.

6.4 Parameter Measurement

Koester & Kepler (2015) gave DBWD parameters (selected from SDSS DR10 and DR12 with $\mathrm{S/N} > 10$) by applying theoretical model fitting, and discussed their relationships and distributions. Using the DB parameter model provided by D. Koester, we also measured the parameters ($T_{\rm eff}$ and log g) of the DB spectra from LAMOST DR5, employing template matching at wavelengths of He I lines. The fitting results for all the DB spectra were inspected carefully. Figure 12 displays an example of this check. The average fitting errors of $T_{\rm eff}$ and log g were 3.7% and 1.4%, respectively.

The parameters of DBWDs measured by both Kepler et al. (2016) and us were compared. Figure 13 demonstrates that the σ of the residuals for $T_{\rm eff}$ is 2482 K, while that for log g is 0.12 cgs. For the DB spectra whose parameters were illustrated in this figure, the average fitting errors of $T_{\rm eff}$ measured by Kepler et al. (2016) and KONG2018 are 204 and 493 K, respectively; while the average fitting errors of log g are 0.058 and 0.069 cgs, respectively. Although the methods are different, the measurements of parameters rely on the same DB models provided by Koester. Therefore, when considering the uncertainties of the $T_{\rm eff}$ and log g, errors from the two methods should contribute to the residuals, which are about $\sqrt{204^2 + 493^2} \approx 534$ K and $\sqrt{0.058^2 + 0.069^2} \approx 0.091$ cgs, respectively. The σ of the residuals of log g is comparable with the uncertainties of the measurements. Obviously, both the errors of $T_{\rm eff}$ from Kepler et al. (2016) and KONG2018 are relatively small; however, the errors from Kepler et al. (2016) are much smaller. It reminds us that the fitting errors of $T_{\rm eff}$ should be scaled when using them to conduct further study.

We show the distribution of $T_{\rm eff}$ and log g in panels (a) and (b) of Figure 14, respectively. The histogram of $T_{\rm eff}$ was used at 500 K intervals, while that of log g was 0.05. A majority of DBWDs from LAMOST DR5 gathered at $T_{\rm eff} \approx 15\,000$K and log $g \approx 8.0$.

Given the strong intensity in the ultraviolet waveband of the WD spectra, we cross-matched all DBWDs from both LAMOST DR5 and SDSS DR14 with the *Galaxy Evolution Explorer* (*GALEX*), obtaining the far-ultraviolet (FUV) and near-ultraviolet (NUV) magnitudes. Those with errors of $<$0.3 mag for both FUV and NUV and $\mathrm{S/N}_g > 10$ were selected to demonstrate the relationship between $T_{\rm eff}$ and FUV $-$ NUV colors:

$$\mathrm{FUV} - \mathrm{NUV} = -9.5 \times 10^{-5} T_{\rm eff} + 2.15 \,. \qquad (1)$$

In total, only 69 DB spectra from LAMOST DR5 were included. See Figure 15 for more details.

We also looked up the masses and ages of DBWDs based on the $T_{\rm eff}$ and log g values in the Synthetic Colors and Evolutionary Sequences of Hydrogen- and Helium-Atmosphere White Dwarfs website[5] (Holberg & Bergeron 2006; Kowalski & Saumon 2006; Tremblay et al. 2011; Bergeron et al. 2011). The online catalog will include these parameters with descriptions in Table 11.

6.5 3D Velocity

The *Gaia* satellite (Gaia Collaboration et al. 2016) released DR2 (Gaia Collaboration et al. 2018) in April, 2018 providing proper motions in right ascension (RA) and declination (DEC), parallaxes (Luri et al. 2018), and photometry (Arenou et al. 2018).

All DBWDs from SDSS DR14 and LAMOST DR5 were cross-matched with data from *Gaia* DR2. We adopted 1/parallax from *Gaia* DR2 to obtain the distances. Most of the DBWDs survived within a distance of 500 pc from the Sun. By applying the criteria of parallax > 0, parallax_error $<$ parallax/5 (Luri et al. 2018) and $\mathrm{S/N}_g > 10$, 1200 DBWDs remained.

We calculated the 3D velocities of the remaining DBWDs using radial velocities (RVs), RAs and DECs obtained from the LAMOST 1D pipeline, and parallaxes and proper motions along the direction of RA and DEC from *Gaia* DR2. By employing the local standard of rest (LSR) from Huang et al. (2015), the 3D velocities (u, v and w in Galactic coordinates) and the locations of the DBWDs are illustrated in Figure 16.

Moreover, we investigated the dispersion of the 3D velocities at different mass levels for the DBWDs. Table 13 shows that low-mass DBWDs, displaying the kinematics of old stars, have a higher velocity dispersion ($\sim$1.5 km s^{-1}). The dispersion decreased along all three

[5] http://www.astro.umontreal.ca/~bergeron/CoolingModels/

directions as the masses increased. This is consistent with the conclusion in Wegg & Phinney (2012), the authors of which used proper motion from PG in the velocity calculation. We employed more precise data and derived a conclusion for DBWDs, which is similar to that for DA WDs given by Wegg & Phinney (2012).

For better illustration, an animation of the 3D velocity distribution was produced, which is available online. From Figure 16 and the online animation (`www.raa-journal.org/docs/Supp/uvw.gif`), one can see that the known DBWDs seem to be more cluttered in the neighborhood of the Sun. Furthermore, most observed DBWDs are concentrated near the Galactic anti-center, and their motions seem to be disorganized.

7 SUMMARY AND CONCLUSIONS

In this study, we spectroscopically identified 287 DBWDs from 351 spectra in the LAMOST DR5, including 53 new objects, using ML, i.e., LASSO and an SVM. The DBWD features were obtained by a combination of all CPSs, as provided in Figure 8 and Table 8. We then constructed DB templates using DBWDs from SDSS DR14 and LAMOST DR5, and added them to the stellar templates of the LAMOST 1D pipeline. By experimenting with several control groups of data, we proposed methods that allow the pipeline to classify DBWDs more accurately. The difference in numbers of DBWDs between SDSS DR14 and LAMOST DR5 was analyzed from three aspects: limiting magnitude, source selection and data quality. Finally, we measured the parameters of all DBWDs using DB models provided by D. Koester. Most DBWDs were found to have $T_{\rm eff} \sim 15\,000$ K ranging from 14 000 K to 26 000 K, and log g ranging from 7.5 to 8.8. Using the *Gaia* DR2, we calculated the 3D locations and velocities of the DBWDs from SDSS DR14 and LAMOST DR5, and have shown them in Figure 16 and an online animation. Their velocity dispersion decreased with increasing mass, which was consistent with the pattern of DA WDs.

At the same time, the application of DB templates and features may require some other optimization to obtain more comprehensive classification results. We need to consider the relationship between the χ^2 values corresponding to the same template and the distributions of χ^2. This will be the subject of our next investigation in this area.

Acknowledgements This work was funded by the National Basic Research Program of China (973 program, 2014CB845700) and the National Natural Science Foundation of China (Grant No. 11390371/4). The Guo Shou Jing Telescope (the Large Sky Area Multi-object Fiber Spectroscopic Telescope, LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project was provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences. We thank Detlev Koester for providing us with his grid of DBWD model spectra. The masses and ages of DBWDs are available at `http://www.astro.umontreal.ca/~bergeron/CoolingModels`. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for DPAC was provided by national institutions, in particular institutions participating in the *Gaia* Multilateral Agreement.

References

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543

Abolfathi, B., Aguado, D. S., Aguilar, G., et al. 2018, ApJS, 235, 42

Akbani, R., Kwek, S., & Japkowicz, N. 2004, in European Conference on Machine Learning (Berlin, Heidelberg: Springer Berlin Heidelberg), 39

Alam, S., Albareti, F. D., Allende Prieto, C., et al. 2015, ApJS, 219, 12

Arenou, F., Luri, X., Babusiaux, C., et al. 2018, A&A, 616, A17

Beauchamp, A., Wesemael, F., Bergeron, P., et al. 1996, in Astronomical Society of the Pacific Conference Series, 96, Hydrogen Deficient Stars, eds. C. S. Jeffery, & U. Heber, 295

Bergeron, P., Wesemael, F., Dufour, P., et al. 2011, ApJ, 737, 28

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

Cortes, C., & Vapnik, V. 1995, Machine learning, 20, 273

Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 1197

Duan, K.-B., & Keerthi, S. S. 2005, in International Workshop on Multiple Classifier Systems, Which is the Best Multiclass SVM Method? An Empirical Study, eds. Oza, N. C., Polikar, R., Kittler, J., & Roli, F., Springer, 278

Eisenstein, D. J., Liebert, J., Harris, H. C., et al. 2006, ApJS, 167, 40

Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72

Fontaine, G., Brassard, P., & Bergeron, P. 2001, PASP, 113, 409

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1

Gentile Fusillo, N. P., Rebassa-Mansergas, A., Gänsicke, B. T., et al. 2015, MNRAS, 452, 765
Girven, J., Gänsicke, B. T., Steeghs, D., & Koester, D. 2011, MNRAS, 417, 1210
Gunn, J. E., Carr, M., Rockosi, C., et al. 1998, AJ, 116, 3040
Guo, J., Zhao, J., Tziamtzis, A., et al. 2015, MNRAS, 454, 2787
Holberg, J. B., & Bergeron, P. 2006, AJ, 132, 1221
Huang, Y., Liu, X.-W., Yuan, H.-B., et al. 2015, MNRAS, 449, 162
Kepler, S. O., Kleinman, S. J., Nitta, A., et al. 2007, MNRAS, 375, 1315
Kepler, S. O., Pelisoli, I., Koester, D., et al. 2015, MNRAS, 446, 4078
Kepler, S. O., Pelisoli, I., Koester, D., et al. 2016, MNRAS, 455, 3413
Kleinman, S. J., Kepler, S. O., Koester, D., et al. 2013, ApJS, 204, 5
Koester, D., & Kepler, S. O. 2015, A&A, 583, A86
Kong, X., Luo, A.-L., Li, X.-R., et al. 2018, PASP, 130, 084203
Kowalski, P. M., & Saumon, D. 2006, ApJ, 651, L137
Lee, Y. S., Beers, T. C., Sivarani, T., et al. 2008, AJ, 136, 2022
Liebert, J., Bergeron, P., & Holberg, J. B. 2005, ApJS, 156, 47
Luo, A.-L., Zhao, Y.-H., Zhao, G., et al. 2015, RAA (Research in Astronomy and Astrophysics), 15, 1095
Luri, X., Brown, A. G. A., Sarro, L. M., et al. 2018, A&A, 616, A9
MacQueen, J. 1967, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1: Statistics (Berkeley, Calif.: University of California Press), 281
Ren, J., Luo, A., Li, Y., et al. 2013, AJ, 146, 82
Ren, J. J., Rebassa-Mansergas, A., Luo, A. L., et al. 2014, A&A, 570, A107
Ren, J.-J., Rebassa-Mansergas, A., Parsons, S. G., et al. 2018, MNRAS, 477, 4641
Silvestri, N. M., Lemagie, M. P., Hawley, S. L., et al. 2007, AJ, 134, 741
Tibshirani, R. 1996, Journal of the Royal Statistical Society. Series B (Methodological), 267
Tremblay, P.-E., Bergeron, P., & Gianninas, A. 2011, ApJ, 730, 128
Voss, B., Koester, D., Napiwotzki, R., Christlieb, N., & Reimers, D. 2007, A&A, 470, 1079
Wegg, C., & Phinney, E. S. 2012, MNRAS, 426, 427
Wei, P., Luo, A., Li, Y., et al. 2013, MNRAS, 431, 1800
Woosley, S. E., & Heger, A. 2015, ApJ, 810, 34