RAA 2015 Vol. **15** No. **10**, 1671–1694 doi: 10.1088/1674–4527/15/10/005
http://www.raa-journal.org *http://www.iop.org/journals/raa*

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

Identifying Carbon stars from the LAMOST pilot survey with the efficient manifold ranking algorithm

Jian-Min Si[1,2,3], Yin-Bi Li[1], A-Li Luo[1,3], Liang-Ping Tu[4], Zhi-Xin Shi[1,2,3], Jian-Nan Zhang[1], Peng Wei[1,3], Gang Zhao[1], Yi-Hong Wu[2], Fu-Chao Wu[2] and Yong-Heng Zhao[1]

[1] Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China; *ybli@nao.cas.cn, lal@nao.cas.cn*
[2] National Key Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100095, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[4] School of Science, Liaoning University of Science and Technology, Anshan 144051, China

Received 2015 January 21; accepted 2015 March 27

Abstract Carbon stars are excellent kinematic tracers of galaxies and can serve as a viable standard candle, so it is worthwhile to automatically search for them in a large amount of spectra. In this paper, we apply the efficient manifold ranking algorithm to search for carbon stars from the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) pilot survey, whose performance and robustness are verified comprehensively with four test experiments. Using this algorithm, we find a total of 183 carbon stars, and 158 of them are new findings. According to different spectral features, our carbon stars are classified as 58 C-H stars, 11 C-H star candidates, 56 C-R stars, ten C-R star candidates, 30 C-N stars, three C-N star candidates, and four C-J stars. There are also ten objects which have no spectral type because of low spectral quality, and a composite spectrum consisting of a white dwarf and a carbon star. Applying the support vector machine algorithm, we obtain the linear optimum classification plane in the $J-H$ versus $H-K_s$ color diagram which can be used to distinguish C-H from C-N stars with their $J-H$ and $H-K_s$ colors. In addition, we identify 18 dwarf carbon stars with their relatively high proper motions, and find three carbon stars with FUV detections likely have optical invisible companions by cross matching with data from the Galaxy Evolution Explorer. In the end, we detect four variable carbon stars with the Northern Sky Variability Survey, the Catalina Sky Survey and the LINEAR variability databases. According to their periods and amplitudes derived by fitting light curves with a sinusoidal function, three of them are likely semiregular variable stars and one is likely a Mira variable star.

Key words: methods: data analysis — methods: statistical — stars: carbon — binaries — stars: variables

1 INTRODUCTION

Carbon stars were first recognized by Secchi (1869), and are defined as cases whose optical spectra are dominated by carbon molecular absorption bands, such as CN, CH or Swan bands of C2, and SiC2 and C3 in cooler stars. Based on the spectral features in each band, they are classified as a visual carbon star, an infrared carbon star, an extreme carbon star, a silicate carbon star and so on. In this paper, we just focus on searching for and studying visual carbon stars with the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) spectra. The optical spectra of carbon stars with late spectral type are similar to M-type stars, but their atmosphere contains more carbon than oxygen, because most of the carbon in M-type stars is depleted by CO and more oxygen is left to form molecular bands of TiO which are specific features of M-type stars. However in carbon stars, most of the oxygen is exhausted by CO, and more carbon remains showing C2, CN or CH molecular bands, which are dominant features that are used to distinguish a carbon star from other types of stars.

Carbon stars are peculiar and rare objects compared with normal stellar objects, and they are excellent kinematic tracers of the Galaxy. Dean (1976) analyzed kinematic properties of 425 carbon stars with radial velocities and spectral types, and concluded that the majority of these carbon stars were dynamically similar to dwarf F5 stars. Metzger & Schechter (1994) measured radial velocities of 179 carbon stars toward the Galactic anticenter with accuracies of 4 km s^{-1}, derived their distances using K-band photometry, and estimated an average Galactocentric distance of 13.9 kpc. From the velocities of carbon stars, they found that one carbon star was moving radially outward with respect to the local standard of rest at a velocity of 6.6 km s^{-1}. Recently, the rotation curve of the Milky Way has been studied based on carbon stars with radial velocities in many literatures (Battinelli et al. 2013; Demers & Battinelli 2007; Demers et al. 2009), and they extended the rotation curve of the Milky Way to different distances. In addition, carbon stars can also serve as viable standard candles. Richer et al. (1984) estimated the distance modulus of NGC 205 to be 24.5 using the mean apparent magnitude of carbon stars in NGC 205, and Richer & Crabtree (1985) made use of carbon stars to derive the distance modulus of NGC 300, which is 25.87 magnitude.

So far, many carbon stars have been found from different sky surveys. Totten & Irwin (1998) found 48 cool carbon stars in the Galactic halo from the APM survey using color relations between B_j, R, O and E. Gigoyan et al. (2012) systematically searched for faint high-latitude carbon stars from the low-resolution spectral database of the Digitized First Byurakan Survey (DFBS), and identified 13 new faint high-latitude carbon stars including five C-N stars, five C-H stars and three possible dwarf carbon (dC) stars. Mauron (2008) identified 58 carbon star candidates using JHK_s colors from the Two Micron All Sky Survey (2MASS), and found 18 new carbon stars from them. From the year 2002 to 2013, carbon stars were systematically searched for in the Sloan Digital Sky Survey (SDSS). Margon et al. (2002) and Downes et al. (2004) respectively found 39 and 251 faint high-latitude carbon stars using SDSS photometry, and also Green (2013) identified 1220 high Galactic latitude carbon stars using the Cross-Correlation Function (CCF) method.

We (Si et al. 2014) applied the label propagation algorithm to search for carbon stars in Data Release Eight (DR8) of SDSS, and found 202 new carbon stars. However, the method of label propagation has high time complexity, and needs a large amount of memory. In this paper, we use the efficient manifold ranking (EMR) method proposed by Xu et al. (2011) to search for carbon stars from the LAMOST pilot survey, which is extremely efficient and scalable for a large dataset, and has similar performance as the label propagation algorithm when applied to searching for carbon stars .

The structure of this paper is organized as follows. In Section 2, we present the EMR algorithm in detail, and analyze its time and space complexity. In Section 3, we analyze the algorithmic efficiency and performance using four experiments with SDSS DR8 stellar spectra, then apply the method to the LAMOST pilot survey, and find 183 carbon stars. In Section 4, we classify these carbon stars into four groups with spectral features, and then obtain a linear optimum classification plane in the

$J-H$ versus $H-K_s$ color diagram, and use it to distinguish the C-H and C-N stars. In this section, we also identify 18 possible dC stars, three possible carbon star binaries with FUV detections, and four possible variable carbon stars. Finally, a brief conclusion is provided in Section 5.

2 EMR ALGORITHM

Let a massive dataset be represented as $X = \{x_1, x_2, \ldots, x_q, x_{q+1} \ldots, x_n\}$, where q is the number of query samples, n-q is the number of unlabeled samples and generally $q \ll n$. We aim to efficiently search for objects which have the same class as queries from massive unlabeled samples. One of the efficient solutions is to rank the unlabeled samples by making full use of relationships between unlabeled samples, and samples ranking at the top are the results of the search. Because $q \ll n$ and n is quite large, relationships between massive pairs of unlabeled samples should be considered. Initially, we can assign query samples a high score, and then calculate the score of unlabeled samples subject to the constraint that similar samples have similar scores and the query samples keep high scores. Finally, samples are ranked in descending order by their scores. Let $R = \{r_1, r_2, \ldots, r_n\}$ be the ranking scores that need to be calculated, $Y = \{y_1, y_2, \ldots, y_n\}$ is the initial ranking scores, where if x_i is in the query $y_i = 1$, otherwise $y_i = 0$. The cost function associated with $R_{\rm cost}$ can be represented by the following formula

$$\min_{R_{\rm cost}} \mathcal{Q}(R_{\rm cost}) = 1/2\left(\sum_{i,j=1}^{n} W_{i,j} \parallel \frac{1}{\sqrt{D_{ii}}} r_i - \frac{1}{\sqrt{D_{jj}}} r_j \parallel^2 + \mu \sum_{i=1}^{n} \parallel r_i - y_i \parallel^2 \right), \quad (1)$$

where $\mu > 0$ is the regularization factor and D is a diagonal matrix $D_{ii} = \sum_{j=1}^{n} W_{i,j}$. The first term can ensure that neighbors have similar rankings, and the second term can keep queries near the top. There are generally two solutions for $R_{\rm cost}$ by minimizing the cost function. One is the optimal analytical solution defined as follows

$$R^* = (I_n - \alpha S)^{-1} Y, \quad (2)$$

where $\alpha = \frac{1}{1+\mu}$, I_n is an identity matrix with $n \times n$, and $S = D^{-1/2} W D^{-1/2}$ is normalization of W. However, the analytical solution does not work when n is large because it needs a large amount of memory, and is very time consuming to invert a large matrix with a time complexity of $O(n^3)$. The iterative solution is another approach, which can be obtained by differentiating the cost function. It is efficient and needs a relatively small memory compared with the analytical solution, but it still has a high time complexity of $O(kn^2)$ because of the construction of the K-nearest neighbors (KNN) graph.

$$R^* = \alpha S \times R + (1 - \alpha) Y\,. \quad (3)$$

2.1 Scalable Graph Construction

Most of the datasets with high dimensions have a manifold structure embedded, and the KNN graph is generally used to describe the manifold structure. The KNN graph can be denoted by adjacency matrix W with element W_{ij}, which indicates similarity between samples i and j. An apparent shortcoming of the KNN graph is the large time complexity $O(kn^2)$ and the large amount of memory required. Anchors can be applied to construct a scalable graph to overcome this shortcoming. Let a data set be $X = \{x_1, x_2, \ldots, x_n\}$, and anchors be $U = \{u_1, u_2, \ldots, u_d\}$, where $d \ll n$. According to the principal of mean shift, each data point of X can be estimated by a weighted combination of its KNN, so the weights between the data point and its KNN from anchors can be used to represent the data point, namely $x_i = \sum_{k=1}^{d} z_{ki} \times u_k$ where z_{ki} denotes the normalized weights between data

point x_i and anchor u_k, i.e. $z_k = \{z_{k1}, z_{k2}, \ldots, z_{kd}\}$ is the representation of the data point x_i. The weights can be well measured using the Nadaraya-Watson kernel, so we get

$$z_{ki} = \frac{K\left(\frac{|x_i - u_k|}{\lambda}\right)}{\sum_{l=1}^{d} K\left(\frac{|x_i - u_l|}{\lambda}\right)}, \tag{4}$$

where λ is the smoothing parameter and $K(t) = \begin{cases} 0.75(1-t^2) & \text{if} |t| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$ The parameter λ is important, which determines the number of neighbors used to estimate the data point. In this paper, we adopt the distance between data point x_i and its k_{th} nearest neighbor from anchors as λ, namely $\lambda(x_i) = |x_i - u_{[k]}|$ where $u_{[k]}$ is the k_{th} nearest anchor of x_i. Once we obtain the weighted matrix $Z = \{z_1, z_2, \ldots z_n\}$ where z_i is a d-dimensional vector with element z_{ki}, we can construct the scalable graph $W = Z^T Z$. If the data point x_i and x_j share one or more anchors, they are correlative, namely $W_{ij} > 0$, otherwise not. The sparsity of the graph is determined by the parameter λ, and the larger λ is, the sparser the graph is. In general, λ is adapted to be distance from data point x_i to its k_{th} anchor, and k, ranging from 5 to 15, is proper empirically. So, high sparseness of Z ensures the graph W is sparse. One advantage is that we need not keep the $n \times n$ matrix in memory, and only keep the more sparse $d \times n$ matrix Z in memory, which is very useful for large scale problems. Its computation time is more efficient, which can be seen in Section 3.4.

2.2 Solution and Complexity

Let $H = ZD^{-1/2}$ and $S = H^T H$, then we can obtain the analytical solution as follows.

$$R^* = (I_n - \alpha H^T H)^{-1} Y = \left[I_n - H^T \left(HH^T - \frac{1}{\alpha} I_d\right)^{-1} H\right] Y. \tag{5}$$

The issue of inverting an $n \times n$ matrix is converted to inverting the $d \times d$ matrix, and $D_{ii} = \sum_{j=1}^{n} W_{i,j}$ can be calculated by $Z^T Z e$, where e is an n dimensional vector with elements that have a value of 1.

The time complexity is $O(dn + d^3)$, and the space complexity is $O(kn)$, so for a large scale problem on the order of a million data points, a common personal computer can satisfy the computational requirements.

2.3 Relevance Feedback

For the retrieval problem, relevant results can generally be used as new queries to improve the retrieval results, because although queries are assigned to the same class, there are differences caused by noises or deformations. In large datasets, the improvement is more apparent, and the high efficiency of the EMR algorithm makes multiple relevance feedback possible.

3 EXPERIMENTS AND RESULTS

3.1 Features

Features are key for a machine learning algorithm. A good definition of a feature which describes the object well can improve the algorithm's performance, but a bad one can degrade it, which can be seen in Subsection 3.4. For spectra, there are two important features. One is the continuum, and the other is spectral lines. So here, we adopt spectra and continuum-subtracted spectra as features that are analyzed by our algorithm. For spectra, a median filter is applied with a width of 5 Å, which can delete narrow strong lines, as illustrated in Figure 1. The continuum-subtracted spectra are obtained

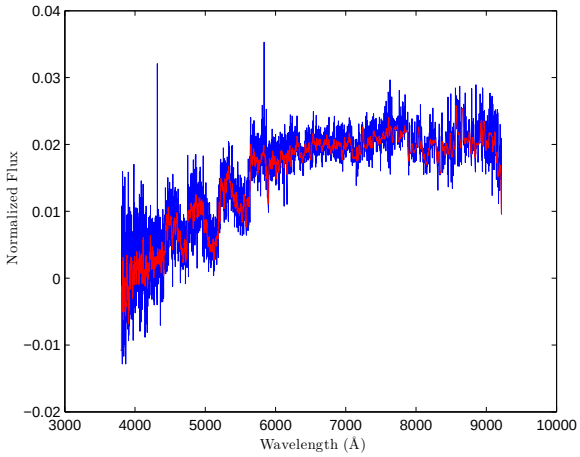


Fig. 1 Median filtered spectra with a width of 5 Å.

using filtered flux to divide the pseudo-continuum derived by the median filter with a width of 300 Å. From Figure 1, we can see that it is efficient to delete strong narrow lines, and we also find that the red and blue ends of the spectra have a large amount of noise, so we delete parts of the spectra with a length of 200 Å from both of these two ends. Based on these two features, we constructed two weighted matrixes Z_1 and Z_2, and the final weighted matrix $Z = Z_1 + Z_2$.

3.2 Anchor Construction

Principal component analysis (PCA) (Jolliffe 2002) is one of the most important and practical methods for dimension reduction and feature extraction, and k-means clustering (Wu et al. 2008) is one of the most useful data mining algorithms. These are widely applied in many fields including astronomy. Here, we use PCA and k-means clustering to construct anchors. Initially, PCA is applied to the whole dataset, and 100 principal components are obtained. Then, all of the data points are mapped on these principal components, and the coefficients are returned. In the end, k-means clustering is applied to these coefficients, and the centers of clusters are adopted as our anchors. However, the final result of k-means is greatly affected by initial seeds, so we perform k-means clustering twice using different random initializations, and both of the obtained centers are implemented as our final anchors. We then divide the whole dataset into two equal numbers of samples, and apply k-means clustering to both of them.

3.3 Efficiency and Performance of the Algorithm

We use 656 801 stellar spectra from SDSS DR8 to test the performance of the algorithm in searching for carbon stars. Green (2013) systematically searched for carbon stars using a CCF and classified results of the SDSS pipeline, then built the largest carbon star catalog identified spectroscopically. By a cross matching with positions of this carbon star catalog, we obtain 1313 carbon star spectra which are classified as 'STAR,' so we can test the algorithm supposing that there are only 1313 carbon star spectra in the 656 801 stellar spectra. Recall and precision are two important methods to evaluate an algorithm in terms of information retrieval. We choose 20 carbon spectra as queries with different signal-to-noise ratios (SNRs) that are listed in Table 1.

There is no doubt that the number of anchors is important for this algorithm, so we calculate the recalls of the top 2000 using different numbers of anchors, which are plotted in Figure 2. It is

Table 1 Queries Used to Verify the Performance of the Algorithm

Query ID	Plate	MJD	Fiber ID	SNR
1	453	51915	53	9.41
2	613	52345	344	6.08
3	696	52209	133	19.45
4	1067	52616	602	5.73
5	1274	52995	359	5.09
6	1307	52999	116	24.09
7	1311	52765	571	10.44
8	1326	52764	502	5.39
9	1465	53082	570	22.55
10	1486	52993	368	4.69
11	1521	52945	596	8.05
12	1687	53260	83	6.85
13	1881	53261	165	7.71
14	2083	53359	93	4.81
15	2183	53536	447	36.82
16	2619	54506	279	29.08
17	2795	54563	618	6.05
18	2866	54478	351	71.97
19	3232	54882	307	36.75
20	726	52207	239	5.76

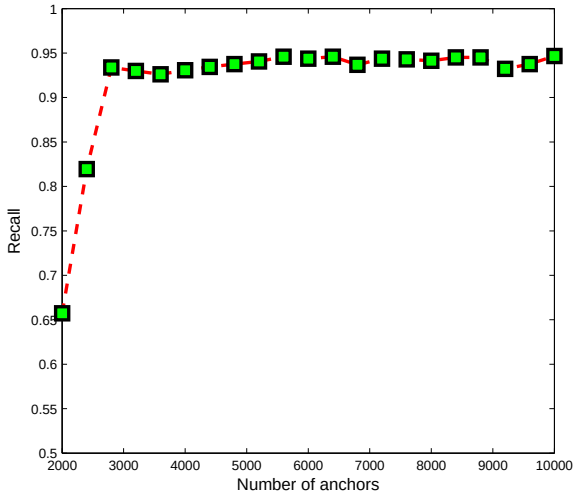


Fig. 2 Recalls of different numbers of anchors.

shown that recall significantly increases when the number of anchors changes from 2000 to 3000. The rate of increase becomes slow when the number of anchors changes from 3000 to 6000, but there is almost no change when the number of anchors changes from 6000 to 10 000. Considering that the running time increases sharply with an increase in the number of anchors, we finally choose 6000 anchors.

In order to explain the importance of the features, we calculate the recalls of the top 2000 and the top 3000 using 20 queries with different features, which are plotted in Figure 3. In Figure 3, the blue bars indicate the results of using filtered flux, the green bars indicate the results of continuum-subtracted flux, and the red bars indicate the results of filtered flux combined with continuum-

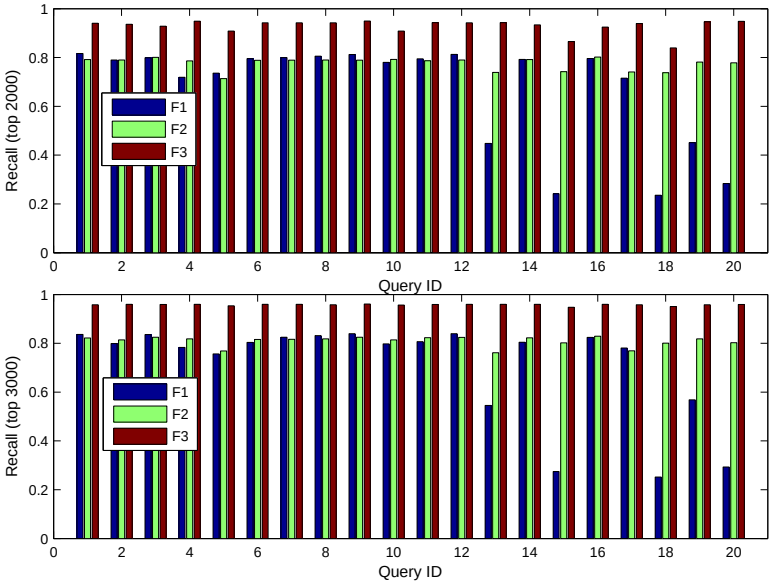


Fig. 3 Recalls using different queries based on different features. The upper panel shows recalls of the top 2000 and the bottom panel shows recalls of the top 3000.

subtracted flux. From Figure 3, we can see that the results of using continuum-subtracted flux are stable compared with the results of using filtered flux. The results would be quite bad if we choose queries poorly when using the feature of filtered flux, but the combined feature is quite stable and can greatly improve the performance. Using the combined feature, recalls of the top 2000 for all queries with the exception of query 18 are more than 90%, many of which are larger than 94%, and all the recalls of the top 3000 are more than 95%, some of which are up to 98%.

The parameters used are also important for an algorithm, and sometimes dominate the performance of the algorithm, causing difficulties in tuning the algorithm. Here, we check the performance of the algorithm with different parameters illustrated in Figure 4, which are curves of recall versus precision for the top 1 to 5000 with different parameters. We can see that the algorithm is quite robust, and they all have high recalls and precisions with different parameters, even if $K = 3$. It can also be seen that when $K = 15$ or $K = 3$ and α = 0.99, recalls are a little poorer than those with other parameters. In our paper, we choose α = 0.85 and $K = 6$ to search for carbon stars from the LAMOST pilot survey.

In order to display the role of relevance feedback, we first choose one query, and recalculate relevance feedback four times by increasing the most relevant samples to 5, 10, 20 and 50 as new queries step by step according to the searching results. Their recalls and precisions for the top 1 to 5000 are plotted in Figure 5. It is clearly seen that relevant feedback can improve the results from one query to five queries.

In order to test whether the algorithm is fast and scalable for a large dataset, we calculate running time for each process of constructing 6000 anchors by applying PCA and k-means clustering using a 12-core Intel(R) Xeon(R) 3.47 GHz machine with 96 GB of RAM. The results are listed in Table 2. The algorithm is encoded and executed in MATLAB, and we can see that the main running time is spent on constructing anchors and Z, and the total time is less than half an hour. The time to calculate $R_{\rm cost}$ is only 19 seconds, and this allows us to calculate relevance feedback many times. Compared with the label propagation algorithm we used to search for carbon stars and DZ white dwarfs (WDs)

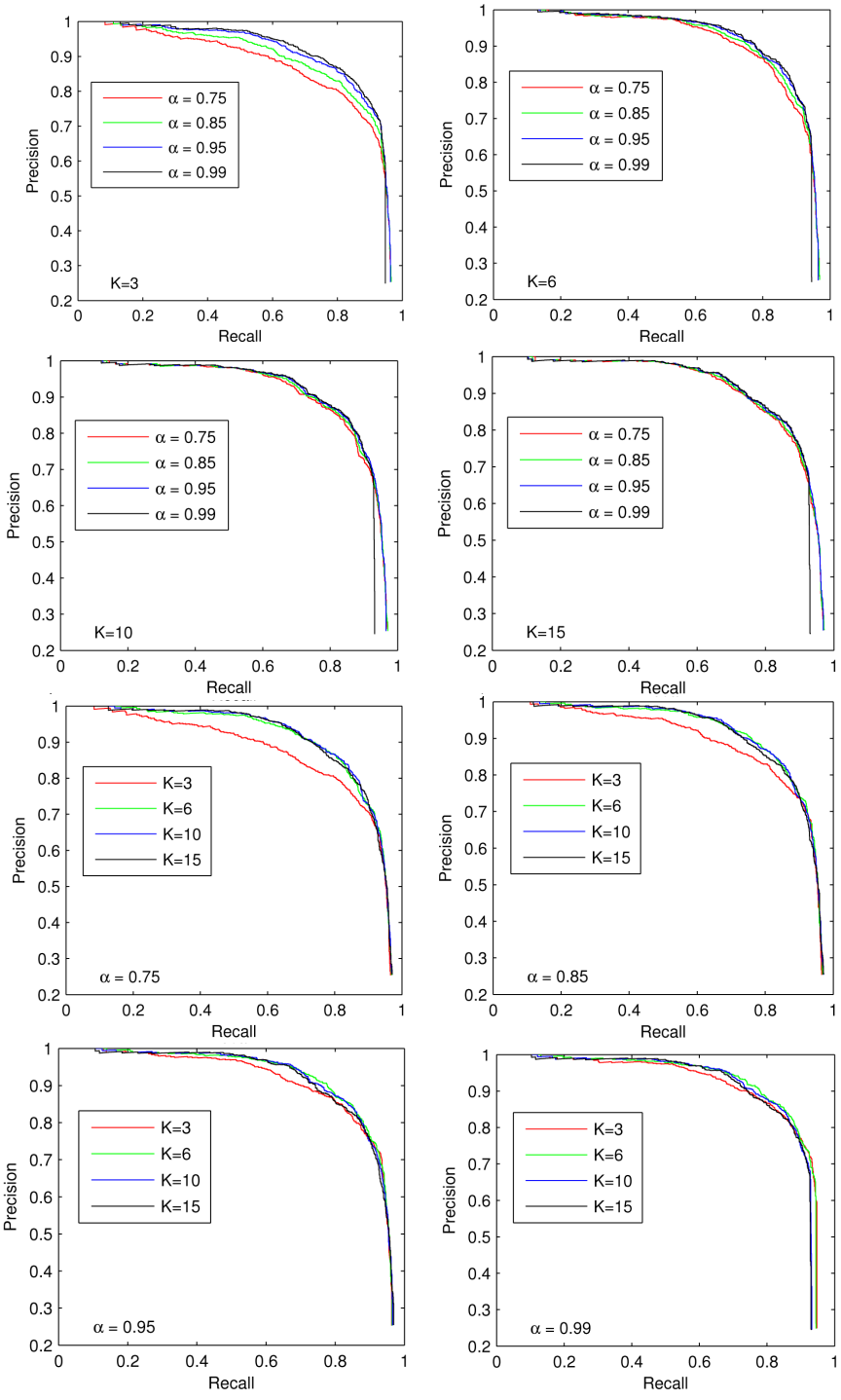


Fig. 4 Recalls and precisions of the top 1 to 5000 cases corresponding to different values for parameters α and K.

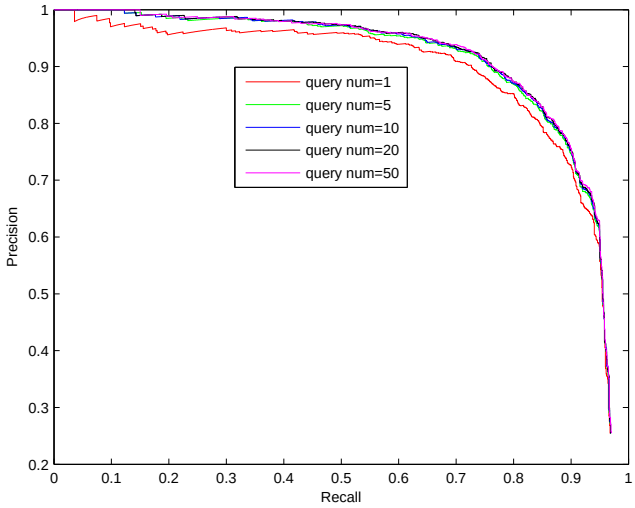


Fig. 5 Relevance feedback performance when increasing queries.

Table 2 Running Time

Process	Time (s)
Anchor construction[a]	1306
Z calculation	334
$R_{\rm cost}$ calculation	19
Total	1659

Notes: Constructing 6000 anchors using PCA and k-means clustering.

from SDSS DR8 (Si et al. 2014), the time efficiency of this algorithm is significantly improved under the condition that the performance has no significant decrease, as illustrated in Figure 6. It is 55.3 times as fast as the label propagation algorithm in searching for carbon stars from a dataset with more than 650 000 spectra in the same running environment, and this advantage is more significant when the dataset is larger.

In summary, the algorithm is efficient and scalable for a large dataset, and can be practically applied in searching for carbon stars from a large spectral dataset.

3.4 Results of Searching for Carbon Stars from the LAMOST Pilot Survey

LAMOST has the potential to efficiently survey a large volume of space for stars and galaxies, and can acquire 4000 spectra as faint as $r = 19$ magnitude in a single exposure at resolution $R = 1800$ (Zhao et al. 2012). The performance of LAMOST was greatly improved after the year 2009. The accuracy of the optical fiber positioning has been better than 1 arcsecond, and after being finely tuned, the stability and the overall efficiency of the associated spectrographs have also been improved[1]. In order to check the performance and feasibility of the science goals, LAMOST has successfully finished the pilot survey which began on 2011 October 24, and ended in June 2012 (Luo et al. 2012). During the pilot survey, 640 000 spectra were observed, and 319 000 spectra were finally released (Luo et al. 2012). We aim to search for carbon stars from all the spectra observed in the pilot survey.

[1] *http://www.lamost.org/public/news/6?locale=en*

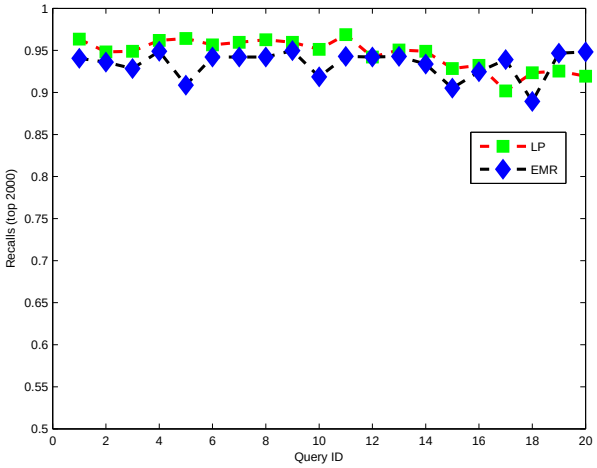


Fig. 6 Recalls of the label propagation algorithm and the EMR algorithm. LP indicates the label propagation algorithm and EMR indicates the efficient manifold ranking algorithm.

Table 3 The New Carbon Star Catalog

Designation	u (mag)	g (mag)	r (mag)	i (mag)	z (mag)	J (mag)	H (mag)	K_s (mag)	SpT[a]	d[b]	New[c]
J065609.26+104247.8	16.94	14.79	13.15	12.83	13.09	11.47	10.92	10.82	CH	d	Y
J015629.10+042344.1	16.88	15.34	14.59	14.34	14.25	13.30	12.90	12.86	CH	d	Y
J093547.66+270939.3	19.29	16.73	15.62	15.29	15.14	14.04	13.53	13.39	CH	d	N
J101754.72+251201.1	16.75	15.32	14.64	14.43	14.34	13.4	13.02	12.92	CH	d	Y
J134816.26-004921.5	16.27	14.97	13.11	12.82	13.46	11.54	10.99	10.89	CH	d	Y
J064858.00+285421.3	20.99	17.81	15.96	15.29	14.84	13.34	12.53	12.28	CR	d	Y
J101946.89+252932.8	16.25	15.18	13.48	14.57	13.51	11.93	11.43	11.35	CR	d	Y
J093608.72+121634.1	21.75	19.4	18.25	17.86	17.65	16.48	16.05	15.05	CR	d	Y
J123204.75+270952.2	19.74	17.19	16.01	15.58	15.45	14.45	13.91	13.75	CR?	d	Y
J081157.14+143533.0	15.71	15.95	15.74	15.56	15.37	14.22	13.53	13.33	BINARY	u	N
J000215.77+311117.9	17.58	16.03	14.66	14.93	14.39	13.03	12.50	12.44	CH	u	Y
J065136.94+131350.0	17.43	15.57	12.83	14.42	12.52	10.64	9.89	9.71	CR	u	Y
J065314.97+110720.4	16.28	13.64	14.45	11.98	12.47	10.38	9.70	9.58	CH?	u	Y
J003858.71+394504.2	18.04	16.43	17.04	14.16	14.87	12.82	12.29	12.22	CH?	u	Y
J003511.27+402231.4	NA[d]	NA	NA	NA	NA	13.65	13.13	13.01	CH	u	Y
J055520.99+260812.2	NA	NA	NA	NA	NA	9.24	7.82	7.02	CN?	u	N
J010629.35+381440.0	NA	NA	NA	NA	NA	12.22	11.54	11.34	CH	u	Y
...	...	...	...	...	...	...	...	...	...	...	...

Notes: [a] The marker '?' indicates an uncertain observation, and 'CH?', 'CN?' and 'CR?' indicate possible C-H, C-N and C-R stars respectively. [b] 'd' indicates dC stars, and 'u' indicates an uncertain observation. [c] 'Y' indicates a new finding and 'N' indicates the star has been published. [d] 'NA' indicates the parameter is not available. Notes: The entire table can be found on *http://www.raa-journal.org/docs/Supp/ms2148table3.csv*. A portion is shown here for guidance regarding its form and content.

We firstly use 20 SDSS carbon stars as initial queries and 138 carbon star spectra are obtained by visually inspecting the top 1000 results. Then, we select 50 carbon spectra with higher quality as new queries. By manually checking the top 5000 results, we identify 183 carbon stars listed in Table 3, and 158 of them are new findings that were cross identified with the SIMBAD, NED and ADS databases and are marked 'Y' in the last column of Table 3. We plot their locations in Galactic coordinates and in equatorial coordinates in Figure 7, and it is shown that the distribution

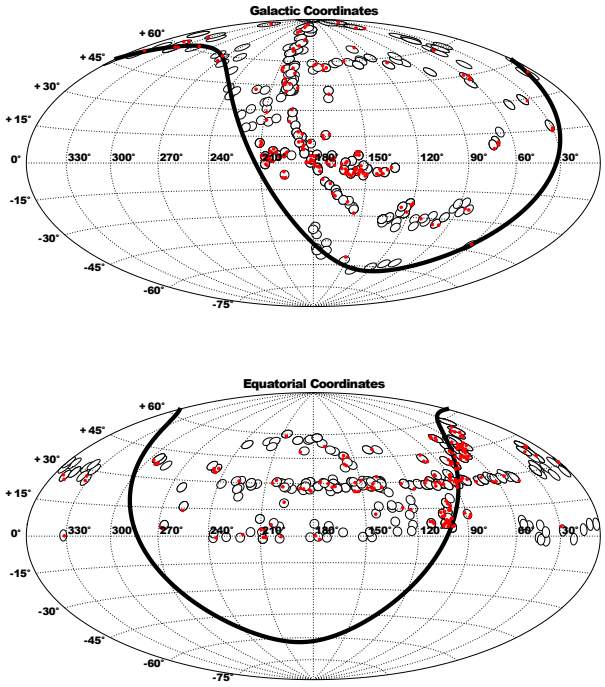


Fig. 7 Positions of our carbon stars and observed regions by the LAMOST pilot survey in Galactic coordinates and equatorial coordinates. The red points are positions of our carbon stars, the circles are positions of observed plates, the black bold curve in the upper panel is the equatorial plane, and the black bold curve in the lower panel is the Galactic plane.

of our carbon stars only depends on the spatial distribution of the observations and shows no regular pattern.

4 DISCUSSION

4.1 Spectral Classification

With the discovery of more carbon stars, we know that carbon stars span a wide variety of different populations and origins, which are indicated by their different spectral characteristics. In the past decades, many studies have proposed various classification systems for carbon stars (Cannon & Pickering 1918; Keenan & Morgan 1941; Yamashita 1972, 1975), but they did not include all types of carbon stars. Keenan (1993) revised the MK carbon star classification system, and carbon stars were divided into five types in this system. This classification system is widely accepted, and used in many studies (Wallerstein & Knapp 1998; Barnbaum et al. 1996; Goswami et al. 2010; Goswami 2005; Lloyd Evans 2010). We classify our carbon stars using a revised classification system, and the classification criteria are summarized as follows:

(1) There is a strong G band from the CH molecule at λ4300 Å with a secondary P branch head at λ4342 Å. The P branch forms the most prominent features that can be used to distinguish C-H stars from C-R stars.

(2) A strong Ca I line at λ4226 Å compared with the CN band at λ4215 Å is a useful indicator for C-R stars.
(3) Lines of atomic hydrogen and s-process element Ba II at λ4554 Å, λ4935 Å and λ6496 Å are distinctly seen in C-H stars, but not in C-R stars.
(4) There is a strong suppression of light below λ 5000 Å, and little or no flux shortward of λ 4400 Å, which are the key features that distinguish C-N stars from C-H and C-R stars.
(5) There is a larger enhancement of s-process elements for C-N stars than for C-R stars.
(6) A high isotope ratio of ^{13}C to ^{12}C is the main characteristic of C-J stars, which can be measured by the ratio of strengths in bands of $^{13}C^{12}C$ at λ6168 Å and $^{12}C^{12}C$ at λ6192 Å, and the equivalent width ratio of bands associated with $^{13}C^{14}N$ at λ6260 Å and $^{12}C^{14}N$ at λ6206 Å.
(7) The C-HD stars are characterized by weak or an absence of hydrogen lines and the G-band of CH, and also characterized by stronger bands of CN and C_2 than normal carbon stars.

De Mello et al. (2009) concluded that carbon stars with C_j-index $\geq$ 4 are surely C-J stars, thus we firstly identify C-J stars by calculating the C_j-index for our carbon stars using a method described by Margon et al. (2002) and De Mello et al. (2009), which quantizes the isotope ratio of ^{13}C and ^{12}C. Margon et al. (2002) and De Mello et al. (2009) defined the C_j-index based on two well correlated parameters as described in criterion (6) shown above. One is the ratio of strengths of isotopic bands associated with $^{13}C^{12}C\lambda$6168 Å and $^{12}C^{12}C\lambda$6192 Å, and the other is the equivalent width of isotopic bands associated with $^{12}C^{14}N\lambda$6206 Å and $^{13}C^{14}N\lambda$6260 Å. In our carbon stars, there are four C-J stars with the C_j-index $\geq$ 4, and their parameters are listed in Table 4. Figure 8 shows an example of the local continuum and normalized flux of the two isotopic bands.

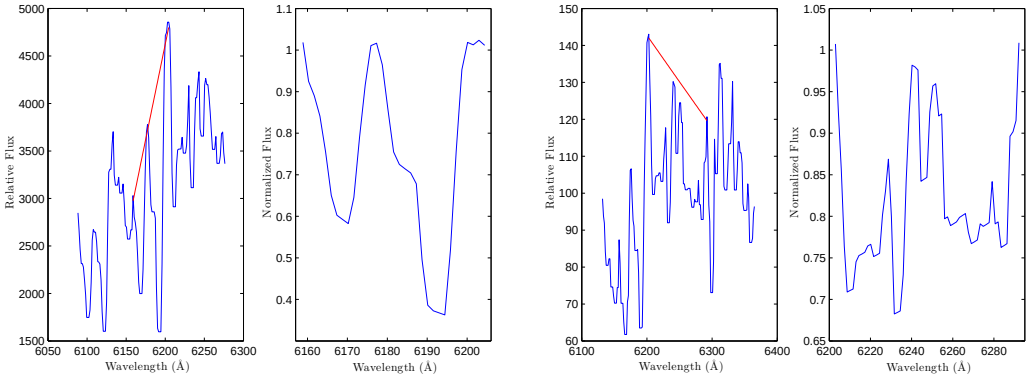
For the remaining stars, 58 stars are identified as C-H stars and 11 stars as C-H star candidates with criteria (1), (2) and (3), 56 stars are identified as C-R stars and ten stars as C-R star candidates with criteria (1), (2), (3), (4) and (5), and 30 stars are identified as C-N stars and three as C-N star candidates with criteria (4) and (6). In addition, there are ten objects which were not assigned spectral types because of the low quality of their spectra, and we are unable to find a C-HD star with criterion (7). Figure 9 shows spectra of four kinds of carbon stars in our samples, which include a C-J star LAMOST J220514.58+000845.5, a C-N star LAMOST J052611.18+382237.6, a C-H star LAMOST J091451.96+332901.6 and a C-R star LAMOST J065136.94+131350.0. Their characteristic spectral lines are indicated in Figure 10, and the local spectra near these spectral lines are also plotted.

In addition, we find a composite spectrum J081157.14+143533.0 consisting of a WD, that is identified as type PG 1159, and a carbon star plotted in Figure 11. J081157.14+143533.0 displays strong CIV absorption lines at 4660 Å and He II at 4686 Å which are significant features for typical PG 1159 WDs, and shows strong CN absorption bands in the red end which are dominant features of carbon stars. Deciding whether it is a physical binary needs more observations in future epochs. We have also plotted the SDSS spectrum of this star in Figure 11, and the composite SDSS spectrum was decomposed into a WD with type of PG 1159 and a carbon star which was described in Si et al. (2014).

Table 4 C_j-index for C-J Stars

Designation	p1[a]	p2[b]	C_j-index
J220514.58+000845.5	0.5375	0.7536	5
J063328.79+274522.5	0.6248	0.6696	5
J052826.57+352943.9	0.7431	1.1299	6
J072530.61+074603.2	0.5477	0.4939	4

Notes: [a] 'p1': the ratio of strengths of isotopic bands associated with $^{13}C^{12}C\lambda$6168 Å and $^{12}C^{12}C\lambda$6192 Å. [b] 'p2': the equivalent width of isotopic bands associated with $^{12}C^{14}N\lambda$6206 Å and $^{13}C^{14}N$ λ6260 Å.



(a) (b)

Fig. 8 The left figures of panels (a) and (b) are local spectra of isotopic bands of $^{13}C^{12}C\lambda$6168 Å and $^{12}C^{12}C\lambda$ 6192 Å, and $^{12}C^{14}N\lambda$6206 Å and $^{13}C^{14}N\lambda$6260 Å respectively, and their local continua are shown by two red lines (*color online*). The right plots of the two panels are their normalized spectra.

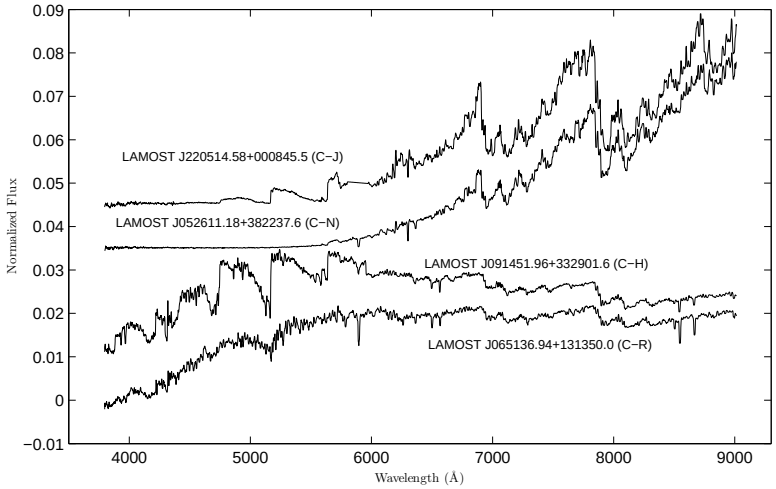


Fig. 9 Spectra of four LAMOST carbon stars, which include a C-J star, a C-N star, a C-H star and a C-R star.

4.2 Identifications of dC Stars

Carbon stars are defined as those with molecular absorption bands of C_2, CN or CH in their optical spectra, and it has long been assumed that carbon stars are always giants as carbon is thought to reach the photosphere only during dredge-up in asymptotic giant branch stars (Green 2013). In 1977, the first dC star, G77−61, was discovered based on its high proper motion, and strong C_2 and CH bands (Dahn et al. 1977). Until 2003, 30 dC stars had been subsequently found based on their relatively high proper motions, and near-infrared wide band colors, which were summarized by Lowrance et al. (2003). Different from classical giant carbon stars, the origins of those dC stars are most likely

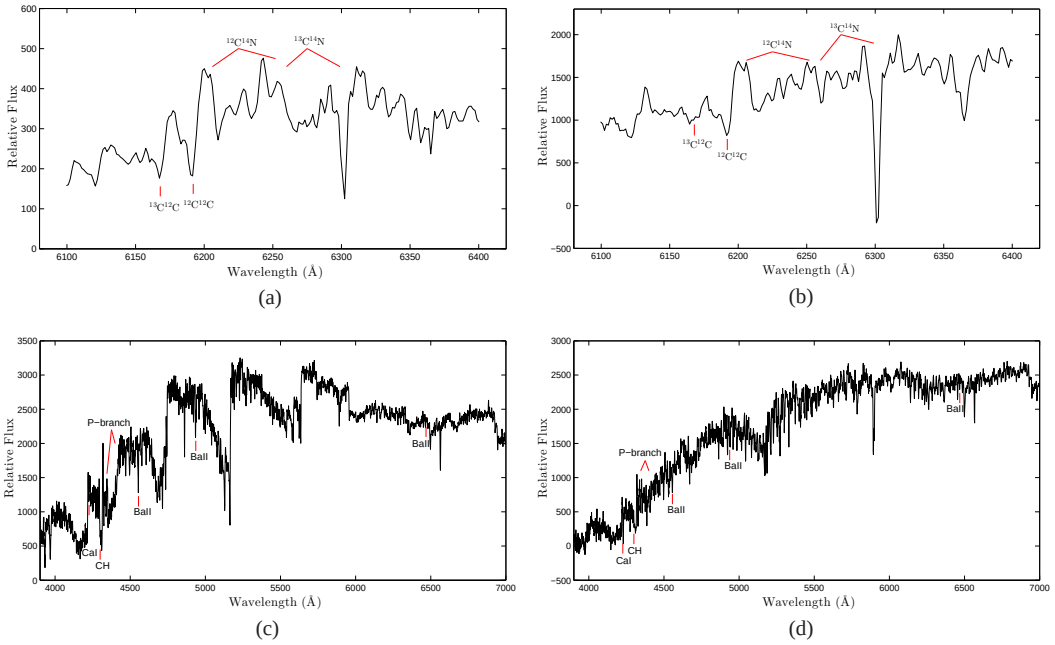


Fig. 10 Prominent spectral feature associated with each of the four carbon stars in Fig. 9. (a) LAMOST J220514.58+000845.5 (C-J), (b) LAMOST J052611.18+382237.6 (C-N), (c) LAMOST J091451.96+332901.6 (C-H), (d) LAMOST J065136.94+131350.0 (C-R).

explained by close binary systems, where the dC star has received material from a now 'invisible' companion (probably a WD with $T_{\rm eff}$ below 5000 K) when the companion was ascending up to the asymptotic giant branch as a carbon giant (Totten et al. 2000; Heber et al. 1993). From 2002 to 2013, Margon et al. (2002), Downes et al. (2004) and Green (2013) searched for faint high-latitude carbon stars from SDSS, and reported 39, 251 and 1220 carbon stars respectively. Among those stars, there are 17, 110 and 729 dC stars respectively, which approximately account for 43.6%, 50% and 69.4% of their total samples of carbon stars. Such a significant fraction of nearby dC stars in faint high-latitude carbon stars demonstrates that they are numerically dominant in the Galaxy, which is different from what was previously assumed. Therefore, it is imperative to investigate dC stars from our carbon samples, which can enlarge the amount of known dC stars.

Until recently, the only way to distinguish between a dC and a carbon giant was luminosity, and hence one needed the parallax or a distance indicator, such as proper motion (Lowrance et al. 2003), to identify dC stars. Except for the two dC stars CBS 311 (Liebert et al. 1994) and PG 0824+289 (Heber et al. 1993), which were identified by their spectra, all known dC stars have been detected by their relatively high proper motions. Of course, other luminosity discriminators based on spectroscopy or photometry have also been proposed, but none of them are currently a reliable criterion and further ancillary clues are needed (Lowrance et al. 2003). Green et al. (1992) suggested that the near-infrared JHK colors and the appearance of an unusually strong C_2 band head at λ6169 Å might be a luminosity indicator to distinguish dC stars. They proposed that dC stars were defined as those with $J-H<0.75$ and $H-K>0.25$. However, Margon et al. (2002) pointed out that JHK photometry is not yet a reliable luminosity indicator and more carbon samples are needed for further confirmation this criterion. They also suggested that the λ6169 band feature may be

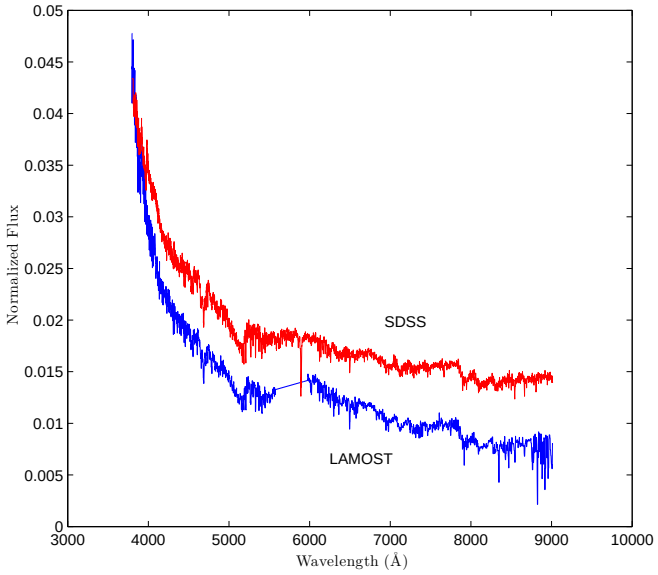


Fig. 11 Spectra of J081157.14+143533.0. The upper one was observed by SDSS and the bottom one was observed by LAMOST.

temperature as well as luminosity sensitive, so it is also not yet a reliable luminosity discriminator. Lowrance et al. (2003) pointed out that some of the dCs have $J-H$ colors like those of giants, therefore they cannot be distinguished by the $J-H$ versus $H-K_s$ photometry relationship, and they suggested that the $R-J$ versus $J-K_s$ diagram may be used to select possible dC stars.

In this paper, we plan to use the high 'proper motion' criterion to distinguish dC stars from our 183 carbon samples. Green (2013) identified 729 faint high-latitude dC stars from SDSS DR8 with significant proper motions, and they define a high proper motion as follows: (1) at least one USNO-B detection and one SDSS detection per source ($n_{\rm fit} > 2$, which is stored in the proper motion catalog in the SDSS schema); (2) proper motion in at least one coordinate is larger than 3σ, where σ is the proper motion uncertainty in that coordinate; (3) total proper motion is larger than 11 mas yr^{-1}. In order to use the above criteria, we first cross match our carbon stars with the proper motion catalog of SDSS DR10 with a radius of two arcseconds, and obtain proper motions of 80 carbon stars. Then, we obtain 71 faint high-latitude carbon stars using an r band magnitude larger than 13 and $|b| > 30^{\circ}$. Finally, we apply the previous three criteria to the 71 samples, and 18 of them are identified as dC stars, plotted in Figure 12, of which position and proper motion are listed in Table 5.

4.3 Locations in the Diagram of $J-H$ versus $H-K_s$ Color

The locations of carbon stars in the two color diagram of $J-H$ and $H-K$ can be used to distinguish C-H stars from C-N stars, and can also be used to determine the fraction of dC star candidates. Totten et al. (2000) concluded that C-N, C-H and dC stars could be well separated by the two-color diagram of $J-H$ versus $H-K$ from the SAAO photometric system, and described their three corresponding regions. In addition, Gigoyan et al. (2012) made use of this infrared photometry method as a supplementary diagnostic tool for classifying carbon stars.

Here, we try to compile more carbon stars with known types to analyze the locations of different types of carbon stars in the diagram of $J-H$ versus $H-K_s$ color from the 2MASS photometric sys-

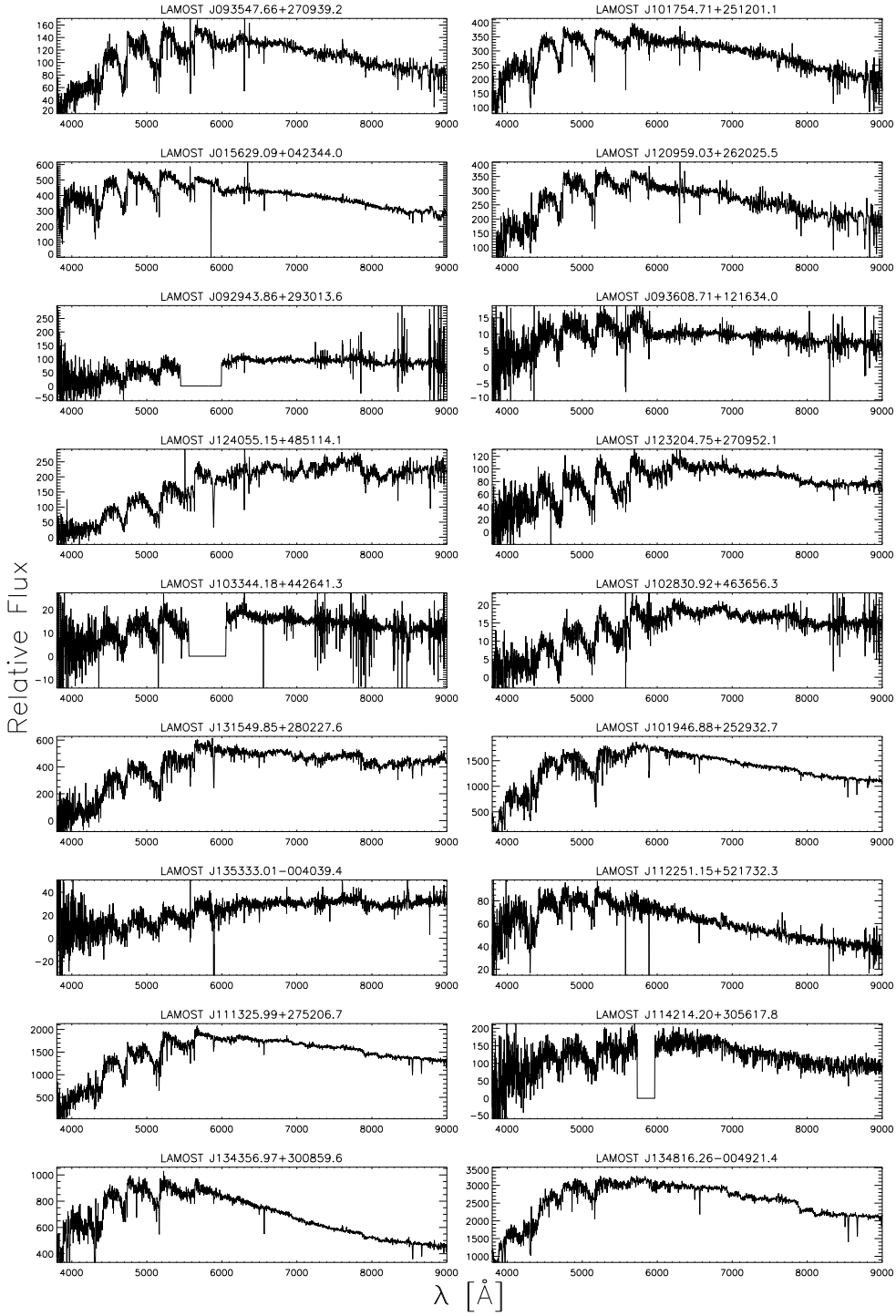


Fig. 12 Spectra of 18 dC stars.

Table 5 The Catalog of dC Stars

Designation	RA	Dec	$\mu_\alpha \cos(\delta)$ (mas yr^{-1})	μ_δ (mas yr^{-1})	$\mu_\alpha \cos(\delta)$_err (mas yr^{-1})	μ_δ_err (mas yr^{-1})	μ (mas yr^{-1})
J015629.10+042344.1	29.12124	4.3956	25.48	–16.94	2.3	2.3	30.59
J092943.86+293013.6	142.4328	29.5038	–50.19	–83.62	2.63	2.63	97.53
J093547.66+270939.3	143.9486	27.1608	17.08	–93.48	2.49	2.49	95.03
	143.9486	27.1608	18.15	–93.78	2.48	2.48	95.52
J093608.72+121634.1	144.0363	12.2761	3.85	–15.98	3.21	3.21	16.44
	144.0363	12.2762	2.08	–16.37	3.39	3.39	16.5
	144.0363	12.2761	2.54	–16.54	3.32	3.32	16.73
J101754.72+251201.1	154.478	25.2003	–3.88	–13.17	2.52	2.52	13.73
	154.478	25.2003	–3.22	–14.35	2.52	2.52	14.71
	154.4779	25.2003	–5.1	–14.22	2.52	2.52	15.11
J101946.89+252932.8	154.9453	25.4924	–43.49	–58.38	2.54	2.54	72.8
	154.9453	25.4924	–42.41	–60.41	2.54	2.54	73.81
J102830.93+463656.4	157.1289	46.6157	–25.3	–35	3.5	3.5	43.19
	157.1289	46.6156	–26.1	–38.86	3.55	3.55	46.81
J103344.19+442641.3	158.4341	44.4448	4.79	–29.58	3.44	3.44	29.96
J111326.00+275206.8	168.3584	27.8681	43.64	–217.13	2.34	2.34	221.47
J112251.15+521732.3	170.7131	52.2922	–11.4	–7.6	2.52	2.52	13.7
J114214.21+305617.9	175.5592	30.9384	–25.24	–5.14	2.63	2.63	25.76
J120959.03+262025.6	182.4959	26.3404	–9.24	–7.61	2.58	2.58	11.97
J123204.75+270952.2	188.0198	27.1643	–28.56	–146.41	2.38	2.38	149.17
J124055.15+485114.2	190.2298	48.8539	11.1	1.88	2.76	2.76	11.25
J131549.85+280227.7	198.9577	28.041	6.4	9.59	2.59	2.59	11.53
	198.9577	28.041	6.59	11.55	2.55	2.55	13.3
J134356.97+300859.6	205.9873	30.1499	–12.73	0.62	2.36	2.36	12.75
J134816.26–004921.5	207.0677	–0.8227	–12.88	–3.63	2.33	2.33	13.38
J135333.02–004039.4	208.3875	–0.6776	–52.57	–29.08	2.91	2.91	60.07
	208.3876	–0.6776	–52.13	–30.59	2.98	2.98	60.44

tem. In total we obtain 190 carbon stars with known spectral type in literature (Totten & Irwin 1998; Barnbaum et al. 1996), and 137 of them have JHK_s magnitudes from the 2MASS photometric system. The Galactic dust reddening and extinction corrections were calculated by the method presented by Schlafly & Finkbeiner (2011). They predicted magnitudes in five bands for each MARCS synthetic spectrum with the method of Gunn et al. (1998), and constructed a synthetic grid of magnitudes. For each star, magnitude in a single SDSS band can be predicted by linearly interpolating this synthetic magnitude grid, and the difference between the predicted colors and the measured colors from the SDSS imaging can give the reddening estimates. Considering the impact of extinction, we remove five of them with K_s band extinction larger than 0.5. The two-color diagram of $J-H$ versus $H-K_s$ for the 132 carbon stars is shown in Figure 13. The left panel is a two-color diagram plotted in the SAAO photometry system by transforming the J and K_s magnitude of the 2MASS system to the SAAO system using the transformation formula derived by Koen et al. (2007), and the superimposed dotted boundaries are the locations of different carbon types, which were defined in Figure 3 of Totten et al. (2000). The right panel is a two-color diagram in the 2MASS system, and the superimposed dotted boundaries were plotted by applying an affine transformation to the locations in the SAAO system using the transformation formula derived by Koen et al. (2007). From Figure 13, we can see that there are a few C-N stars located outside of their superimposed dashed boundary regions given by Totten et al. (2000) in the SAAO system, and in the 2MASS system, there are more C-N stars outside of their corresponding region, which indicates that C-N stars may cover larger regions. We can also see that the C-J stars are located in the same region as the C-N stars, and the C-R stars are in the region where C-N and the C-H stars are also positioned, which can also be seen in figure 5 of Downes et al. (2004). In addition, dC regions only contain a few dC stars, and the

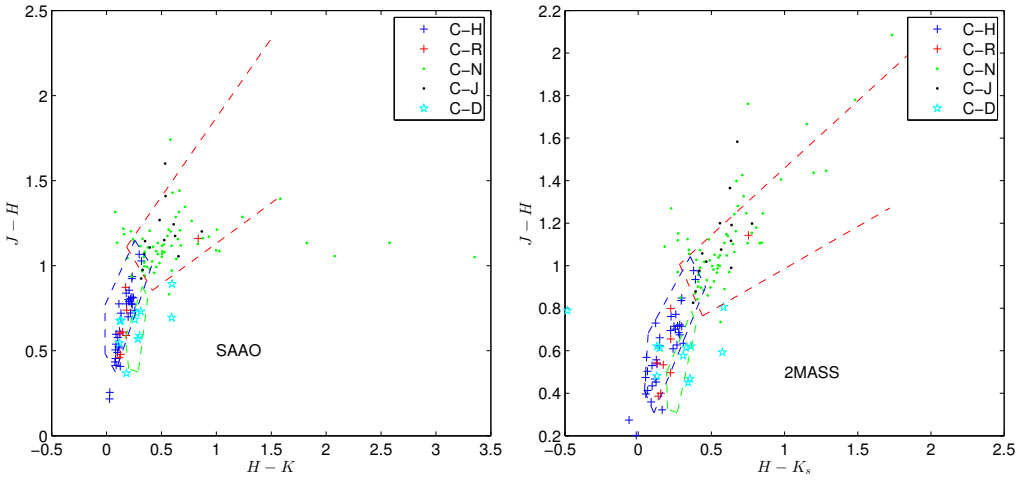


Fig. 13 Distributions for different types of carbon stars in the diagram of $J-H$ and $H-K$ colors in the SAAO system and $J-H$ and $H-K_s$ colors in the 2MASS system.

large majority of them are located outside of this region. In the end, we can also conclude that the JHK_s color of 2MASS can be used to separate C-N stars from C-H stars, which might be explained in that JHK_s colors are a good indicator of effective temperature (Wang & Jiang 2014), and C-N stars are cooler than C-H stars.

Because the C-N, C-H and dC stars could be separated well by the two-color diagram of $J-H$ versus $H-K_s$, we adopt a support vector machine (SVM) method, which is one of the most practical and widely applied methods in this field, which was proposed by Cortes & Vapnik (1995). When applying SVM to this study, the goal is to obtain an optimum linear classification plane that can distinguish C-N from C-H stars. In the $J-H$ versus $H-K_s$ color diagram, we find that the locations of five C-N stars with K_s band extinction larger than 0.5 are far from other points in the graph, and they might be selected as support vectors which could seriously affect the fitting of the classification plane. To overcome this problem, we primarily exclude them from 137 stars.

Figure 14 shows the selected support vectors indicated by black open circles, and the linear classification plane $(J-H) = -0.6851 \times (H-K_s) + 1.0974$ marked by the black solid line. The diagram of $J-H$ versus $H-K_s$ color of our 187 carbon stars with the JHK_s photometry from the 2MASS is plotted in Figure 15. It is clearly seen that the distributions are consistent with those in Figure 13, and the optimum linear classification plane obtained by the SVM algorithm can distinguish the C-N stars from the C-H stars well, which also indicates that our classification results in Subsection 4.1 are reliable. In addition, there are two identified dwarfs located in the region where dC stars are located given by Totten et al. (2000), which is indicated with green dashed lines, and another 19 dC stars are located outside of this region.

4.4 GALEX Detected Stars

The Galaxy Evolution Explorer (GALEX)[2] (Martin et al. 2005), a NASA small explorer mission launched on April 2003, performed the first UV sky imaging and spectroscopic survey in two bands (NUV: 1344 – 1786 Å and FUV: 1771–2831 Å). The primary goal of the GALEX survey was to

[2] *www.galex.caltech.edu/index.html*

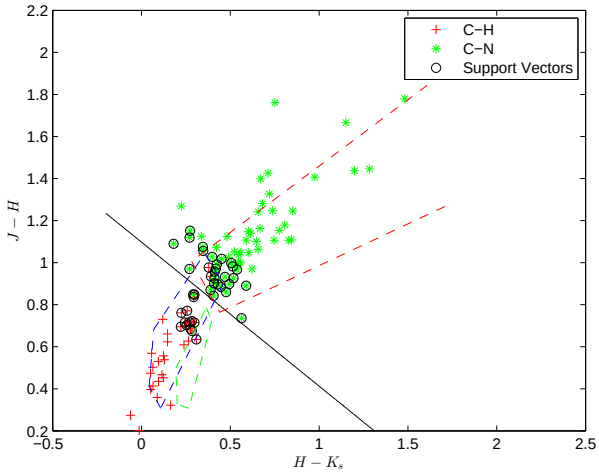


Fig. 14 The optimum classification plane for C-N and C-H stars in the near-infrared color diagram of $J-H$ versus $H-K_s$ using the SVM algorithm.

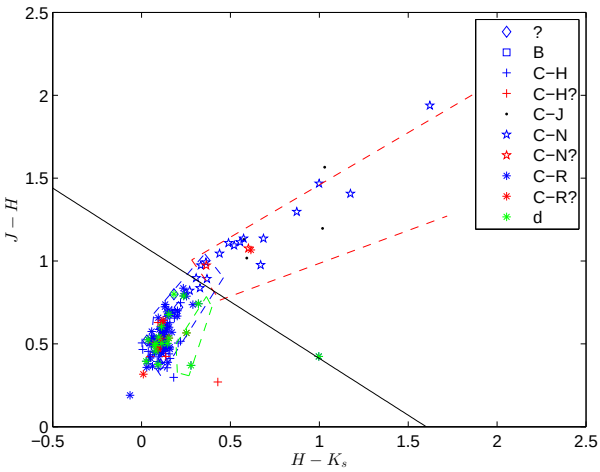


Fig. 15 Distributions of our carbon stars in the diagram of $J-H$ versus $H-K_s$ colors from 2MASS.

study star formation and evolution in galaxies in UV bands, which makes it feasible for detecting hot WDs in unresolved binaries with main-sequence companions as early as G and K types, and cooler WDs with companions that are early M type or later (Green 2013).

Through the CASjob tool of GALEX DR6, we find 81 GALEX NUV-detected carbon stars with a search radius of 3 arcseconds, which correspond to 48 distinct stars because of repeated observations. As evidence for our capacity to detect carbon star binaries with a hot WD companion in GALEX, 48 NUV-detected carbon stars are approximately 26% of our 183 findings, which is about nine times (3%) the spurious match rate, so the vast majority of the 48 stars are true detections.

Of these 48 GALEX-detected carbon stars, 37 are G types. These 37 detections represent about 70% of all (53) the G type carbon stars in our sample, an extremely high detection frac-

Table 6 Objects with NUV and FUV Detections

Designation	nuv_mag (mag)	nuv_magerr (mag)	fuv_mag (mag)	fuv_magerr (mag)
J083021.22+154319.6	22.21	0.17	23.08	0.28
	22.11	0.17	23.01	0.32
J101423.22+302200.4	21.07	0.01	25.20	0.25
J101946.89+252932.8	20.32	0.11	22.45	0.41
	20.80	0.23	21.49	0.36

tion. Only three of them have FUV detections listed in Table 6, of which J083021.22+154319.6 and J101946.89+252932.8 were observed twice. According to Green (2013), a high NUV detection rate of G type carbon stars is not evidence for hot WD companions, and FUV detections could indicate hot WD companions. Therefore, the three FUV G type detections could have hot WD companions. Besides having a hot WD component, UV brightness may arise from the active regions, transition regions, or chromospheric emission of young and active objects. However, none of the three G type FUV-detected carbon stars show emission lines, which is a remarkable feature in such stellar spectra.

4.5 Variability of Our Carbon Stars: Exploring the NSVS, Catalina and LINEAR Databases

In order to study the variability of our carbon stars, we search for our 183 carbon stars in the Northern Sky Variability Survey (NSVS) database[3] (Woźniak et al. 2004b), the Catalina Sky Survey database[4] (Drake et al. 2014) and the LINEAR database[5] (Stokes et al. 2000) by cross matching within a five arcsecond radius.

The NSVS survey was conducted from Los Alamos, New Mexico, and acquired photometry data for approximately 14 million objects. With a 1 yr baseline and typically 100 to 500 measurements for each object, this survey is the most extensive variability survey of the northern sky, and some data in the southern sky are also available in the range $-38^\circ < \delta < 0^\circ$, although with fewer epochs. In a median field, bright unsaturated stars have a point to point magnitude scatter of about 0.02 magnitude and position errors are within 2 arcseconds (Woźniak et al. 2004b).

The Catalina Sky Survey (CSS) database began in 2004. It initially aimed at studying near-earth objects (NEOs), or more specifically, potentially hazardous asteroids (PHAs), and was also used to study variable stars (Drake et al. 2014). The CSS combines data taken from the Mount Lemmon Survey (MLS) in Tucson, Arizona and the Siding Spring Survey (SSS) in Siding Spring, Australia. In this paper, we concentrate on the second public data release data (CSDR2) of CSS, which covers a time span from April 2005 to June 2012. The CSDR2 offered photometry data for 500 million objects (about 40 billion measurements), and the photometry data encompass different observation epochs, Catalina V magnitudes and errors. The sky coverage of this survey is limited to the range $-75^\circ < \delta < 70^\circ$ and $|b| > 15^\circ$, for an area of 33 000 square degrees (Drake et al. 2014).

The LINEAR database began in 1998 and ended in 2009, and compiled photometry data for about 25 million objects (over 5 billion measurements). The sky coverage of this survey is smaller than that of Catalina, but extends over more than about 10 000 square degrees in the northern hemisphere. The photometry errors are typically 0.2 mag at Sloan r band magnitude ~ 18 (Stokes et al. 2000).

In our carbon stars, 79 were found to have entries in the NSVS data, 80 were found to have entries in the Catalina data, and 42 were found to have entries in the LINEAR data. In order to study their variability, a periodogram of the data was obtained first. Then, we used the sinusoidal function defined in Equation (6) to fit the light curves by the Levenberg-Marquardt non-linear least

[3] *http://skydot.lanl.gov/nsvs/nsvs.php*

[4] *http://www.lpl.arizona.edu/css/*

[5] *https://astroweb.lanl.gov/lineardb/*

Table 7 Four Variable Carbon Stars

Designation	P1[a] (d)	P2[b] (d)	P3[c] (d)	A1[d] (mag)	A2[e] (mag)	A3[f] (mag)	Class
J040401.78+271545.4	370.4	364.3	NA	1.18	1.23	NA	SR
J064815.90+080240.7	200.6	NA	NA	1.12	NA	NA	SR
J133557.08+062355.0	NA	242.4	243.0	NA	1.46	1.42	SR
J220514.58+000845.5	NA	222.5	221.3	NA	2.40	2.22	Mira

Notes: [a] Period derived by NSVS data. [b] Period derived by Catalina data. [c] Period derived by LINEAR data. [d] Amplitude derived by NSVS data. [e] Amplitude derived through Catalina data. [f] Amplitude derived by LINEAR data. [g] 'NA' means parameter is not available.

squares fitting method, and finally retain four objects that have clear periodicity, which represents a proportion of about 2% of the objects analyzed. Other stars were rejected for three reasons as described in Mauron et al. (2014): (1) the number of data points was smaller than 15; (2) the light variation was weak, which could be ascertained by magnitude errors; (3) the light curve was irregular and could not be fitted with the sinusoidal function. The photometry data, fitted sinusoidal function and periodicity are shown in Figure 16, and their fitted periodicity and amplitude are listed in Table 7.

$$y = a1 \times \sin\Big[a2 \times (x - a3)\Big] + a4\,. \tag{6}$$

The four variable carbon stars are J064815.90+080240.7, J040401.78+271545.4, J133557.08 +062355.0 and J220514.58+000845.5 respectively. The star J064815.90+080240.7 is a C-N star with an H_α emission line, which was identified by Woźniak et al. (2004a), and classified as a semiregular variable star with a period of 202 d. Our fitted period is about 200.6 d which is consistent with the result estimated by Woźniak et al. (2004a). The amplitude of the V-magnitudes is 0.55 mag, which verifies that it is a semiregular variable star according to the candidate criteria proposed by Usatov & Nosulchik (2008). The star J040401.78+271545.4 is also a C-N star, which is our new finding, and has entries in both the NSVS data and the Catalina data. The fitted result of NSVS data suggests that its period is 370.4 d, and the amplitude of V-magnitude is 1.18. In addition, the fitted result of Catalina data shows that its period is 364.3 d, and the amplitude of the V-magnitude is 1.23. Similarly, according to the criteria of Usatov & Nosulchik (2008), the object J040401.78+271545.4 is also a semiregular variable star. The stars J133557.08+062355.0 and J220514.58+000845.5 both have entries in the Catalina and LINEAR data, and their periods derived from the two data sets are very consistent and are listed in Table 7. With the criterion mentioned in Mauron et al. (2014), the star J133557.08+062355.0 with Catalina amplitude lower than 1.5, is considered as a semiregular variable star, while the Catalina V-magnitude amplitudes of star J220514.58+000845.5 suggest that it is a Mira star.

5 CONCLUSIONS

Carbon stars are excellent kinematic tracers of galaxies, and can be used to derive the rotation curve of the Milky Way. They can also serve as a viable standard candle for galaxies that can be used to derive the distance of these galaxies. So, it is quite useful to automatically search for them from large datasets. In this paper, we apply the EMR algorithm to search for carbon stars from the LAMOST pilot survey.

In the paper, we analyze the performance of the EMR algorithm with four test experiments using the SDSS DR8 stellar spectra. The first experiment tests the robustness of spectral features used in this paper, and the feature of median filtered spectra combined with continuum-subtracted spectra can significantly improve the performance. Using this combined feature, recalls of the top 2000 cases for all queries with one exception of query 18 are more than 90%, many of which are up to 95%, and all the recalls of the top 3000 are more than 95%, some of which are up to 98%.

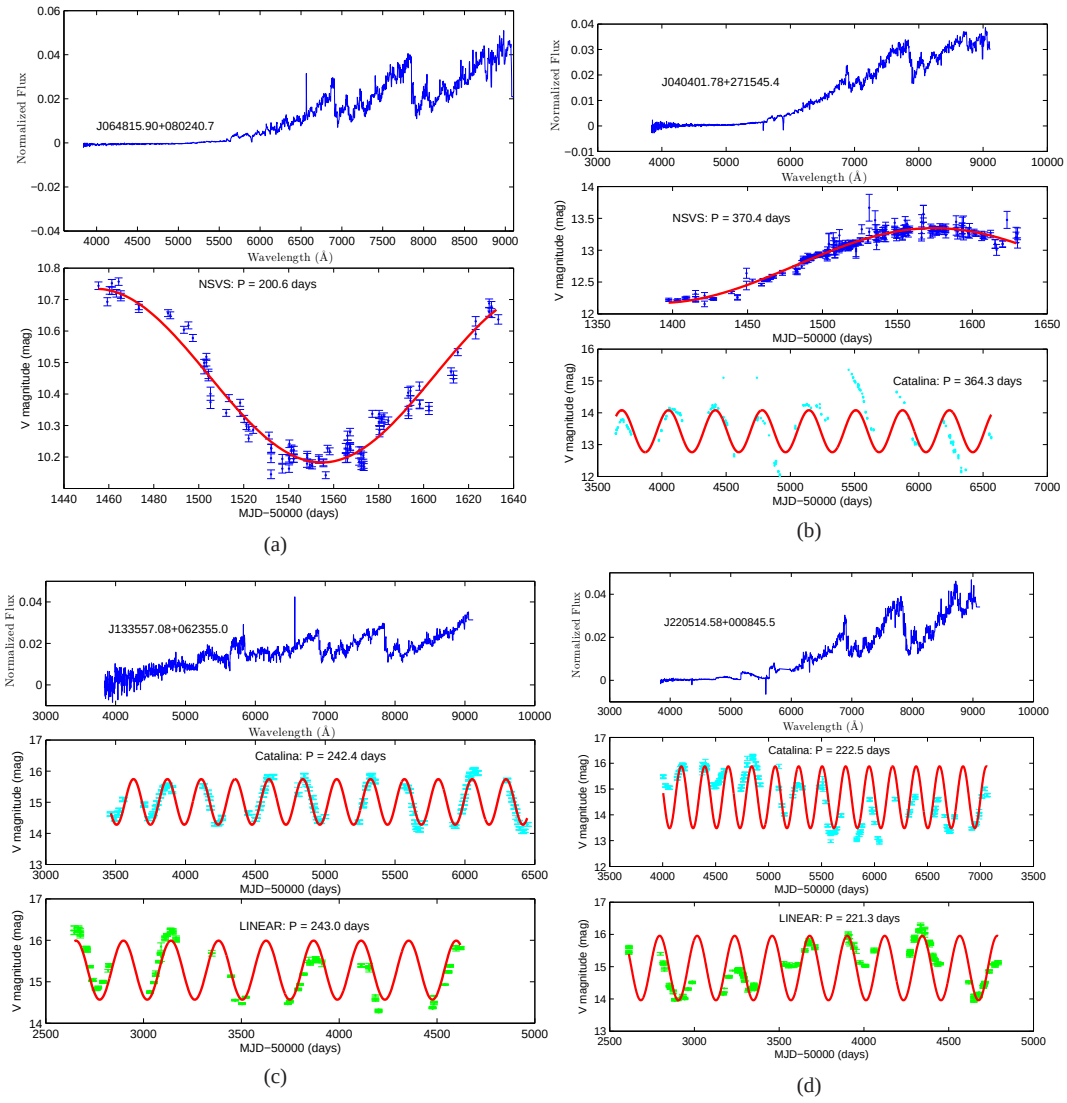


Fig. 16 Light curves of four variables with the NSVS data (*blue*), the Catalina data (*cyan*) and the LINEAR data (*green*), and the fitted sinusoids. (a) J064815.90+080240.7, (b) J040401.78+271545.4, (c) J133557.08+062355.0, (d) J220514.58+000845.5.

The second experiment tests the effect of parameters on the performance of the algorithm, and the algorithm is robust to parameter variability. In the third experiment, we test the performance of the relevance feedback, and confirm that it can also improve the performance of the algorithm. In the fourth experiment, running times of each step of the algorithm are calculated, and we can conclude that the algorithm is quite fast and scalable for a large dataset. In summary, the EMR algorithm is quite efficient and scalable when searching for carbon stars with a large amount of spectra.

After applying the EMR algorithm, we find a total of 183 carbon stars from the LAMOST pilot survey, and 158 of them are new findings. They are classified as 58 C-H stars, 11 C-H star candidates, 56 C-R stars, ten C-R star candidates, 30 C-N stars, three C-N star candidates, and four C-J stars

based on their spectral features. There are also ten objects which have no spectral types because of low quality spectra, and one binary consisting of a WD and a carbon star. Locations of these carbon stars in the diagram of $J-H$ versus $H-K_s$ color have been checked carefully, and we can conclude that the JHK_s colors of 2MASS can separate C-N stars from C-H stars, and we classify the C-N and C-H stars on the $J-H$ and $H-K_s$ color diagram with the optimum linear classification planes obtained by the SVM method.

We identify 18 dC stars from our 183 carbon star samples with three proper motion criteria, and also find three possible carbon star binaries which may have optical invisible companions, which could be a hot WD by cross matching with the GALEX. In addition, four variable carbon stars are found through fitting their light curves, which are obtained from the NSVS database, the CSS database and LINEAR database. Three of them are likely semiregular variable stars and one of them is likely a Mira star.

Acknowledgements We thank an anonymous referee for very useful comments that improved the presentation of the paper. The work was funded by the National Natural Science Foundation of China (Grant Nos. 11390371, 11303036, 11390374, 11233004 and 61202315). 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 has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences. The LAMOST Pilot Survey Web site is *http://data.lamost.org/pdr*. This research also makes use of the Sloan Digital Sky Survey (SDSS) spectra. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is *http://www.sdss3.org/*. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University. This research also makes use of data products from the Two Micron All Sky Survey 2MASS (University of Massachusetts and IPAC/California Institute of Technology, funded by NASA and NSF), the Northern Sky Variability Survey NSVS (Los Alamos National Laboratory and University of Michigan), the Catalina Sky Survey (California Institute of Technology, NASA), and the Lincoln Near-Earth Asteroid Research LINEAR program (Massachusetts Institute of Technology Lincoln Laboratory, NASA and US Air Force). This research makes use of Simbad and Vizier tools offered by the Centre de Données de Strasbourg (Institut National des Sciences de l'Univers, CNRS, France).

References

Barnbaum, C., Stone, R. P. S., & Keenan, P. C. 1996, ApJS, 105, 419

Battinelli, P., Demers, S., Rossi, C., & Gigoyan, K. S. 2013, Astrophysics, 56, 68

Cannon, A. J., & Pickering, E. C. 1918, Annals of Harvard College Observatory, 91, 1

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

Dahn, C. C., Liebert, J., Kron, R. G., Spinrad, H., & Hintzen, P. M. 1977, ApJ, 216, 757

De Mello, A. B., Lorenz-Martins, S., de Araújo, F. X., Bastos Pereira, C., & Codina Landaberry, S. J. 2009, ApJ, 705, 1298

Dean, C. A. 1976, AJ, 81, 364
Demers, S., & Battinelli, P. 2007, A&A, 473, 143
Demers, S., Battinelli, P., & Forest, H. 2009, in IAU Symposium, Vol. 254, IAU Symposium, ed. J. Andersen, Nordströara, B. m, & J. Bland-Hawthorn, 20P
Downes, R. A., Margon, B., Anderson, S. F., et al. 2004, AJ, 127, 2838
Drake, A. J., Graham, M. J., Djorgovski, S. G., et al. 2014, ApJS, 213, 9
Gigoyan, K. S., Russeil, D., Mickaelian, A. M., Sarkissian, A., & Avtandilyan, M. G. 2012, A&A, 544, A95
Goswami, A. 2005, MNRAS, 359, 531
Goswami, A., Karinkuzhi, D., & Shantikumar, N. S. 2010, MNRAS, 402, 1111
Green, P. 2013, ApJ, 765, 12
Green, P. J., Margon, B., Anderson, S. F., & MacConnell, D. J. 1992, ApJ, 400, 659
Gunn, J. E., Carr, M., Rockosi, C., et al. 1998, AJ, 116, 3040
Heber, U., Bade, N., Jordan, S., & Voges, W. 1993, A&A, 267, L31
Jolliffe, I. 2002, Principal Component Analysis (Wiley Online Library)
Keenan, P. C. 1993, PASP, 105, 905
Keenan, P. C., & Morgan, W. W. 1941, ApJ, 94, 501
Koen, C., Marang, F., Kilkenny, D., & Jacobs, C. 2007, MNRAS, 380, 1433
Liebert, J., Schmidt, G. D., Lesser, M., et al. 1994, ApJ, 421, 733
Lloyd Evans, T. 2010, Journal of Astrophysics and Astronomy, 31, 177
Lowrance, P. J., Kirkpatrick, J. D., Reid, I. N., Cruz, K. L., & Liebert, J. 2003, ApJ, 584, L95
Luo, A.-L., Zhang, H.-T., Zhao, Y.-H., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 1243
Margon, B., Anderson, S. F., Harris, H. C., et al. 2002, AJ, 124, 1651
Martin, D. C., Fanson, J., Schiminovich, D., et al. 2005, ApJ, 619, L1
Mauron, N. 2008, A&A, 482, 151
Mauron, N., Gigoyan, K. S., Berlioz-Arthaud, P., & Klotz, A. 2014, A&A, 562, A24
Metzger, M. R., & Schechter, P. L. 1994, ApJ, 420, 177
Richer, H. B., & Crabtree, D. R. 1985, ApJ, 298, L13
Richer, H. B., Crabtree, D. R., & Pritchet, C. J. 1984, ApJ, 287, 138
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Secchi, A. 1869, Astronomische Nachrichten, 73, 129
Si, J., Luo, A., Li, Y., et al. 2014, Science China Physics, Mechanics, and Astronomy, 57, 176
Sloan, G. C., Kraemer, K. E., Matsuura, M., et al. 2006, ApJ, 645, 1118
Stokes, G. H., Evans, J. B., Viggh, H. E. M., Shelly, F. C., & Pearce, E. C. 2000, Icarus, 148, 21
Totten, E. J., & Irwin, M. J. 1998, MNRAS, 294, 1
Totten, E. J., Irwin, M. J., & Whitelock, P. A. 2000, MNRAS, 314, 630
Usatov, M., & Nosulchik, A. 2008, Open European Journal on Variable Stars, 87, 1
Wallerstein, G., & Knapp, G. R. 1998, ARA&A, 36, 369
Wang, S., & Jiang, B. W. 2014, arXiv:1405.1171
Woźniak, P. R., Williams, S. J., Vestrand, W. T., & Gupta, V. 2004a, AJ, 128, 2965
Woźniak, P. R., Vestrand, W. T., Akerlof, C. W., et al. 2004b, AJ, 127, 2436
Wu, X. D., Kumar, V., Quinlan, R. J., et al. 2008, Knowl Inf Syst, 14, 1
Xu, B., Bu, J. J., Chen, C., et al. 2011, SIGIR, 34, 525
Yamashita, Y. 1972, Annals of the Tokyo Astronomical Observatory, 13, 169
Yamashita, Y. 1975, Annals of the Tokyo Astronomical Observatory, 15, 47
Zhao, G., Zhao, Y.-H., Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, RAA (Research in Astronomy and Astrophysics), 12, 723