Machine learning classification of \textit{Gaia} Data Release 2

Yu Bai\textsuperscript{1}, Ji-Feng Liu\textsuperscript{1,2} and Song Wang\textsuperscript{1}

\textsuperscript{1} Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China; ybai@nao.cas.cn
\textsuperscript{2} College of Astronomy and Space Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

Received 2018 July 13; accepted 2018 August 16

\textbf{Abstract} Machine learning has increasingly gained more popularity with its incredibly powerful ability to make predictions or calculate suggestions for large amounts of data. We apply machine learning classification to 85,613,922 objects in the \textit{Gaia} Data Release 2, based on a combination of Pan-STARRS 1 and AllWISE data. The classification results are cross-matched with the Simbad database, and the total accuracy is 91.9\%. Our sample is dominated by stars, \(\sim98\%\), and galaxies make up 2\%. For the objects with negative parallaxes, about 2.5\% are galaxies and QSOs, while about 99.9\% are stars if the relative parallax uncertainties are smaller than 0.2. Our result implies that using the threshold of \(0 < \sigma_\pi / \pi < 0.2\) could yield a very clean stellar sample.

\textbf{Key words:} methods: data analysis — stars: general — \textit{Gaia} catalog

1 INTRODUCTION

The ESA space mission \textit{Gaia} is performing an all-sky astrometric, photometric and radial velocity survey at optical wavelengths (Gaia Collaboration et al. 2016). The primary objective of the \textit{Gaia} mission is to survey more than one billion stars, in order to investigate the origin and subsequent evolution of our Galaxy. Its second data release (\textit{Gaia} DR2; Gaia Collaboration et al. 2018) includes \~\textasciitilde1.3 billion objects with valid parallaxes. These parallaxes are obtained with a complex iterative procedure, involving various assumptions (Lindegren et al. 2012). Such procedure may produce parallaxes for galaxies and quasi-stellar objects (QSOs), which should present no significant parallaxes (Liao et al. 2018).

In addition, \textit{Gaia} uses two fields of view to observe, and this, in principle, might lead to a global parallax bias (van Leeuwen 2005; Butkevich et al. 2017; Liao et al. 2018). Separating galaxies and QSOs from stars allows us to characterize parallax bias in the \textit{Gaia} catalog, and to provide a clean and accurate stellar sample for further investigation. Traditionally, the classification of objects involves magnitude and color criteria, but the criteria become too complex to be described with functions in a multidimensional parameter space. By contrast, this parameter space can be effectively explored with machine learning (ML) algorithms, which have aided astronomers in dealing with complex problems in modern astrophysics (Huertas-Company et al. 2008, 2009; Manteiga et al. 2009; Bai et al. 2018a; Pashchenko et al. 2018).

ML provides us with an alternative option to classify billions of objects that cannot be followed-up spectroscopically. Bai et al. (2018a) applied supervised ML to star/galaxy/QSO classification based on the combination of SDSS and LAMOST spectral surveys (the Sloan-LAMOST (SL) classifier). Actually, the class labels of the training objects are from spectroscopy and are regarded as true. Narrow line QSOs are classified as galaxies by both the SDSS and LAMOST pipelines because the template of a QSO in the pipelines is a theoretical one with broad emission lines. A classifier built with the random forest algorithm generated the best performance in terms of time cost and total accuracy. Several blind tests were also performed on objects observed by RAVE,
6dFGS and UVQs. Accuracies were higher than 99% for the stars and galaxies, and higher than 94% for the QSOs. In this paper, we apply the SL classifier to Gaia DR2 to investigate potential extragalactic objects. The data and classification are described in Section 2. Section 3 gives the result and analysis, and a summary is presented in Section 4.

2 DATA AND CLASSIFICATION

In order to use the SL classifier, we build a nine-dimensional color space that includes $g - r$, $r - i$, $i - J$, $J - H$, $H - K$, $K - W_1$, $W_1 - W_2$, $w_1 mag_1 - w_1 mag_3$ and $w_2 mag_1 - w_2 mag_3$ (Bai et al. 2018a). The optical colors are extracted from Data Release 1 of the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS 1; hereafter PS1) archive data. PS1 has carried out a set of imaging sky surveys including the 3r Steradian Survey, in which the mean 5σ point source limiting sensitivities are 23.3, 23.2 and 23.1 mag in $g$, $r$, $i$ bands, respectively (Chambers et al. 2016). We cross-matched the Gaia DR2 with PS1 using $\text{panstarrs1}_\text{best}_\text{neighbour}$, the pre-computed PS1 cross-match table provided in the Gaia archive (Marrese et al. 2017). The table includes 810 359 898 entries, which represent the most likely matches between PS1 and Gaia DR2, as determined by angular distances, position errors, epoch differences and density of sources in PS1.

In order to obtain infrared colors, we cross-matched the Gaia DR2 with the AllWISE catalog using $\text{allwise}_\text{best}_\text{neighbour}$, which includes 300 207 917 matches (Marrese et al. 2017). Here we select objects with signal-to-noise ratios higher than 2 in the $W_1$ and $W_2$ bands. As a result, the cross-matchings yield 85 613 922 objects with the nine colors. We feed the nine-color matrix to the SL classifier, and the classifier returns the types and the probabilities ($P$) for stars, galaxies and QSOs. The sum of $P$ for the three types is 100%, and the type with the highest $P$ is adopted by the SL classifier as the output type. Therefore, the $P$ of the adopted type is higher than 33.33%.

Traditionally, QSOs are separated from other AGNs mainly by their absolute $B$ magnitudes. The QSOs in training data of the SL classifier are identified with the QSO spectral templates. The different definitions of QSOs may cause many galaxies in our sample to be classified as QSOs in literatures. Therefore, the galaxies and QSOs given by the SL classifier are combined and hereafter called galaxies. The results include 83 891 260 stars and 1 722 662 galaxies.

3 RESULT AND ANALYSIS

3.1 Comparison with Simbad

We cross-match these objects with the Simbad database in order to estimate the probability of incorrect classifications. The Simbad database returns 308 864 galaxies, 191 497 stars and 10 987 unclassified objects.

The distributions of the output probabilities are presented in Figure 1. We defined the accuracies as the ratios between the numbers of Simbad types and those given by the SL classifier. The total accuracy is 91.9%. More than 99.1% of the galaxies in our sample are also classified as galaxies in the Simbad database, and more than 83% of the stars in our sample are classified as stars.

The classification accuracy of the stars is lower than those of the spectrally resolved samples in Bai et al. (2018a). Stars in the training sample of the SL classifier are mainly from LAMOST, which is dominated by stars located near the Galactic Anticenter. This selection effect may make the SL classifier perform better when applied to lightly reddened objects. Objects located in the heavily reddened direction of the Galaxy are probably hard to recognize by the SL classifier.

3.2 Sky Distribution

We present distributions of the classification results in Figure 2. It is expected that the Galactic plane is dominated by stars, and the percentages of galaxies become higher at high latitudes. The relatively high percentages in the most central Galactic plane may be due to the low density of stars in this region (left panel in Fig. 2). The low completeness of PS1 caused by high extinction (Chambers et al. 2016) may result in the low density of stars in the most central Galactic plane. Additionally, WISE photometry is limited by confusion near the Galactic plane due to high source density (Wright et al. 2010). In the distribution of galaxies, we can find overdense areas corresponding to some galaxy clusters (Jarrett 2004), e.g., Abell 624, Perseus-Pisces Supercluster and Shapley Concentration.

3.3 Relative Error of Parallax

The relative parallax uncertainty, $\sigma_\pi/\pi$, is an important parameter that can be used to constrain bias caused by
Fig. 1 The ML probability distributions for galaxies (left panel) and stars (right panel). The red lines are the numbers of Simbad classifications higher than the corresponding probabilities. The y axis is displayed on a log scale for clarity. The blue lines are the accuracies compared to Simbad.

Fig. 2 The distributions of classification results in Galactic coordinates. Upper left panel: the percentage of galaxies per degree$^2$. Upper right panel: the density of stars. Lower panel: the density of galaxies.

Fig. 3 The stacked distributions of $\sigma_{\pi}/\pi$. Left panel: the distribution in the range between –2 and 3. Right panel: the distribution in the range between 0 and 1.
the Lutz–Kelker Effect (LKE; Trumpler & Weaver 1953; Lutz & Kelker 1973; Bai et al. 2018b). We present the stacked distributions of $\sigma_{\pi}/\pi$ in Figure 3. The sample of galaxies makes up $\sim$2.5% of objects with parallaxes less than zero. The percentages of stars decrease sharply in the range $\sim 0.6 < \sigma_{\pi}/\pi < 0.0$, and reach a minimum of 96.7%. Since there is no negative uncertainty, the negative $\sigma_{\pi}/\pi$ means negative parallax.

The percentages of stars decline with the increase of $\sigma_{\pi}/\pi$ for objects with positive parallaxes. The galaxies make up less than 1% of objects in the range $0 < \sigma_{\pi}/\pi < 0.6$. The sample of galaxies makes up $\sim 2.5\%$ of objects with parallaxes less than zero. The percentages of stars decrease sharply in the range $\sim 0.6 < \sigma_{\pi}/\pi < 0.0$, and reach a minimum of 96.7%. Since there is no negative uncertainty, the negative $\sigma_{\pi}/\pi$ means negative parallax.

The percentages of stars decline with the increase of $\sigma_{\pi}/\pi$ for objects with positive parallaxes. The galaxies make up less than 1% of objects in the range $0 < \sigma_{\pi}/\pi < 0.6$. The sample of galaxies makes up $\sim 2.5\%$ of objects with parallaxes less than zero. The percentages of stars decrease sharply in the range $\sim 0.6 < \sigma_{\pi}/\pi < 0.0$, and reach a minimum of 96.7%. Since there is no negative uncertainty, the negative $\sigma_{\pi}/\pi$ means negative parallax.

The percentages of stars decline with the increase of $\sigma_{\pi}/\pi$ for objects with positive parallaxes. The galaxies make up less than 1% of objects in the range $0 < \sigma_{\pi}/\pi < 0.6$. The sample of galaxies makes up $\sim 2.5\%$ of objects with parallaxes less than zero. The percentages of stars decrease sharply in the range $\sim 0.6 < \sigma_{\pi}/\pi < 0.0$, and reach a minimum of 96.7%. Since there is no negative uncertainty, the negative $\sigma_{\pi}/\pi$ means negative parallax.

4 SUMMARY

We apply the SL classifier to 85 613 922 objects in Gaia DR2, based on colors built from the PS1 and AllWISE. The classification shows that about 98% of the sample are stars and 2% are galaxies. This result is cross-matched with the Simbad database in order to estimate the probability of incorrect classifications, and the total accuracy is 91.9%. The Galactic plane is dominated by stars, and the percentages become higher at high latitudes. We find that about 2.5% of the sample are galaxies for objects with negative parallaxes and the threshold of $0 < \sigma_{\pi}/\pi < 0.2$ could yield a very clean stellar sample including about 99.9% stars.

Acknowledgements This work was supported by the National Program on Key Research and Development Project (Grant No. 2016YFA0400804) and the National Natural Science Foundation of China (Grant Nos. 11603038, 11333004, 11425313 and 11403056). Some of the data presented in this paper were obtained from the Mikulski Archive for Space Telescopes (MAST). STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. Support for MAST for non-HST data is provided by the NASA Office of Space Science via grant NNX09AF08G and by other grants and contracts.

This work has made use of data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

References

Bai, Y., Justham, S., Liu, J.F., et al. 2018a, in preparation
Bai, Y., Liu, J., & Wang, S. 2018b, arXiv:1807.00531
Butkevich, A. G., Klioner, S. A., Lindegren, L., Hobbs, D., & van Leeuwen, F. 2017, A&A, 603, A45
Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv:1612.05560
Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, arXiv:1804.09365
Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., & Le Fèvre, O. 2008, A&A, 478, 971
Huertas-Company, M., Tasca, L., Rouan, D., et al. 2009, A&A, 497, 743
Jarrett, T. 2004, PASA, 21, 396
Liao, S., Lattanzi, M. G., Vecchiato, A., et al. 2018, Proceedings of the International Astronomical Union, S330, 231
Liao, S.-I., Qi, Z.-x., Guo, S.-f., & Cao, Z.-h. 2018, arXiv:1804.08821
Lindegren, L., Lammers, U., Hobbs, D., et al. 2012, A&A, 538, A78
Lutz, T. E., & Kelker, D. H. 1973, PASP, 85, 573
Manteiga, M., Carricajo, I., Rodríguez, A., Dafonte, C., & Arcay, B. 2009, AJ, 137, 3245
Marrese, P. M., Marinoni, S., Fabrizio, M., & Giuffrida, G. 2017, A&A, 607, A105
Pashchenko, I. N., Sokolovsky, K. V., & Gavras, P. 2018, MNRAS, 475, 2326
Trumpler, R. J., & Weaver, H. F. 1953, Statistical astronomy (New York: Dover Publications)
van Leeuwen, F. 2005, A&A, 439, 805
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868