J-PLUS: Support vector machine applied to STAR-GALAXY-QSO classification

C. Wang (存)1,2, Y. Bai (白宇)3, C. López-Sanjuan7,8, H. Yuan (苑海波)3, S. Wang (王松)1, J. Liu (刘继峰)1,2, D. Sobral3, P. O. Baqui5, E. L. Martin15,6,7, C. Andres Galarza10, J. Alcaniz10, R. E. Angulo11,12, A. J. Cenarro7,8, D. Cristóbal-Hornillos7,8, R. A. Dupke10,13,14, A. Ederoclite9, C. Hernández-Monteagudo15,6, A. Marín-Franch7,8, M. Moles7,8, L. Sodré Jr.9, H. Vázquez Ramió7,8, and J. Varela7,8

1 Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, 20A Datun Road, Chaoyang District, Beijing 100012, PR China
e-mail: wangcunshi@nao.cas.cn
2 College of Astronomy and Space Sciences, University of Chinese Academy of Sciences, Beijing 100049, PR China
3 Department of Astronomy, Beijing Normal University, Beijing 100875, PR China
4 Department of Physics, Lancaster University, Lancaster LA1 4YB, UK
5 PPGFis & Núcleo de Astrofísica e Cosmolologia (Cosmo-ufes), Universidade Federal do Espírito Santo, 29075-910 Vitória, ES, Brazil
6 Departamento de Astrofisica, Universidad de La Laguna (ULL), 38206 La Laguna, Tenerife, Spain
7 Consejo Superior de Investigaciones Científicas (CSIC), 28006 Madrid, Spain
8 Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Unidad Asociada al CSIC, Plaza San Juan 1, 44001 Teruel, Spain
9 Instituto de Astronomía, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, 05508-090 São Paulo, Brazil
10 Observatório Nacional – MCTI (ON), Rua Gal. José Cristino 77, São Cristóvão 20921-400, Rio de Janeiro, Brazil
11 Donostia International Physics Centre (DIPC), Paseo Manuel de Lardizabal 4, 20018 Donostia-San Sebastián, Spain
12 IKERBASQUE, Basque Foundation for Science, 48013 Bilbao, Spain
13 University of Michigan, Department of Astronomy, 1085 South University Ave., Ann Arbor, MI 48109, USA
14 University of Alabama, Department of Physics and Astronomy, Gallalee Hall, Tuscaloosa, AL 35401, USA
15 Instituto de Astrofísica de Canarias, La Laguna 38205, Tenerife, Spain

Received 19 September 2021 / Accepted 13 December 2021

ABSTRACT

Context. In modern astronomy, machine learning has proved to be efficient and effective in mining big data from the newest telescopes.

Aims. In this study, we construct a supervised machine-learning algorithm to classify the objects in the Javalambre Photometric Local Universe Survey first data release (J-PLUS DR1).

Methods. The sample set is featured with 12-waveband photometry and labeled with spectrum-based catalogs, including Sloan Digital Sky Survey spectroscopic data, the Large Sky Area Multi-Object Fiber Spectroscopic Telescope, and VERONCAT – the Veron Catalog of Quasars & AGN. The performance of the classifier is presented with the applications of blind test validations based on RAdial Velocity Extension, the Kepler Input Catalog, the Two Micron All Sky Survey Redshift Survey, and the UV-bright Quasar Survey. A new algorithm was applied to constrain the potential extrapolation that could decrease the performance of the machine-learning classifier.

Results. The accuracies of the classifier are 96.5% in the blind test and 97.0% in training cross-validation. The F1-scores for each class are presented to show the balance between the precision and the recall of the classifier. We also discuss different methods to constrain the potential extrapolation.

Key words. methods: data analysis – techniques: spectroscopic – astronomical databases: miscellaneous

1. Introduction

Developments in computer science and the technological applications have changed the ways of data processing and knowledge management. Especially, as a growing realm of technology, machine learning has gained worldwide popularity due to its powerful ability to manage large amounts of data. Machine-learning algorithms can reveal potential patterns and physical meanings that are otherwise indistinguishable by traditional methods. Furthermore, machine learning enables us to construct the structure of each observed quantity and to reveal its manner of working.

In modern astronomy, the newest telescopes now produce large amounts of unprocessed data. The Javalambre Photometric Local Universe Survey (J-PLUS, Cenarro et al. 2019) is designed to observe several thousand square degrees in the optical bands. It has been designed to observe more than 13 million objects with the Javalambre Auxiliary Survey Telescope (JAST80) at the Sierra de Javalambre in Spain, and to enhance knowledge from the Solar System to cosmology1, such as the Coma cluster (Jiménez-Teja et al. 2019),

* Data are only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/659/A144

1 http://j-plus.es/survey/science

Open Access article, published by EDP Sciences, under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
low-metallicity stars (Whitten et al. 2019), and galaxy formation (Nogueira-Cavalcante et al. 2019).

The current classification of sources detected by J-PLUS is morphological, this is to say it aims to distinguish between point-like and extended sources (Cenarro et al. 2019; López-Sanjuan et al. 2019). It is therefore not able to differentiate stars from quasi-stellar objects (QSOs), and it does not include valuable color information from the 12 optical J-PLUS bands. This paper presents the spectrum-based classification for the J-PLUS first data release (DR1) with machine-learning algorithms. The input catalog is a modified version of the J-PLUS data set which has been recalibrated by Yuan (in prep.). This version includes 13 265 168 objects with magnitudes obtained with 12 different filters (Sect. 2.1). From Sects. 2.2–2.4, we label the data set as STAR, GALAXY, and QSO based on the spectroscopy surveys, including the Sloan Digital Sky Survey (SDSS), the Large Sky Area Multi-Object Fiber Spectroscopy Telescope (LAMOST), and VERONCAT – the Veron Catalog of Quasars & AGN (VV13).

Several machine-learning algorithms have been applied for the classification (Sect. 3), including the Support Vector Machine (SVM, Cortes & Vapnik 1995), linear discrimination, the $k$-nearest neighbor ($k$-NN, Cover & Hart 1967; Stone 1977), Bayesian, and decision trees (Quinlan 1986). In the pretraining, we adopted the algorithm with the highest accuracy (Sect. 3.1). In Sect. 3.2, we present the processes to test the parameters of the algorithms and to train the classifier. We also provide the blind test and a new method to constrain potential extrapolation (Sect. 3.3) in our prediction.

We present our result in Sect. 4, including our result catalogs (Sect. 4.1), considerations about ambiguous objects from the classification probabilities (Sect. 4.2), and a comparison between the J-PLUS parameter (Sect. 4.3). In Sect. 5, we discuss different methods to constrain the extrapolation. The classifier is compared with other published classifiers, and the difference is analyzed in detail (Sect. 5.2). Section 5.3 gives an outlook of the Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS, Benítez et al. 2014; Bonoli et al. 2021) and our future work.

2. Data

The rapid advance in telescopes and detectors has led to a significant data explosion in modern astronomy. New technologies help us accelerate information acquisition from the huge datasets. Several studies have focused on developing classifiers, and they have proved that spectral-based methods are more reliable than those only based on photometric data (Bai et al. 2018; Ball et al. 2006).

2.1. J-PLUS

J-PLUS$^2$ is being conducted from the Observatorio Astrofísico de Javalambre (OAJ, Teruel, Spain; Cenarro et al. 2014) using the 83-cm JAST80 and T80Cam, a panoramic camera of 9.2k × 9.2k pixels that provides a $2\,\text{deg}^2$ field of view (FoV) with a pixel scale of 0.55 arcsec pix$^{-1}$ (Marín-Franchet et al. 2015). The J-PLUS filter system is composed of 12 passbands, including five broad and seven medium bands from 3000 to 9000 Å. The J-PLUS observational strategy, image reduction, and main scientific goals are presented in Cenarro et al. (2019). J-PLUS DR1 covers a sky area of 1022 deg$^2$, and the limiting magnitudes are in the range 21–22. For different kinds of objects, the magnitudes of these 12 bands exhibit different distributions$^3$, and such a difference gives us a theoretical foundation for object classification.

Compared to other catalogs, the J-PLUS catalog is an ideal data set for classification owing to its characteristic of both large amounts and multiple wavebands. Multiple bands could provide more information for a single object. In machine learning, these 12-band magnitudes lead to a more expanding training instance space and a smoother training structure. We adopted the 12 band magnitudes as training features, which are $u$, $j0378$, $j0395$, $j0410$, $j0430$, $g$, $j0515$, $r$, $j0660$, $i$, $j0861$, and $z$. We name them mag1 through to mag12.

Recently, Yuan (in prep.) recalibrated the J-PLUS catalog and increased the accuracy of photometric calibration by using the method of stellar color regression (SCR), similar to the method in Yuan et al. (2015). The catalog in Yuan (in prep.) contains 13 265 168 objects, including 4 126 928 objects with all 12 valid magnitudes.

2.2. SDSS

The observation of SDSS has covered one-third of the sky and yielded more than 3 million spectra. We explore the spectroscopy survey sets in data release 16 (DR16; Ahumada et al. 2020). With the help of SDSS Catalog Archive Server Jobs$^4$, the objects with zWarning = 0 were chosen to label the J-PLUS data as “STAR”, “GALAXY”, and “QSO”.

The Apache Point Observatory Galactic Evolution Experiment (APOGEE) has observed more than 100 000 stars in the Milky Way, with reliable spectral information including stellar parameters and radial velocities (Zasowski et al. 2013). We adopted the APOGEE catalog to enlarge the training set.

We cross-matched the J-PLUS catalog with SDSS DR16 using Tool for OPerations on Catalogues And Tables (Topcat, Taylor 2005)$^5$ with a tolerance of one arcsec, and we obtained 45 350 stars, 68 381 galaxies, and 44 745 QSOs from the general catalog, as well as 13 749 stars from APOGEE. After cross-matching with other catalogs, APOGEE contributes 6147 independent stars.

2.3. LAMOST

LAMOST (Cui et al. 2012; Luo et al. 2012; Zhao et al. 2012; Wang et al. 1996; Su & Cui 2004) is located at the Xinglong Observatory in China, which is able to observe 4000 objects in 20 deg$^2$ simultaneously. LAMOST has many scientific projects, and two of them aim to understand the structure of the Milky Way (Deng et al. 2012) and external galaxies. The low-resolution spectra of LAMOST have a limiting magnitude of about 20 mag in the $g$ band for a resolution $R = 500$. Data release 7 (DR7) was adopted to label the sample. We also adopted information from stellar catalogs from DR7, including the A-, F-, G-, and K-type star catalog, as well as the A- and M-star catalogs.

The A-, F-, G-, and K-type star catalog has stars with a $g$ band signal-to-noise ratio higher than 6 in dark nights or 15 in bright nights. The A- and M-star catalogs contain all A and M stars from the pilot and general surveys. For overlapping stars, we followed the priority of the star catalogs and the general catalog.

---

$^2$ www.j-plus.es

$^3$ http://j-plus.es/daterelases/data_release_dr1

$^4$ http://skyserver.sdss.org/casjobs/

$^5$ http://www.star.bris.ac.uk/~mbt/topcat/
Table 1. Constitution of a sample set.

| Catalog            | Star | Galaxy | QSO | Total   |
|--------------------|------|--------|-----|---------|
| SDSS DR16          | 45,350 | 68,381 | 44,745 | 158,476 |
| SDSS APOGEE        | 13,749 | 0      | 0    | 13,749  |
| LAMOST DR7         | 299,907 | 16,004 | 4,758 | 345,975 |
| LAMOST A-, F-, G- and K- type stars | 212,114 | 0      | 0    | 212,114 |
| LAMOST A- stars    | 5145 | 0      | 0    | 5145    |
| LAMOST M- stars    | 25,604 | 0      | 0    | 25,604  |
| VV13               | 0    | 0      | 4744 | 4744    |

Notes. The numbers reveal how many objects there are that correspond to each catalog and class, after crossing with the J-PLUS catalog. The sample contains 468,685 objects with a full 12 mag from the catalogs, with 348,085 STAR, 74,701 GALAXY, and 45,899 QSO objects. There are repeated objects in different catalogs, which cause the inequality of the sum. This table presents all objects for training and testing. See Sect. 3.3.2 for a blind test.

Fig. 1. Comparison between the class in the sample set and the J-PLUS “CLASS_STAR” parameter. The panel on the left-hand side shows the normalized distributions of CLASS_STAR. The panel on the right-hand side shows the relation between the average magnitudes in the g band corresponding to each bin (the left panel) of CLASS_STAR. The white box and black line denote the stellar objects, blue stands for the galaxies, and yellow is for the QSOs.

In the LAMOST DR7 catalog, the cross-match yields 299,907 stars, 16,004 galaxies, and 4,758 QSOs. There are 212,114 matched stars in the A-, F-, G-, and K-type star catalog and 5145 and 25,604 stars in the A- and M-star catalogs, respectively. Nearly all of the stars (except only one star) from the star catalogs are covered in the DR7 general catalog.

2.4. QSO catalog

Quasars in VV13 (VERONCAT – Veron Catalog of Quasars & AGN, the 13th edition) were also employed to enlarge our QSO samples. The catalog contains AGN objects with spectroscopic parameters (including redshift; Véron-Cetty & Véron 2010). The VV13 contains 4,744 QSOs after a one arcsec tolerance cross identification with J-PLUS, 4,593 QSOs are included in SDSS DR16, and 1339 QSOs are in LAMOST DR7. The VV13 catalog provides 108 additional QSOs.

2.5. Sample construction

The machine-learning sample is made up of SDSS, LAMOST, and VV13 (Table 1, see more in Appendix C, and magnitude distributions are in Appendix B). There are 468,685 unique objects with 12 valid magnitudes, including 74,701 galaxies, 45,899 QSOs, and 348,085 stars. These 468,685 objects were all put in training with a 10-fold validation. The blind test set was carried out with 2853 objects in other catalogs, see Sect. 3.3.2.

3. Methodology

Machine learning has developed many algorithms that are able to deal with big data effectively. Three of them, that is to say decision trees, SVM, and k-NN, are the most popular ones.

3.1. Pretraining

A pretraining process with 10-fold validation was adopted in order to determine which algorithm fits our problem best. The No-Free-Lunch theorem (Shalev-Shwartz & Ben-David 2014) tells us that a perfect learning algorithm that can fit every problem does not exist. In the pretraining, we considered the accuracy to be the most important factor of the training performance. The accuracies of the pretraining are shown in Table 2.

In the k-NN algorithm, the label of each data point is defined by its neighborhood. By introducing a metric function, the algorithm can calculate the distance between every two objects. For each object, the nearest k-objects are determined, and its label is defined. This process continues until the labels of all objects are stable. The k-NN gives a reasonable result for a nonlinear or discrete training set, and it has good performance when extrapolating a prediction and separating for outliers. However, the k-NN algorithm cannot present reliable results for unbalance data that are dominated by objects in one or two classes (Shalev-Shwartz & Ben-David 2014). This is one reason why we precluded the algorithm. In our test, we adopted a 10-NN algorithm with a Euclid norm, and no hyperparameters or weights were involved.
### Table 2. Accuracy and time cost for algorithms.

| Algorithm     | Accuracy | Time cost |
|---------------|----------|-----------|
| Decision tree | 92.6%    | 96 s      |
| Linear discrim | 86.9%    | 26 s      |
| Bayesian      | 74.3%    | 10 s      |
| SVM           | 96.4%    | 90 min    |
| k-NN          | 95.7%    | 23 min    |
| AdaBoost      | 92.0%    | 3 min     |
| Random forest | 96.2%    | 7 min     |

**Notes.** The last column is the rough training time cost for the training sample. The “s” stands for second and “m” is for minute. See Fisher (1936) for details about the linear discrimination algorithm.

Decision tree is a nonparametric supervised learning method. The tree in the algorithm is built by the threshold calculated from the sample. For each node of a tree, a gain function defines the loss of the prediction (Quinlan 1986). If the loss function is low, the node is split. This procedure continues until all objects in the training set are labeled. The time cost of decision tree is low (Shalev-Shwartz & Ben-David 2014), but the gain function may lead to a bias or overfitting for the unbalanced data set. Random forest (RF, Breiman 2001) and bagging tree are enhanced decision tree algorithms that can decrease overfitting.

In our work, we tested three tree algorithms without hyperparameters. The decision tree algorithm is based on the Gini index (Quinlan 1986), and the maximum split is 100. The RF (Breiman 2001) algorithm also contains 30 learners and maximum splits to 468 684. The AdaBoost algorithm (Freund & Schapire 1997) contains 30 learners with a learning rate of 0.1, and it has a maximum split of 20.

For each model, we examined the accuracy and training time (Table 2), and the SVM algorithm provides the highest validation accuracy. Since the model accuracy is the primary factor in our consideration, we decided to adopt the SVM algorithm even if it needs a relatively long training time. The training time becomes significant in other situations, such as transient detection.

#### 3.2. SVM

SVM is a binary classification method (see Cortes & Vapnik 1995; Boser et al. 1992 for details). The theory of SVM is presented in Cristianini & Shawe-Taylor (2000) and Shalev-Shwartz & Ben-David (2014).

In brief, the SVM algorithm generates a super surface in the instance space by maximizing the margin. The margin is defined by the smallest distance between the object and the super surface. Given a super surface, the algorithm divides the instance space into two parts and labels the object in each part. The algorithm then compares each label with the sample and calculates the loss function. The margin is maximized when the loss function reaches its minimum. For our classification problem, there are 12 dimensions in the instance space.

SVM is a binary classification algorithm, while we are facing a multi-classification algorithm. The coding method can change a multi-classification problem into several binary classifications, such as one-versus-one coding and one-versus-all coding. For a $k$-classification problem, one-versus-one coding finds all binary combinations of the labels. After making a democratic decision, the algorithm produces the predicted label. One-versus-one coding needs $\frac{k(k-1)}{2}$ binary classifications to reach the aim. One-versus-all coding singly picks one label out and defines it as a positive class, and the rest ($k - 1$) of the labels are negative. After $k$ times binary classifications, the one-versus-all coding presents the labels by democratic decision. One-versus-one coding has a higher accuracy in our classification.

The Gaussian kernel, also known as the radial basis function (RBF) kernel, is an important parameter in the SVM algorithm construction. It can accelerate the optimization of the margin in the SVM algorithm. In the Gaussian kernel, a kernel scale is an adjustable parameter that measures the distance to the half-space. A small kernel scale constrains the kernel function in low variation, and further parameterizes the margin exquisitely. The farther the data points are located from the margin, the less they weigh. In order to find the best kernel scale, we tested the scale from 0.5 to 1, with a step size of 0.05. For each kernel scale, we trained a classifier and calculated its accuracy. Finally, we conclude that 0.75 is the best kernel scale (Fig. 2).

The magnitude uncertainties in J-PLUS DR1 (Yuan, in prep.) depend on the observing condition and the photometric calibration. In our training process, we employed uncertainties as the training weight to describe the reliability of the data.

The confusion matrix is shown in Fig. 3. The total cross-validation accuracy is 97%. The low accuracy of QSO may be due to its relatively small sample size.

#### 3.3. Validation

Model validation has been designed to show the effectiveness and to avoid potential overfitting. Extrapolation is significant in model validation. It has been proved that the prediction accuracy might decrease when extrapolating outside the feature space region of training samples (S. Wang, priv. comm.). The other validating procedure is a blind test, which can reveal the potential overfitting of the classifier. Moreover, the appropriate training data size would be implied by comparing the training and blind test accuracy.

**3.3.1. Extrapolation**

Applying any extrapolation may cause low accuracy due to the nonrepresentativeness between training data and predicting data (S. Wang, priv. comm.). Here, we use the density contour.
of the training sample to define the potential extrapolation. A
dozen three-dimensional density contour surfaces were
generated based on the distribution of training data. These surfaces
were used as the boundary of the potential extrapolation. The
magnitude combinations are (mag1, mag2, mag3), (mag2, mag3,
mag4), ..., and (mag12, mag1, mag2), and an example is shown in
Fig. 4. We present all the contour surfaces in Appendix A. We
then define the potential extrapolation with these 12 contour sur-
faces for the prediction. There are 3,496,867 (84.73%) objects of
J-PLUS DR1 located inside these contours.

3.3.2. Blind test

We applied a blind test to reveal the classifier’s validation and
its potential overfitting of the training data. The blind test data
set (Table 3) was built by stars from the RA
dial Velocity Experiment (RAVE) and Kepler Input Catalog (KIC), galaxies from
the 2 MASS (Two Micron All Sky Survey) Redshift Survey
(2MRS) and QSOs from the UV-bright Quasar Survey (UVQS).

Table 3. Constitution of a blind test set.

| Catalog | Interpolation | Extrapolation | Total |
|---------|---------------|---------------|-------|
| RAVE    | 29            | 3             | 32    |
| KIC     | 2071          | 64            | 2135  |
| 2MRS    | 606           | 46            | 652   |
| UVQS    | 16            | 18            | 34    |
| Total   | 2722          | 131           | 2853  |

Notes. Every catalog is crossed with J-PLUS and all 12 mag are avail-
able. The extrapolation stands for the objects suffering from potential
extrapolating, and the others are interpolations.

The accuracy distribution and the confusion matrix of the blind
test are shown in Figs. 5 and 6.

Fig. 5. Accuracy distribution for different interpolating data blind sets.
The red bars show the correct objects, while the yellow bars show the
incorrect ones. The numbers are the accuracies.

Fig. 6. Confusion matrix for interpolating the blind test, and the accu-

racy is 96.5%. The colors are the same as in Fig. 3.
Table 4. Parameters for interpolating the blind test.

| Parameters | STAR | GALAXY | QSO |
|------------|------|--------|-----|
| Recall     | 99.4%| 88.5%  | 76.9%|
| Precision  | 96.1%| 97.7%  | 1   |
| $F_1$-score| 95.0%| 92.9%  | 87.0%|

Table 5. Parameters for extrapolating the blind test.

| Parameters | STAR | GALAXY | QSO |
|------------|------|--------|-----|
| Recall     | 99.0%| 61.0%  | 68.3%|
| Precision  | 83.3%| 90.4%  | 1   |
| $F_1$-score| 90.5%| 72.9%  | 81.2%|

We used the UVQS catalog (Monroe et al. 2016) for QSO-blind testing and obtained 34 objects after cross-matching with J-PLUS. There are 18 objects that have fallen into the extrapolating region. UVQS contains UV bright QSOs, while the observation wavelength of VV13 is mainly in optical bands. This difference may cause a bias between training and testing, and further result in misclassifications.

The blind test set was constructed by the independent objects from the four catalogs. We then separated the testing data into the interpolation and extrapolation samples.

We also adopted some other parameters to describe the classifier: recall, precision, and $F_1$-score. We first define true positives (TPs), false positives (FPs), and false negatives (FNs) to demonstrate these parameters. TP is the number that both the
Table 6. J-PLUS classification.

| ID       | RA         | Dec      | Class_star | PredictClass | Probability |
|----------|------------|----------|------------|--------------|-------------|
| 26016–5  | 255.46835  | 22.76001 | 0.9990     | STAR         | 92.96%      |
| 26016–15 | 255.30506  | 22.76083 | 0.9990     | STAR         | 99.91%      |
| 26016–16 | 255.36753  | 22.76113 | 0.9990     | STAR         | 99.20%      |
| 26016–22 | 255.27928  | 22.76202 | 0.9536     | GALAXY       | 79.71%      |
| 26016–29 | 255.45243  | 22.76157 | 0.9814     | GALAXY       | 76.31%      |
| 26016–33 | 255.39194  | 22.76157 | 0.9780     | GALAXY       | 93.95%      |
| 26016–50 | 255.16394  | 22.76439 | 0.0460     | STAR         | 64.35%      |
| 26016–55 | 255.66306  | 22.76168 | 1.0000     | STAR         | 99.49%      |
| 26016–63 | 254.86508  | 22.76506 | 0.7656     | QSO          | 86.82%      |
| 26016–64 | 254.75625  | 22.76496 | 0.9663     | STAR         | 84.58%      |

Notes. This table contains the top ten objects located inside the 12 contours. ID is the object identity from J-PLUS (http://archive.cefca.es/catalogues/jplus-dr1/navigator.html). RA and Dec are the right ascension and declination of the objects. The class_star column comes from the J-PLUS catalog, denoting the probability of stars. The PredictClass column presents our classification. The last column provides the probabilities of the predicted class. The sum of the probabilities for the three classes is equal to 100%. In this work, we developed a SVM algorithm with a one-versus-one strategy and a Gaussian kernel equal to 0.75. The table is uploaded with the paper.

blind test labels and the predicted labels are positive. FP is the number that blind test labels are negative while the predicted labels are positive, FN is the number that blind test labels are positive, while the predicted labels are negative. Readers should recall that \( \frac{TP}{TP+FP} \) shows the fraction of right prediction for a label. Precision = \( \frac{TP}{TP+FP} \) shows the fraction of right prediction, and \( F_1 \)-score = \( \frac{2TP}{2TP+FP+FN} \) shows the harmonic mean of the precision and the recall.

The total accuracy is 96.5% for the interpolating sample (Figs. 5 and 6), and the parameters are shown in Table 4. We present the accuracy distribution corresponding to the magnitudes as well (Fig. 9). See more in Appendix B. The blind test indicates a high reliability of the classifier. For the rest of the sample, the total accuracy is 79.1% (Figs. 7 and 8), which is much lower than the interpolating sample, and the test parameters are shown in Table 5. This indicates that it is significant and effective to constrain extrapolation in prediction.

Fig. 10. Color-color diagram of GALAXYs. Left panel: sample, and right panel: interpolation set. The color is the density of the sample, with a color bin of 0.01 mag$^2$.

Fig. 11. Color-color diagram of QSOs (similar to Fig. 10).

Fig. 12. Color-color diagram of STARs (similar to Fig. 10).
In order to find the abnormal objects from the ambiguous samples, we calculated the Mahalanobis distance (De Maesschalck et al. 2000; Mahalanobis 1936). We then checked whether the objects were far from each label. The objects that have a higher distance to one label than the distance of this label to the other labels were treated as abnormal objects. These objects are not only located outside the region of three classes, but they are also far from all of them. The criteria of the Mahalanobis distance are as follows: 18.4 between STAR to GALAXY, 23.3 between GALAXY to QSO, and 50.3 between QSO to STAR. Table 10 presents 26 abnormal objects.

### 4.3. Comparison with CLASS_STAR

In Fig. 13, we draw the difference between our results and the CLASS_STAR in J-PLUS catalog. The figure indicates that there are differences between the J-PLUS CLASS_STAR and our result. The difference may be caused by the different strategies of classifying the objects: binary classification for J-PLUS based on the point-source detection and our triple classification based on machine learning. The QSOs are probably not distinguished from stars or galaxies with the point-source detection, and such a detection could further result in the difference in the figure. Therefore, the factor CLASS_STAR may not be suitable enough for multi-classifications.

### 5. Discussion

#### 5.1. Different ways to constrain extrapolation

We constructed three methods to constrain extrapolation, including magnitude cuts and two density-dependence methods. The most straightforward thought is defining intervals based on magnitude distributions of our training sample. We can determine whether an object belongs to the intersection of these intervals or not.

We employed kernel distribution (Bowman & Azzalini 1997) to fit the distributions of each dimension in the instance space. The kernel distribution is a kind of probability measure. For each dimension, the objects situated in the middle part of the distribution are defined as interpolations. By cutting down 0.025 for each side of a magnitude distribution, the intervals were constructed, and the objects situated in the middle part of the distribution are defined as interpolations. By cutting down 0.025 for each side of a magnitude distribution, the intervals were constructed, and the objects could be separated into interpolation or extrapolation. This method results in an accuracy of 95.5% for the blind test. After the selection, 2,749,840 interpolating objects

---

**Table 7. Extrapolation objects.**

| ID     | RA      | Dec     | Class_star | PredictClass | Probability |
|--------|---------|---------|------------|--------------|-------------|
| 26016–2 | 255.64129 | 22.73862 | 0.9213     | STAR         | 58.49%      |
| 26016–3 | 255.54013 | 22.75905 | 0.9345     | STAR         | 60.68%      |
| 26016–9 | 255.39760 | 22.76004 | 0.4218     | QSO          | 64.53%      |
| 26016–10 | 255.20027 | 22.76122 | 0.9497     | STAR         | 64.68%      |
| 26016–11 | 255.40765 | 22.76091 | 0.9711     | STAR         | 54.15%      |
| 26016–13 | 255.38474 | 22.76107 | 0.9604     | STAR         | 96.74%      |
| 26016–14 | 255.22779 | 22.76128 | 0.9975     | QSO          | 89.82%      |
| 26016–23 | 255.76778 | 22.76090 | 0.9379     | QSO          | 95.45%      |
| 26016–34 | 255.61690 | 22.76251 | 0.6918     | GALAXY       | 46.79%      |
| 26016–37 | 255.37846 | 22.76318 | 0.9785     | STAR         | 95.16%      |

**Notes.** This table contains the top ten objects located outside the 12 contours. The column labels for this table are the same as in Table 6. The table is uploaded with the paper.

**Table 8. Object numbers and criteria.**

| Criterion | Object number |
|----------|---------------|
| <0.5     | 117,952       |
| <0.4     | 15,583        |
| <0.35    | 943           |
| <0.34    | 155           |

**Notes.** The first column is the upper limit for the highest probability in three classes. The objects were taken from both Tables 6 and 7.
Table 9. Ambiguous objects.

| ID       | RA      | Dec     | Class_star | PredictClass | GALAXY | QSO   | STAR   |
|----------|---------|---------|------------|--------------|--------|-------|--------|
| 26016–20835  | 255.17392  | 23.50692  | 0.0008     | GALAXY        | 33.50% | 32.81% | 33.68% |
| 26016–32840  | 255.67090  | 24.07627  | 0.5004     | STAR          | 33.38% | 32.90% | 33.70% |
| 26015–6320   | 256.61032  | 23.07467  | 0.4304     | STAR          | 32.54% | 33.51% | 33.94% |
| 26010–33609  | 257.27271  | 25.44612  | 0.0078     | GALAXY        | 33.71% | 33.09% | 33.19% |
| 26012–8040   | 256.82760  | 25.80270  | 0.9121     | GALAXY        | 33.95% | 32.38% | 33.65% |
| 26012–24063  | 256.34766  | 26.30697  | 0.4980     | GALAXY        | 33.25% | 33.40% | 33.33% |
| 26028–3296   | 126.56224  | 29.97372  | 0.8632     | STAR          | 32.39% | 33.63% | 33.97% |
| 26037–3692   | 139.61591  | 29.96809  | 0.0011     | GALAXY        | 33.45% | 33.74% | 32.80% |
| 26038–17239  | 144.43648  | 30.72382  | 0.0027     | QSO           | 33.22% | 33.51% | 33.25% |
| 26036–8820   | 146.18454  | 30.19118  | 0.0050     | GALAXY        | 33.45% | 33.22% | 32.82% |

Notes. This table contains the top ten ambiguous objects for criterion 0.34. The first five column labels of this table are the same as in Table 6. The last three column labels provide the probability of the corresponding label. The table is uploaded with the paper.

Table 10. Abnormal objects.

| ID       | RA      | Dec     | CLASS_STAR | PredictClass | GALAXY | Gdis | QSO   | Qdis | STAR   | Sdis |
|----------|---------|---------|------------|--------------|--------|------|-------|------|--------|------|
| 26016–32840  | 255.67090  | 24.07627  | 0.5005     | STAR         | 33.39% | 45.85 | 32.90% | 59.90 | 33.19% | 214.59 |
| 26047–19730  | 121.32316  | 31.91350  | 0.0001     | GALAXY       | 32.76% | 78.09 | 33.94% | 51.34 | 33.30% | 204.47 |
| 26091–3622   | 128.45054  | 34.12899  | 0.4170     | GALAXY       | 32.93% | 201.06 | 33.70% | 61.91 | 33.59% | 224.79 |
| 33209–2012   | 137.46224  | 39.60661  | 0.1433     | GALAXY       | 33.52% | 55.05 | 32.89% | 67.51 | 33.59% | 214.59 |
| 26141–18524  | 169.58419  | 40.46721  | 0.0016     | STAR         | 33.52% | 73.44 | 32.63% | 52.22 | 33.85% | 185.80 |
| 26145–15610  | 284.78880  | 39.72198  | 0.2306     | GALAXY       | 33.87% | 133.31 | 32.82% | 50.71 | 33.32% | 278.23 |
| 33232–7122   | 126.08110  | 41.29631  | 0.1117     | GALAXY       | 33.60% | 98.28 | 33.32% | 67.51 | 33.08% | 251.50 |
| 26151–29824  | 273.32033  | 41.61314  | 0.0134     | GALAXY       | 33.74% | 229.37 | 32.77% | 56.60 | 33.49% | 255.62 |
| 26207–8959   | 137.26562  | 52.62772  | 0.0027     | QSO          | 32.94% | 66.90 | 33.75% | 59.96 | 33.31% | 312.61 |
| 26241–25002  | 137.26562  | 52.62772  | 0.5029     | GALAXY       | 33.31% | 129.51 | 33.41% | 70.59 | 33.28% | 278.00 |

Notes. This table contains the top ten abnormal objects. The column labels of this table are the same as in Table 9. The column Gdis, Qdis, and Sdis provide the Mahalanobis distance to the group GALAXY, QSO, and STAR. The table is uploaded with the paper.

were left, and there was 65.79% of the J-PLUS catalog (Figs. 14 and 15). This method is precluded due to its low accuracy and its unrepresentative of the interpolation boundary.

The ideal approach is to draw a 12-dimension density contour to select the interpolating sample. The adopted method (Sect. 3.3.1) is an approximation of such an ideal approach. The last method is four contours instead of 12 contours, which are (mag1, mag2, mag3), (mag4, mag5, mag6), (mag7, mag8, mag9), and (mag10, mag11, mag12). These rough contours result in an accuracy of 96.1%, and they left 3,702,268 interpolating objects (Figs. 16 and 17). This method is also precluded due to its low accuracy.

5.2. Comparison of different classifiers

Bai et al. (2018) used a RF algorithm to gain a classifier with an accuracy of 99%. We also tested RF, but its accuracy is lower than SVM. The different results of these two works are probably due to the different sample sizes and wavebands. The accuracy of the blind test is similar to the training accuracy, implying that there is no obvious overfitting in our training process (Shalev-Shwartz & Ben-David 2014).

The sample size may also influence the training accuracy. In our method, the sample size is 468,685, while in Bai et al. (2018), the number is 2,973,855. Both SVM or RF have a finite Vapnik-Chervonenkis dimension (VCdim; Shalev-Shwartz & Ben-David 2014). If a sample size goes to infinity, the training error and the validation error converge to the approximation error. This implies that there exists a limited accuracy of a classifier. In our work, the training error (97%) is similar to the validation error (96.5%). Therefore, if we enlarge the sample size, the accuracy may not increase significantly.

Bai et al. (2018) applied nine-dimensional color spaces including infrared bands, while we used 12 optical magnitudes. More and broader bands involved in the training would lead to a higher total accuracy. The accuracy in our work is slightly lower.
This is probably due to the strong correlation in the 12 bands. We calculated a correlation matrix (Fig. 18) for these 12 bands from their photometric results.

The minimum of the correlation coefficient stands at (5,4), mag4 (J0410), and mag5 (J0430) in Fig. 18. These high correlation values indicate that all of these wavebands are highly correlated. The correlation may be explained by not only the distance of the object that causes a similarity in all magnitudes, but also the overlapping of filter profiles. From the band plot in J-PLUS\(^6\), the filter profile of \(u, g, r, i, z\) have overlapped with other narrow bands, implying that they are not strongly independent. The wavebands adopted in Bai et al. (2018) cover a larger range, and the correlations are probably weaker.

The constitution of the data set may also influence the accuracy of a classifier. In Ball et al. (2006), a tree algorithm was developed to output the probability of a star, galaxy, and nsng (neither star nor galaxy object). In Bai’s and Ball’s training samples, there was a significant bias in the sample set. The sample construction of our SVM classifier and the two mentioned classifiers is shown in Table 11. Bai et al. (2018) and Ball et al. (2006) concluded that the biased sample can also present a training accuracy of better than 95%.

### 5.3. Future work

The advantages of J-PLUS are 12 optical filters and a large amount of data. The ongoing J-PAS has an all-time system of 56 optical narrowband filters, making it one of the most promising surveys in the world. The way we work on J-PLUS can be copied to J-PAS. The more bands applied, the more precise a classifier could be.

Baqui et al. (2021) developed different classifiers to label the mini-JPAS (Bonoli et al. 2021), including RF and Extremely Randomized Trees (ERT). MiniJ-PAS is a previous project to...
Fig. 18. Correlation matrix of the 12 bands. The numbers in the axes imply the bands, e.g., 1 corresponds to mag1. The 12 mag are mag1 (r), mag2 (J0378), mag3 (J0395), mag4 (J0410), mag5 (J0430), mag6 (g), mag7 (J0515), mag8 (r), mag9 (J0660), mag10 (i), mag11 (J0860), and mag12 (z).

Table 11. Constitution of different algorithms.

| Class          | This work | Bai et al. (2018) | Ball et al. (2006) |
|----------------|-----------|-------------------|-------------------|
| Galaxy        | 15.94%    | 27.11%            | 75.77%            |
| QSO(nsng)     | 9.79%     | 1.47%             | 11.16%            |
| Star          | 74.27%    | 71.42%            | 13.07%            |

Notes. In Bai’s paper, the second row is QSO, while in Ball’s work, it is nsng.

test J-PAS. Their work has gained good performance with Area Under the Curve (AUC) greater than 0.95 in different classifiers. AUC is equal to the positive probability.

SVM is inferior when the instance space has too many dimensions, or when the data set is too large for calculation. More works are required to test the time cost of SVM when we apply larger data with more features. Although SVM is a good algorithm; its performance in J-PAS still needs to be tested considering the computational complexity and no-free-lunch theorem.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (NSFC) through grants NSFC-11988101/11973054/11933004 and the National Programs on Key Research and Development Project (grant No. 2019YFA0405504 and 2019YFA0405000), Strategic Priority Program of the Chinese Academy of Sciences under grant number XDB41000000. H.B. Yuan acknowledges funding from National Natural Science Foundation of China (NSFC) through grants NSFC-12173007. E.L.M. acknowledges support from the Agencia Estatal de Investigación del Ministerio de Ciencia e Innovación (AEI-MCINN) under grant PID2019-109522GB-C33. Based on observations made with the JAST80 telescope at the Observatorio Astrofísico de Javalambre (OAJ), in Teruel, owned, managed, and operated by the Centro de Estudios de Física del Cosmos de Aragón (CEFCA). We acknowledge the OAJ Data Processing and Archiving Unit (UPAD) for reducing the OAJ data used in this work. Funding for the J-PLUS Project has been provided by the Governments of Spain and Aragón through the Fondo de Inversiones de Teruel; the Aragón Government through the Research Groups E96, E103, and E16_17R; the Spanish Ministry of Science, Innovation and Universities (MICU/AEI/FCEDER, UE) with grants PGC2018-097585-B-C21 and PGC2018-097585-B-C22; the Spanish Ministry of Economy and Competitiveness (MINECO) under AYA2015-66211-C2-1-P, AYA2015-66211-C2-2, AYA2015-30789, and ICTS-2009-14; and European FEDER funding (FCDD10-4E-867, FCDID13-4E-2685). The Brazilian agencies FINEP, Fapesp, and the National Observatory of Brazil have also contributed to this project. Guoshoujing 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. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the US 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 http://www.sdss.org/. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chelan Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofisica 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-Plank-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestre Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatario 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. This work is supported by the CSST project on “stellar activity and late evolutionary stage”.
References

Ahumada, R., Allende Prieto, C., Almeida, A., et al. 2020, ApJS, 249, 3
Bai, Y., Liu, J., Wang, S., & Yang, F. 2018, AJ, 157, 9
Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. 2006, ApJ, 650, 497
Baqui, P. O., Marra, V., Casarini, L., et al. 2021, A&A, 645, A87
Benítez, N., Dupke, R., Moles, M., et al. 2014, ArXiv e-prints [arXiv:1403.5237]
Bertin, E., & Arnouts, S. 1996, A&SAS, 117, 393
Bonoli, S., Marín-Franch, A., Varela, J., et al. 2021, A&A, 653, A31
Boser, B. E., Guyon, I. M., & Vapnik, V. N. 1992, Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT ’92 (New York: Association for Computing Machinery), 144
Bowman, A. W., & Azzalini, A. 1997, Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations (Oxford: Oxford University Press), 18
Breiman, L. 2001, Stat. Sci., 16, 199
Cenarro, A. J., Moles, M., Marín-Franch, A., et al. 2014, Proc. SPIE, 9149, 9149I
Cenarro, A. J., Moles, M., Cristóbal-Hornillos, D., et al. 2019, A&A, 622, A176
Cortes, C., & Vapnik, V. 1995, Mach. Learn., 20, 273
Cover, T., & Hart, P. 1967, IEEE Trans. Inf. Theory, 13, 21
Cristianini, N., & Shawe-Taylor, J. 2000, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods (Cambridge: Cambridge University Press)
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, Res. Astron. Astrophys., 12, 1197
De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. 2000, Chemom. Intell. Lab. Syst., 50, 1
Deng, L.-C., Newberg, H. J., Liu, C., et al. 2012, Res. Astron. Astrophys., 12, 735
Fisher, R. A. 1936, Ann. Eugen., 7, 179
Freund, Y., & Schapire, R. E. 1997, J. Comput. Syst. Sci., 55, 119
Wang, S.-G., Su, D.-Q., Chu, Y.-Q., Cui, X., & Wang, Y.-N. 1996, Appl. Opt., 35, 5155
Huchra, J. P., Macri, L. M., Masters, K. L., et al. 2012, ApJS, 199, 26
Jiménez-Teja, Y., Dupke, R. A., Lopes de Oliveira, R., et al. 2019, A&A, 622, A183
López-Sanjuan, C., Vázquez Ramió, H., Varela, J., et al. 2019, A&A, 622, A177
Luo, A.-L., Zhang, H.-T., Zhao, Y.-H., et al. 2012, Res. Astron. Astrophys., 12, 1243
Mahalanobis, P. C. 1936, Proc. Natl. Inst. Sci., 2, 49
Marín-Franch, A., Taylor, K., Cenarro, J., Cristóbal-Hornillos, D., & Moles, M. 2015, IAU Gen. Assem., 29, 2257381
Monroe, T. R., Prochaska, J. X., Tejos, N., et al. 2016, AJ, 152, 25
Nogueira-Cavalante, J. P., Dupke, R., Coelho, P., et al. 2019, A&A, 630, A88
Quinlan, J. R. 1986, Mach. Learn., 1, 81
Shalev-Shwartz, S., & Ben-David, S. 2014, Understanding Machine Learning: From Theory to Algorithms (Cambridge: Cambridge University Press)
Steinmetz, M., Matijevič, G., Enke, H., et al. 2020, AJ, 160, 82
Stone, C. J. 1977, Ann. Stat., 5, 595
Su, D.-Q., & Cui, X.-Q. 2004, Chin. J. Astron. Astrophys., 4, 1
Taylor, M. B. 2005, in Astronomical Data Analysis Software and Systems XIV, eds. P. Shopbell, M. Britton, & R. Ebert, ASP Conf. Ser., 347, 29
Véron-Cetty, M. P., & Véron, P. 2010, A&A, 518, A10
Whitten, D. D., Placco, V. M., Beers, T. C., et al. 2019, A&A, 622, A182
Yuan, H., Liu, X., Xiang, M., et al. 2015, ApJ, 799, 133
Zasowski, G., Johnson, J. A., Frinchaboy, P. M., et al. 2013, AJ, 146, 81
Zhao, G., Zhao, Y.-H., Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, Res. Astron. Astrophys., 12, 723
Appendix A: Density contours of the sample set

We present all 12 three-dimensional contours of the predictions.

Fig. A.1. First two contours for extrapolation constraining.
Fig. A.2. Third and fourth contour for extrapolation constraining.
Fig. A.3. Fifth and sixth contour for extrapolation constraining.
Fig. A.4. Seventh and eighth contour for extrapolation constraining.
Fig. A.5. Ninth and tenth contour for extrapolation constraining.
Fig. A.6. Last two contours for extrapolation constraining.
Appendix B: Magnitude distributions

We present the magnitude distributions for each class, magnitude, and for both samples and interpolations. The red line indicates STAR, the green is for GALAXY, and the blue is for QSO.

Fig. B.1. Magnitude distributions for the interpolation objects. The STARs are red, the GALAXYs are green, and the QSOs are blue. The x-axis shows the magnitude, and the y-axis shows the probability.
Fig. B.2. Magnitude distribution for sample objects. The axes and line colors are the same as the interpolations.
Fig. B.3. Magnitude distributions in the g-band of our training. The top panels show the sample set and blind test set from left to right. The middle panels show the interpolation and extrapolation objects, and the bottom panel presents the J-PLUS catalog distribution.
Fig. B.4. Magnitude distributions in the g-band of each label. The three panels show the label STAR, GALAXY, and QSO from the top left to the bottom, respectively.
Appendix C: Sample of our training

We present the training sample in Table C.1, and the subclasses of STARs are included. The overlap of the stars in the sample is presented in Table C.2. The overlap of galaxies between SDSS DR16 and LAMOST DR7 is 9,871. The QSO overlaps between each catalog are 4,593 for VV13 and SDSS DR16, 1,339 for VV13 and LAMOST DR7, and 3,802 for SDSS DR16 and LAMOST DR7. Though the overlapping is so enormous for the LAMOST catalogs, there is still one independent object in the A-, F-, G-, and K-type star catalog. This catalog can provide more information. Also, there are 6,147 and 108 independent objects in APOGEE and VV13.

Table C.1. Sample set.

| ID       | R.A.  | Dec.  | class | subclass | catalog         |
|----------|-------|-------|-------|----------|-----------------|
| 25998-16 | 117.49152 | 39.45784 | STAR  | G8       | LAMOST M        |
| 25998-143| 117.58813 | 39.46421 | STAR  | M3       | LAMOST A-F-G-K- |
| 25998-309| 116.08787 | 39.47680 | STAR  | F7       | LAMOST          |
| 25998-495| 117.13277 | 39.48703 | QSO   |          | SDSS            |
| 25998-942| 116.99474 | 39.50345 | GALAXY|          | SDSS            |
| 25998-8981| 116.47530 | 39.83560 | STAR  | G2       | LAMOST M        |
| 25998-9909| 116.52973 | 39.82853 | STAR  |          | APOGEE          |
| 26036-5884| 145.42144 | 30.01548 | STAR  | A2V      | LAMOST A-F-G-K- |
| 26025-6501| 125.43704 | 30.12984 | QSO   |          | VV13            |
| 26025-7265| 124.35765 | 30.17180 | STAR  |          | SDSS            |

Notes. The first four columns are the same as in Table 6. The “subclass” is labeled from the LAMOST DR7 catalog, indicating the subclass of stars. The blank in the subclass means that the subclass is missing or it is not a star. The column “catalog” shows the origin catalog, where SDSS means SDSS DR16 and LAMOST means LAMOST DR7.

Table C.2. Sample overlap of STARs.

| Catalog | APOGEE | LAMOST | AFGK | LA  | LM  |
|---------|--------|--------|------|-----|-----|
| SDSS    | 91     | 7,266  | 3,993| 960 | 333 |
| APOGEE  | 7,564  | 6,419  | 69   | 294 |     |
| LAMOST  | 212,114| 5,145  | 2,5604|     |     |
| AFGK    | 1,408  | 421    |      |     |     |
| LA      | 6      |        |      |     |     |

Notes. The overlap of each catalog in the class STAR. SDSS stands for SDSS DR16, and LAMOST stands for LAMOST DR7. AFGK, LA, and LM are the LAMOST A-, F-, G-, and K-type star, LAMOST A-star, and LAMOST M-star catalog, respectively.