The Astrophysical Journal Supplement Series, 246:8 (8pp), 2020 January https://doi.org/10.3847/1538-4365/ab5a7c
© 2020. The American Astronomical Society. All rights reserved.

RR Lyrae Star Candidates from SDSS Databases by Cost-sensitive Random Forests

Jingyi Zhang[1,2], Yanxia Zhang[1], and Yongheng Zhao[1]
[1] Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, People's Republic of China
zyx@bao.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China
Received 2019 September 18; revised 2019 October 23; accepted 2019 October 25; published 2020 January 6

Abstract

With the increase of known RR Lyrae stars, it is reliable to create classifiers of RR Lyrae stars based on their photometric data or combined photometric and spectroscopic data. Nevertheless the total number of known RR Lyrae stars is still too small compared with the large survey databases. So classification of RR Lyrae stars and other sources belongs to imbalanced learning. Based on Sloan Digital Sky Survey (SDSS) photometric and spectroscopic data, we apply cost-sensitive Random Forests fit for imbalanced learning to preselect RR Lyrae star candidates. Only with photometric data, $u-g$, $g-r$, $r-i$, $i-z$ is the best input pattern. While also considering physical parameters ($T_{\rm eff}$, [Fe/H], log(g)), the optimal input pattern is $T_{\rm eff}$, [Fe/H], log(g), $u-g$, $g-r$, $r-i$, $i-z$, at this moment for cost-sensitive Random Forests, the performance metrics of completeness, contamination, and Matthews correlation coefficient are 0.975, 0.019, and 0.975, respectively. It indicates that adding stellar physical parameters is helpful for identifying RR Lyrae stars from other stars. We apply the best classifiers on the SDSS photometric data and combined photometric data with physical parameters to select RR Lyrae star candidates. Finally 11,041 photometric candidates with spectral type A and F are obtained, and then 304 candidates with physical parameters are selected out. Among the 304 candidates, a small part are HB stars, BS stars, RGB stars, and peculiar stars, and the rest are unknown in the Simbad database. These candidates may be used as the input catalog for time-series follow-up observations.

Unified Astronomy Thesaurus concepts: RR Lyrae variable stars (1410); Stellar types (1634); Astronomy databases (83); Metallicity (1031); Random Forests (1935); Astrostatistics (1882); Astronomy data analysis (1858)

Supporting material: machine-readable table

1. Introduction

RR Lyrae stars have been studied for over a century now, and were first discovered in nearby globular clusters. These stars have periods of 0.2–1.1 days and present brightness variations of the order of a magnitude. They pulsate in radius and luminosity over short periods with the brightness rising quickly to its peak followed by a slow and gradual drop (Kayal & Benacquista 2013). RR Lyrae stars all have an average absolute magnitude of ∼0.5, and serve as standard candles to determine distances by comparing the apparent and absolute magnitudes. Further, they are the RR Lyrae stars that fix the cosmological distance scale and witness the evolution of the universe as mainly Population II stars. RR Lyrae stars may be taken as good tracers to study the structure, formation, and evolution of the Galactic halo, which helps us understand the history of our Milky Way as well as that of other galaxies. Therefore it is an important issue to collect as many RR Lyrae stars as possible and separate them from other sources.

Many previous works focused on the Sloan Digital Sky Survey (SDSS) colors to separate RR Lyrae stars from the large databases. Krisciunas et al. (1998) concentrated on the identification of the RR Lyrae stars based on the SDSS colors. Ivezić et al. (2000) selected 148 RR Lyrae star candidates by SDSS color criteria for about 100 deg^2 of sky. Ivezić et al. (2005) also picked out RR Lyrae star candidates, which were efficiently recognized even with single-epoch data.

Not only the colors, but also the physical parameters are used to study the properties of RR Lyrae stars. Metallicity is concerned with the chemical composition of a star. Since RR Lyrae stars are repeatedly converting between singly and doubly ionized helium, the abundance of helium is of importance in determining properties of an RR Lyrae star. Furthermore, an RR Lyrae star is a relatively low-mass star with low metallicity. Therefore, RR Lyrae stars are considered to directly reflect the original abundance of heavy elements in the gas cloud. Many studies demonstrate that many properties are related to the metallicity of RR Lyrae stars. For instance, when metallicity decreases, periods of RRab and RRc stars increase slightly. Such relations of the properties of RR Lyrae stars in globular clusters and in the galactic field still exist. Nemec et al. (2013) studied the spectroscopic iron-to-hydrogen ratios, radial velocities, atmospheric parameters, and new photometric analyses of 44 RR Lyrae stars. They concluded that there were the empirical Fourier-based P–[Fe/H] relations for non_Blazhko and most Blazhko RRab stars. Haschke et al. (2012) presented for the first time a detailed spectroscopic study of chemical element abundances of metal-poor RR Lyrae stars in the Large and Small Magellanic Cloud. They also determined abundance ratios for 10 chemical elements and found that spectral synthesis of the α-elements in metal-poor field RR Lyrae stars revealed a mean α-enhancement of 0.36 ± 0.25 dex in very good agreement with these other sources.

SDSS survey provides photometric data and spectroscopic data. The basic atmospheric parameters (effective temperature $T_{\rm eff}$, surface gravity log(g), and metallicity [Fe/H]) for stars are also presented. Based on SDSS databases, we may study properties of RR Lyrae stars and other stars. Given stellar parameters, we further explore the difference between RR Lyrae stars and other stars.

In this work, our primary goal is to study classification characteristics of RR Lyrae stars and get the RR Lyrae star candidates. This paper is organized as follows. Section 2 provides a brief description of the SDSS survey. How to get the samples is also presented in detail. In Section 3, we describe the cost-sensitive Random Forest algorithm and apply it to the samples. Section 4 discusses the performance of the approach in our case and analyzes the results. Finally, we draw the conclusions and present future work.

2. The Data

SDSS (York et al. 2000) is a major multifilter imaging and spectroscopic survey supported by a 2.5 m optical telescope, which provides a wide field of view, and has five filters located in the u, g, r, i, and z bands, which range from 0.36 to 0.90 μm. The processed data include all photometric and spectroscopic observations. Compared to the prior SDSS Data Release, SDSS Data Release 15 (DR15) provides more robust and precise photometric and spectroscopic data.

All SDSS data are available from the Catalog Archive Server (CAS) database. First, all standard stars are collected from the Stripe82 table with photometry. The number of SDSS standard stars is 1,006,843. We refer to this particular sample throughout this paper as Sample 1. A cross-match between them and the sppParams database is performed within a radius of $3\farcs0$. Then we can get the counterparts for sources with stellar parameters, which contains 27,735 standard stars with stellar parameters, hereinafter considered as Sample 2.

The *Gaia* survey provides 140,784 RR Lyrae stars. In order to obtain as many RR Lyrae stars as possible, we primarily use the Optical Gravitational Lensing Experiment (OGLE; Soszyński et al. 2016) catalogs of RR Lyrae stars, and also collect RR Lyrae stars by the CTRS (Drake et al. 2013a, 2013b, 2014, 2017; Torrealba et al. 2015), ASAS (Pojmański 1997; Richards et al. 2012), ASAS-SN (Jayasinghe et al. 2018), ATLAS (Tonry et al. 2018), IOMC (Alfonso-Garzón et al. 2012), LINEAR (Palaversa et al. 2013), and NSVS (Kinemuchi et al. 2006), as well as from the works based on the Simbad database (Wenger et al. 2000). The whole number of known RR Lyrae stars and *Gaia* RR Lyrae stars amounts to 213,476. To understand the statistical properties of RR Lyrae stars with SDSS photometries, the whole RR Lyrae star sample is cross-matched with the SDSS database in $3\farcs0$. The number of SDSS RR Lyrae stars is 4239. Hereinafter it is taken as Sample 3. Then we further perform cross-matching of the sppParams database with the whole RR Lyrae star sample to get physical parameters from SDSS. Finally we obtain 3932 SDSS RR Lyrae stars with stellar parameters, short for Sample 4. Also we obtain 239,302 sources from sppParams table to select RR Lyrae star candidates using the algorithm of this work. For these sources, we only keep spectral type A and F stars avoiding other types of star pollution since most RR Lyrae stars belong to spectral type A and F stars. We name this particular sample throughout this paper Sample 5.

3. The Method

3.1. Random Forest

Random Forest is an improvement over bagged decision trees (Breiman 2001). There is a problem in decision trees that they are greedy. Decision trees select the feature using a greedy algorithm that we make the locally optimal choice at each stage with the intent of finding a global optimum. So the decision trees may have many structural similarities and high correlation in their predictions. Random forest improves the algorithm for the way that the subtrees are learned so that the resulting predictions have less correlation. There are many advantages of Random Forest (Gao et al. 2009). Among many machine-learning algorithms, the accuracy of Random Forest is higher. Given a large database, it can handle efficiently. In the forest building progress, it produces an inherent unbiased estimate of the generalization error. In addition, it has an efficient way of estimating missing data and maintains accuracy when a large fraction of the data are missing. Also, it is helpful for imbalanced data.

3.2. Cost-sensitive Random Forest

Although Random Forest is a powerful machine-learning algorithm owing to its practical advantages, the standard Random Forest is not suited to some cases, like imbalanced data. In such cases, classification methods tend to be biased toward the majority class. These algorithms are inefficient in this case mainly because they seek to maximize a measure of performance such as accuracy which is no longer proper for skewed data. Accuracy treats equally the correctly and incorrectly classified examples of different classes. When dealing with the balanced data, the class weight of the positive and the negative class is usually the same. Similar to the previous work (Zhang et al. 2018), we are more interested in minority. If not setting different costs for different class, the prediction is inclined to the majority. A learning algorithm ensembling cost-sensitivity is used to deal with imbalanced data. In other words, a learning algorithm factors in the costs when building a classifier. For adjusting the imbalance of cost, we set Random Forest with the parameter C of class i to class_weight $[i] * C$, called cost-sensitive learning. So far a lot of research has been done in this area. For example, Yin & Yuping (2014) proposed a cost-sensitive algorithm based on Random Forest and their results showed that the cost-sensitive Random Forest achieved higher performance. Here the program of cost-sensitive Random Forest is adopted from the SCIKIT-LEARN library (Pedregosa et al. 2011).

3.3. Evaluation Metric

Completeness, contamination, and Matthews correlation coefficient (MCC) are used to evaluate the performance of algorithms. As provided by Sesar et al. (2007), the completeness is defined as the fraction of recovered RR Lyrae stars in the whole RR Lyrae sample and the contamination is defined as the fraction of false RR Lyrae stars in the predicted RR Lyrae sample. The MCC metric was first introduced by Matthews (1975) to assess the performance of protein secondary structure prediction. MCC is used in the machine-learning algorithm as a measure of the quality of binary classifications. It takes into consideration true and false positives and negatives and is generally regarded as a balanced measure that can be used even if the classes are of very different sizes. It returns a value between -1 and $+1$. A coefficient of $+1$ represents a perfect prediction, 0 an average random prediction, and -1 the worst possible prediction. In the study of Sabri et al. (2017), they compared the four metrics: MCC, Area Under ROC Curve (AUC), Accuracy, and F1. Accuracy and F1 metrics gave various performance evaluation for different classifiers and these two metrics were sensitive to the data imbalance. While

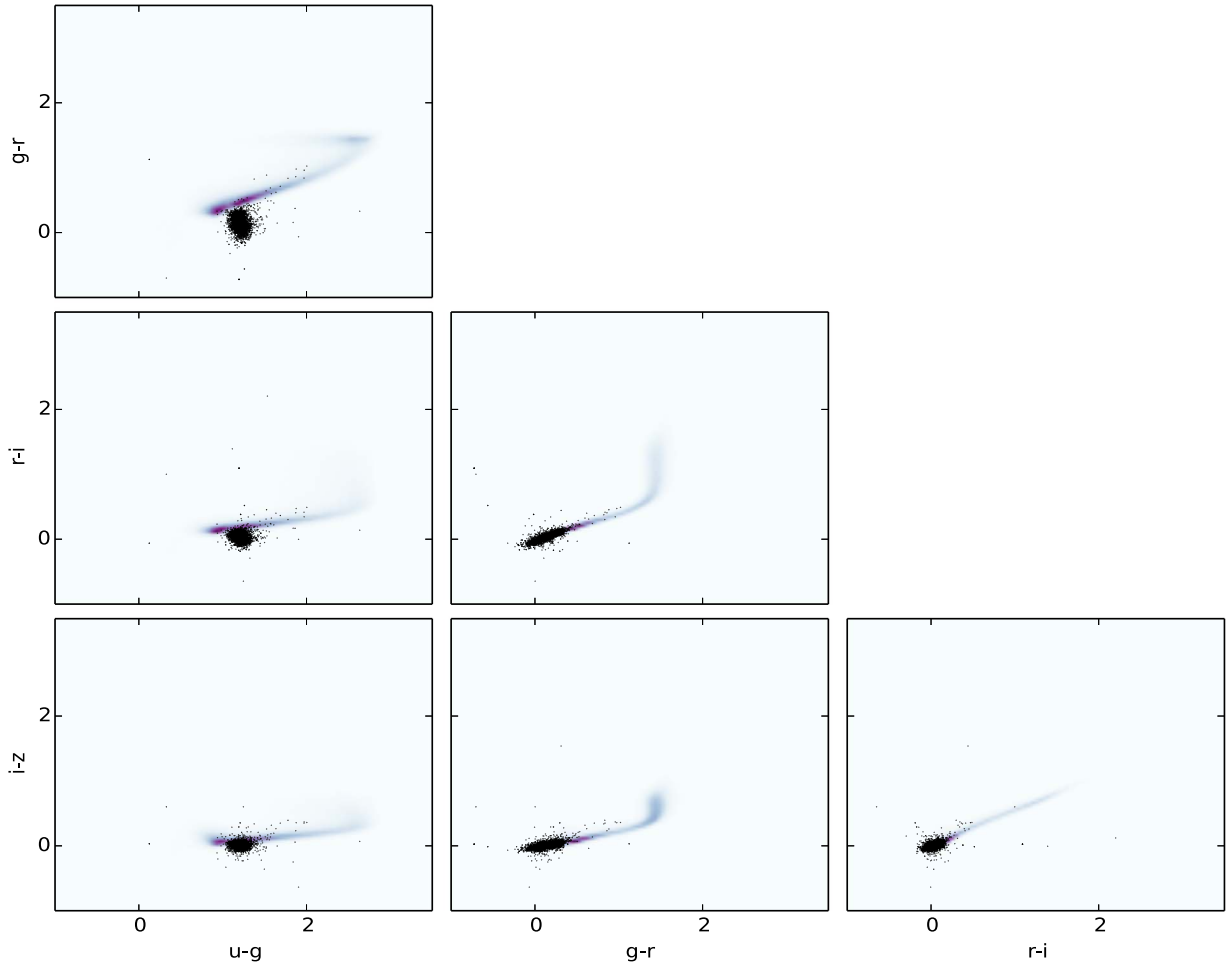

Figure 1. Contour and scatter plot of different parameters. The scatter points represent RR Lyrae stars, and the contour plot represents the other stars.

metrics MCC and AUC had shown constant performance for different classifiers. Therefore MCC and AUC were robust to data imbalance. There was a limitation of using AUC, which was no explicit formula to compute AUC. However, MCC had a close form and it was very well suited to evaluate the optimal classifier for imbalanced data. As a result, here we use the completeness, contamination and MCC to evaluate the performance of a classifier.

4. Results and Discussion

Our aim is to select RR Lyrae star candidates from the whole star sample. Sample imbalance, the performance of a classifier may influence the performance of selecting RR Lyrae stars. Our work belongs to the imbalance problem. Here we plan to apply cost-sensitive Random Forest to choose RR Lyrae stars.

The samples are randomly split into training and test sets, whose number ratio is 7:3. For cost-sensitive Random Forest, the model parameters and imbalance weighting parameter need be adjusted in order to achieve the best performance. The n estimators range from 10 to 100. The imbalance weighting parameter for the negative class is c, whose value goes from 0.1 to 1 and its step is 0.1. We use the tenfold cross-validation during the validation stage.

First the distribution of the SDSS standard stars and all RR Lyrae stars in the color spaces is indicated in Figure 1. From Figure 1, it is still difficult to discriminate RR Lyrae stars from other stars in the 2D color spaces although the number of known RR Lyrae stars increase compared to the work (Zhang et al. 2018). So we use Sample 1 and Sample 3 to build cost-sensitive Random Forest classifiers. The different color combinations ($u-g$, $g-r$), ($u-g$, $g-r$, $r-i$), and ($u-g$, $g-r$, $r-i$, $i-z$) are adopted as input patterns. For different input patterns, cost-sensitive Random Forest shows different performances. Table 1 shows that the performances of cost-sensitive Random Forest are given by means of metrics. According to this table, it is found that more colors are easy to separate RR Lyrae stars from other stars efficiently. When $u-g$, $g-r$, $r-i$, $i-z$ as input pattern, the optimal completeness, contamination, and MCC are 0.853, 0.135, and 0.858, respectively. So ($u-g$, $g-r$, $r-i$, $i-z$) is the best input pattern for classifying the RR Lyrae stars and other stars. It is also concluded that the performance of a classifier is seriously influenced by the training sample, especially known RR Lyrae stars available, with the work of Zhang et al. (2018) for contrast. That is to say, it is of great value to construct a complete and representative training sample. There is always a big gap between ideal and reality. In reality, we try to obtain a training sample as complete and representative as possible. Moreover the performance of a classifier is directly limited by the training sample. For different training samples, the same classifier has variant performance. No single algorithm always remains invincible. For Fast Boxes, we hypothesize that the

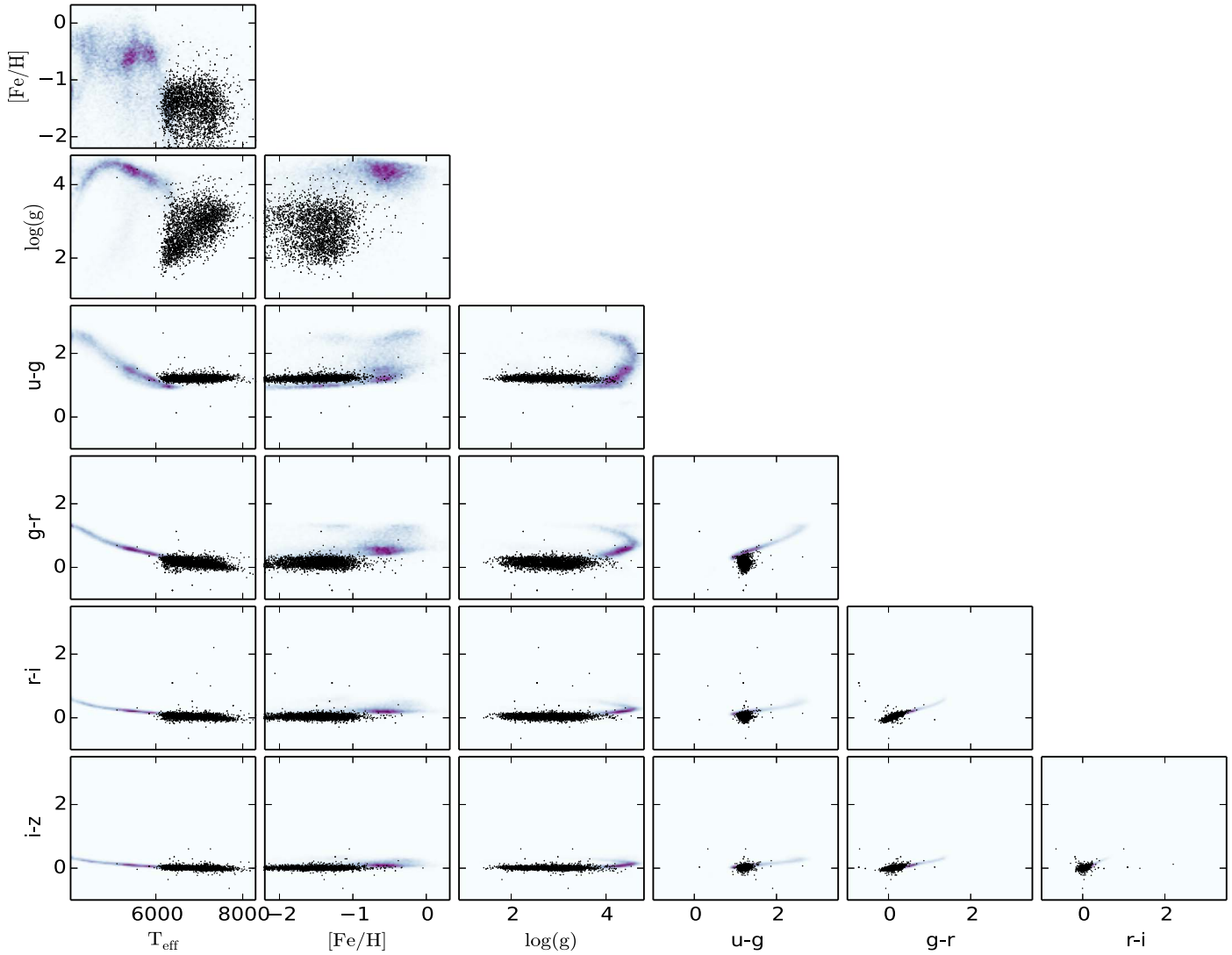

Figure 2. Contour and scatter plot of different parameters. The scatter points represents RR Lyrae stars, and the contour plot represents the other stars.

Table 1
Performance of Cost-sensitive Random Forests with Different Colors

Input Patterns	Completeness	Contamination	MCC
$u-g, g-r$	0.837 ± 0.011	0.173 ± 0.006	0.831 ± 0.006
$u-g, g-r, r-i$	0.853 ± 0.015	0.157 ± 0.011	0.847 ± 0.005
$u-g, g-r, r-i, i-z$	0.853 ± 0.008	0.135 ± 0.007	0.858 ± 0.006

Table 2
Performance of Cost-sensitive Random Forests with Colors and Stellar Parameters

Input patterns	Completeness	Contamination	MCC
$T_{\rm eff}$, $u-g, g-r, r-i, i-z$	0.966 ± 0.006	0.048 ± 0.003	0.953 ± 0.004
[Fe/H], $u-g, g-r, r-i, i-z$	0.961 ± 0.006	0.053 ± 0.006	0.948 ± 0.004
log(g), $u-g, g-r, r-i, i-z$	0.960 ± 0.004	0.033 ± 0.005	0.958 ± 0.005
$T_{\rm eff}$, [Fe/H], $u-g, g-r, r-i, i-z$	0.971 ± 0.004	0.024 ± 0.004	0.970 ± 0.002
$T_{\rm eff}$, log(g), $u-g, g-r, r-i, i-z$	0.971 ± 0.004	0.021 ± 0.004	0.971 ± 0.003
$T_{\rm eff}$, [Fe/H], log(g), $u-g, g-r, r-i, i-z$	0.975 ± 0.003	0.019 ± 0.004	0.975 ± 0.003

positives cluster relative to the negatives. This means that the first step is to cluster the positives and then discriminate the positives from negatives. In detail, we draw a high dimensional axis-parallel box around each cluster and then adjust each boundary locally. If the cluster assumption about the class distributions is not correct, then Fast Boxes could meet

Table 3
Known Sources in Simbad Database

Sources	Type	R.A.	Decl.	$T_{\rm eff}$	[Fe/H]	log(g)	u	g	r	i	z
SDSS J232343.55+531738.3	BS	350.93151	53.293978	8457.781	−0.1002491	4.337116	17.13265	15.59565	15.33409	15.25784	15.22331
SDSS J022351.40+000253.1	HB	35.964172	0.048081	5731.531	−1.451072	1.377947	19.30484	18.07692	17.65233	17.46306	17.39419
SDSS J160553.38+045820.4	HB	241.47241	4.972317	8521.262	−1.595541	3.26835	15.43514	14.16756	16.0089	14.44309	14.50237
SDSS J092253.11+144424.9	RGB	140.7213	14.740269	5183.594	−2.136045	2.330663	17.12849	15.84953	15.26474	14.99032	14.88106
SDSS J111028.59-165026.6	HB	167.61913	−16.840746	5886.702	−1.273482	3.223569	19.21599	18.18251	17.7289	17.56809	17.53925
SDSS J222146.49-000723.4	BS	335.44374	−0.12318257	7960.037	−1.250327	4.279664	18.45947	17.28048	17.24497	17.35355	17.38006
SDSS J112525.62-020804.4	HB	171.35675	−2.1345804	8901.773	−2.020908	3.747	18.4117	17.31343	17.51562	17.70451	17.79642
SDSS J124413.20+350251.5	HB	191.05504	35.047659	7783.068	−2.137797	3.885145	19.38168	18.12946	18.20817	18.26938	18.38394
SDSS J160722.91+104656.8	HB	241.84547	10.782462	8504.273	−1.792795	3.275842	16.86022	15.69523	15.82101	15.95104	16.05928
SDSS J101515.20+070456.9	BS	153.81337	7.082504	8186.453	−0.7701482	4.21986	19.64441	18.49322	18.61356	18.74877	18.90405
SDSS J173118.30+071820.3	RGB	262.82626	7.305667	4935.06	−1.046277	2.498016	18.1761	16.42437	15.68004	15.34589	15.19146
SDSS J201958.72-132540.5	HB	304.99468	−13.427916	7811.433	−1.847607	3.526709	19.00633	17.66319	17.65935	17.71493	17.7593
SDSS J120259.05+140206.5	RGB	180.74608	14.035147	5043.187	−1.361025	1.921489	18.32556	16.71644	16.02215	15.64718	15.58309
SDSS J234632.34-084824.8	Pe	356.63473	−8.806893	5106.36	−2.419248	2.300944	20.0033	18.46357	17.79573	17.55091	17.44113

Table 4
The Catalog of RR Lyrae Star Candidates

Source	R.A.	Decl.	$T_{\rm eff}$	[Fe/H]	log(g)	u	g	r	i	z
SDSS J221247.62+690759.6	333.19851	69.133248	6124.168	−0.1681285	3.48211	17.56385	15.96563	15.22771	14.87916	14.67276
SDSS J165204.91+231252.6	253.02049	23.21464	6941.499	−0.9811056	3.891449	19.95932	18.71199	18.49281	18.38769	18.3522
SDSS J231236.94+210138.9	348.15395	21.027479	6881.329	−1.191111	3.888561	18.20325	17.05122	16.66425	16.52124	16.42668
SDSS J004201.85+004256.4	10.507744	0.71568	6242.705	−1.343125	3.548861	19.80403	18.79634	18.55596	18.42665	18.4069
SDSS J181155.57+234230.6	272.98156	23.708542	5648.04	−0.3218291	3.878797	19.7755	18.19112	17.55933	17.38127	17.25303
SDSS J125141.79+602529.5	192.92416	60.424852	5662.54	−0.5734528	4.367483	20.54886	19.38606	18.95858	18.77245	18.7179
SDSS J111640.57+405450.8	169.16906	40.914116	5674.414	−1.397757	3.334078	20.28753	19.2898	18.9151	18.72502	18.65285
SDSS J235935.22+265351.8	359.89679	26.897733	5745.736	0.08807483	3.747856	16.09045	14.64571	14.10701	13.97607	13.89992
SDSS J081505.54-081541.8	123.77311	−8.261641	8230.285	−1.131375	4.121557	16.30152	14.912	14.75429	14.80451	14.76955
SDSS J211912.47+002538.5	319.80197	0.427362	6135.098	−1.377403	4.433465	19.28309	18.30943	17.92994	17.78351	17.71742

(This table is available in its entirety in machine-readable form.)

problems. So we adopt cost-sensitive Random Forest instead of Fast Boxes in this work.

With the arrival of huge spectroscopic data, the stellar physical parameters are measured. We want to check whether stellar physical parameters affect the classification of RR Lyrae stars and other stars. Figure 2 shows that the colors with stellar parameters or stellar parameter combination can separate the RR Lyrae stars from other stars better. Then Sample 2 and Sample 4 are applied to generate different cost-sensitive Random Forest classifiers. Table 2 displays the performance of cost-sensitive Random Forest with best color input pattern ($u-g$, $g-r$, $r-i$, $i-z$) and stellar parameters ($T_{\rm eff}$, [Fe/H], $\log(g)$). The results show that ($T_{\rm eff}$, [Fe/H], $\log(g)$, $u-g$, $g-r$, $r-i$, $i-z$) can build the best classifier, whose completeness, contamination, and MCC achieve 0.975, 0.019, and 0.975, respectively. It indicates that adding stellar parameters is helpful for identifying RR Lyrae stars by comparison of Tables 1 and 2.

Based on the above results, we use the best input patterns to select RR Lyrae star candidates from the SDSS database. In order to get more accurate candidates, two procedures are processed. First, based on the SDSS standard stars and all RR Lyrae stars for training with $u-g$, $g-r$, $r-i$, $i-z$ as input pattern, we build a cost-sensitive Random Forest classifier to select RR Lyrae candidates from SDSS databases, and get about 11,041 RR Lyrae star candidates from Sample 5. Then with ($T_{\rm eff}$, [Fe/H], $\log(g)$, $u-g$, $g-r$, $r-i$, $i-z$) as an input pattern, another cost-sensitive Random Forest classifier is constructed to select RR Lyrae stars from these 11,041 RR Lyrae star candidates. Finally we obtain 304 RR Lyrae star candidates. Further to check the 304 sources, we cross-match the sources with the Simbad survey database. Among them, there are seven HB stars, three BS stars, three RGB stars, and one peculiar star. The rest of the sources are unknown in Simbad. Table 3 shows the 14 known sources. Table 4 shows the details of all the candidates. It is concluded that RR Lyrae stars are easy to be confused with HB stars, BS stars, and RGB stars since RR Lyrae stars have evolved from the red giant branch (RGB) to their present position on the horizontal branch (HB) when the helium flash initiates the triple-alpha reaction in the degenerate core.

5. Conclusion

Serving as distance indicators, RR Lyrae stars are still a hot and challenging issue in astronomy, and worthy of research, for instance, a tracer used to study the structure and evolution of the Galactic halo, and fixing the cosmological distance scale. By means of the statistical distribution of different stellar physical parameters and photometries, we study the classification characteristics of RR Lyrae stars. In terms of highly skewed data, we adopt cost-sensitive Random Forests with different input patterns to separate the RR Lyrae stars from the other types of stars. Finally, we combine the stellar parameters with photometry to obtain the RR Lyrae star candidates. These candidates may be performed time-series follow-up observation. In the future work, we plan to use the time-domain data. Given time-domain, asynchronous data and the deeper magnitudes of celestial objects, we will select many more RR Lyrae star candidates by automated methods.

We are very grateful to the referee for his constructive comments. This paper is funded by 973 Program 2014CB845700 and the National Natural Science Foundation of China under grants No.11873066 and No.U1731109. We acknowledgment SDSS databases. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS website is www.sdss.org. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) /University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

ORCID iDs

Yanxia Zhang https://orcid.org/0000-0002-6610-5265

References

Alfonso-Garzón, J., Domingo, A., Mas-Hesse, J. M., & Giménez, A. 2012, A&A, 548, A79
Breiman, L. 2001, Machine Learning, 45, 5
Drake, A. J., Catelan, M., Djorgovski, S. G., et al. 2013a, ApJ, 763, 32
Drake, A. J., Catelan, M., Djorgovski, S. G., et al. 2013b, ApJ, 765, 154
Drake, A. J., Djorgovski, S. G., Catelan, M., et al. 2017, MNRAS, 469, 3688
Drake, A. J., Graham, M. J., Djorgovski, S. G., et al. 2014, ApJS, 213, 9
Gao, D., Zhang, Y., & Zhao, Y. 2009, RAA, 9, 220
Haschke, R., Grebel, E. K., Duffau, S., & Jin, S. 2012, AJ, 143, 33
Ivezić, Z., Goldston, J., Finlator, K., et al. 2000, AJ, 120, 963
Ivezić, Z., Vivas, A. K., Lupton, R. H., et al. 2005, AJ, 129, 1096
Jayasinghe, T., Kochanek, C. S., Stanek, K. Z., et al. 2018, MNRAS, 477, 3145
Kayal, K., & Benacquista, M. 2013, AAS, 221, 354.10
Kinemuchi, K., Smith, H. A., Woźniak, P. R., McKay, T. A. & ROTSE Collaboration 2006, AJ, 132, 1202
Krisciunas, K., Margon, B., & Szkody, P. 1998, PASP, 110, 1342
Matthews, B. W. 1975, Biochimica et Biophysica Acta Protein Structure, 405, 442
Nemec, J. M., Cohen, J. G., Ripepi, V., et al. 2013, ApJ, 773, 181
Palaversa, L., Ivezić, Z., Eyer, L., et al. 2013, AJ, 146, 101
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825, http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
Pojmański, G. 1997, AcA, 47, 467
Richards, J. W., Starr, D. L., Miller, A. A., et al. 2012, ApJS, 203, 32
Sabri, B., Fethi, J., Mohammed, E. A., & Quan, Z. 2017, PLoSO, 12, e0177678

Sesar, B., Ivezić, Z., Lupton, R. H., et al. 2007, AJ, 134, 2236
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2016, AcA, 66, 131
Tonry, J. L., Denneau, L., Heinze, A. N., et al. 2018, PASP, 130, 064505
Torrealba, G., Catelan, M., Drake, A. J., et al. 2015, MNRAS, 446, 2251
Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9
Yin, H., & Yuping, H. 2014, Engin. J. Wuhan Univ., 5, 707, http://caod.oriprobe.com/articles/42879322/A_cost_sensitive_algorithm_based_on_random_forest.htm
York, D. G., Adelman, J., Anderson, J. E., et al. 2000, AJ, 120, 1579
Zhang, J., Zhang, Y., & Zhao, Y. 2018, AJ, 155, 108