A Photometric Machine-Learning Method to Infer Stellar Metallicity

Adam A. Miller[1,2,*]

[1] Jet Propulsion Laboratory, California Institute of Technology,
Pasadena, CA 91109, USA
amiller@astro.caltech.edu
http://astro.caltech.edu/~amiller

[2] California Institute of Technology, Pasadena, CA 91125, USA

Abstract. Following its formation, a star's metal content is one of the few factors that can significantly alter its evolution. Measurements of stellar metallicity ([Fe/H]) typically require a spectrum, but spectroscopic surveys are limited to a few$\times 10^6$ targets; photometric surveys, on the other hand, have detected $> 10^9$ stars. I present a new machine-learning method to predict [Fe/H] from photometric colors measured by the Sloan Digital Sky Survey (SDSS). The training set consists of ∼120,000 stars with SDSS photometry and reliable [Fe/H] measurements from the SEGUE Stellar Parameters Pipeline (SSPP). For bright stars ($g' \leq 18$ mag), with 4500 K $\leq T_{\rm eff} \leq$ 7000 K, corresponding to those with the most reliable SSPP estimates, I find that the model predicts [Fe/H] values with a root-mean-squared-error (RMSE) of ∼0.27 dex. The RMSE from this machine-learning method is similar to the scatter in [Fe/H] measurements from low-resolution spectra.

Keywords: photometric surveys, machine learning, random forest, stellar metallicity.

1 Introduction

The Sloan Digital Sky Survey (SDSS, [13]) has cataloged more than one billion photometric sources, while also obtaining nearly 2 million optical spectra [1]. Despite this unprecedented volume of spectra, existing and currently planned instruments have no hope of observing each of the photometrically cataloged stars found by SDSS. Within the next decade, the Large Survey Synoptic Telescope (LSST; [7]) will dwarf SDSS, and other similar surveys, by detecting ∼20 billion photometric sources. The data volume from modern photometric surveys is too large to be examined on a source by source basis. Instead, a prudent analysis of the full data set requires advanced algorithms, such that we can identify the most interesting sources for spectroscopic observations, while also inferring the properties of those for which spectra will never be obtained.

[*] NASA Hubble Fellow.

W. Chu et al. (Eds.): DNIS 2015, LNCS 8999, pp. 231–236, 2015.

Machine-learning methods provide a promising solution to this issue: machines can readily identify patterns within the data, enabling a fast classification of the billions of stars detected in modern imaging surveys. One reason machine-learning methods are appealing is that they are data driven: the relationships they derive between observables and the parameters of interest do not rely on parametric physical models. Thus, in scenarios where we are partially ignorant to the relevant stellar physics, the machines may still be able to infer the desired stellar quantities.

Many studies have utilized machine-learning approaches to classify stellar sources of variable brightness (e.g., [4,11,5]), but only recently have efforts been made to infer fundamental physical properties via machine learning [10]. These efforts build on a long history of methods designed to estimate stellar properties, which are typically measured via spectra, from photometric observations. While the effective temperature of a star, T_{eff}, can be photometrically measured with great accuracy [6], estimates of [Fe/H] prove far more challenging [3].

A star with enhanced metal content (i.e. large [Fe/H]) produces less flux in the blue portion of its optical spectrum. Thus, imaging surveys with blue filters, such as SDSS and LSST, can be used to estimate metallicity via the photometric colors of a star. For samples restricted to F and G dwarf stars, broadband colors are capable of producing a scatter ∼0.2 dex for [Fe/H] [6]. When no restrictions are applied the best estimates from photometric methods produce a scatter of ∼0.3 dex [8].

Here, I present a new machine-learning method, which utilizes the random forest algorithm [2], that is capable of estimating [Fe/H] from the SDSS broadband photometric filters ($u'g'r'i'z'$). I train the model using a sample of ∼120,000 stars that have reliable estimates of [Fe/H] from SDSS spectroscopic observations. The final model enables a precise estimate of [Fe/H] with a low catastrophic error rate.

2 Sample

The training set for the machine learning model is constructed from the sample of stars with existing SDSS optical spectra. Every SDSS optical spectrum obtained through the eighth data release was analyzed by the SEGUE Stellar Parameters Pipeline (SSPP), a suite of algorithms optimized to estimate effective temperature (T_{eff}), surface gravity (log g), and metallicity ([Fe/H]) for stellar sources [9]. Briefly, the SSPP provides estimates of these values using multiple methods that are robustly combined to produce final adopted values of T_{eff}, log g, and [Fe/H], as well as their corresponding uncertainties. For high signal-to-noise ratio (SNR) spectra with 4500 K $\leq T_{\mathrm{eff}} \leq$ 7500 K and log $g > 2$, the SSPP measures T_{eff}, log g, and [Fe/H] with typical uncertainties of 157 K, 0.29 dex, and 0.24 dex, respectively [9]. The pipeline also flags spectra for which it cannot provide reliable estimates of the stellar parameters.

For the training sample, I include only stars that did not raise any flags during SSPP processing. From this sample of 376,073 stars, I further reject sources with flagged SDSS photometry, a single SSPP measurement of [Fe/H], $T_{\rm eff} < 4500$ K or $T_{\rm eff} > 7000$ K, or $g > 18$ mag. Finally, I remove any duplicate spectroscopic observations of the same star. These cuts are made to ensure that both the photometric and spectroscopic uncertainties are small. I summarize the cuts, as well as the number of stars remaining following each cut below:

(1) SSPP flag = nnnnn	(376,073)
(2) No SDSS photometric flags	(217,274)
(3) 4500 K $\leq T_{\rm eff} \leq$ 7500 K	(188,716)
(4) $\geq$2 SSPP [Fe/H] measurements	(182,408)
(5) $g' \leq 18$ mag	(139,176)
(6) Remove duplicates	(119,596).

In sum, there are ~120,000 stars with reliable photometry and spectroscopic measurements of [Fe/H] that are included in the model training set.

3 Model and Results

There are four features to be utilized by the machine learning model, the SDSS photometric colors $(u' - g',\ g' - r',\ r' - i',\ i' - z')$, which will enable the prediction of [Fe/H], as measured from the SDSS spectra. To perform this supervised machine-learning regression between photometric colors and [Fe/H], I adopt the random forest algorithm [2]. In short, random forest regression aggregates the results of multiple decision trees built from randomized bootstrap samples of the training set. At each node of the individual trees, the splitting parameter is selected from a random subset of the four features in the model to minimize the root-mean-squared-error (RMSE) in the resulting branches from the node. After the forest has been fully constructed, the output from each tree is averaged to provide a robust estimate of [Fe/H].

To optimize the model, the sample of ~120,000 sources is split into a training set containing a random subset of 80,000 stars, while the remaining 39,596 sources provide a test set. Tuning parameters for the random forest are adopted following a grid search and 10-fold cross-validation on the training set. The cross-validated RMSE on the training set is 0.269 dex. The results from applying the optimized model to the test set are shown in Figure 1. The RMSE for the test set is 0.273 dex, and the catastrophic error rate (CER), defined as the percentage of predictions that are incorrect by more than 0.75 dex, is 2.3%. As seen in Figure 1, the model shows a tight scatter around the one-to-one regression line. As noted above, the SSPP produces estimates of [Fe/H] with a typical uncertainty of ~0.24 dex. Thus, this machine learning method produces a scatter similar to that from a low-resolution spectrum. However, with $> 10^9$ SDSS photometrically observed sources and ~3 orders of magnitude fewer spectroscopically observed sources, the machine learning method can be applied to a significantly larger swath of stars.

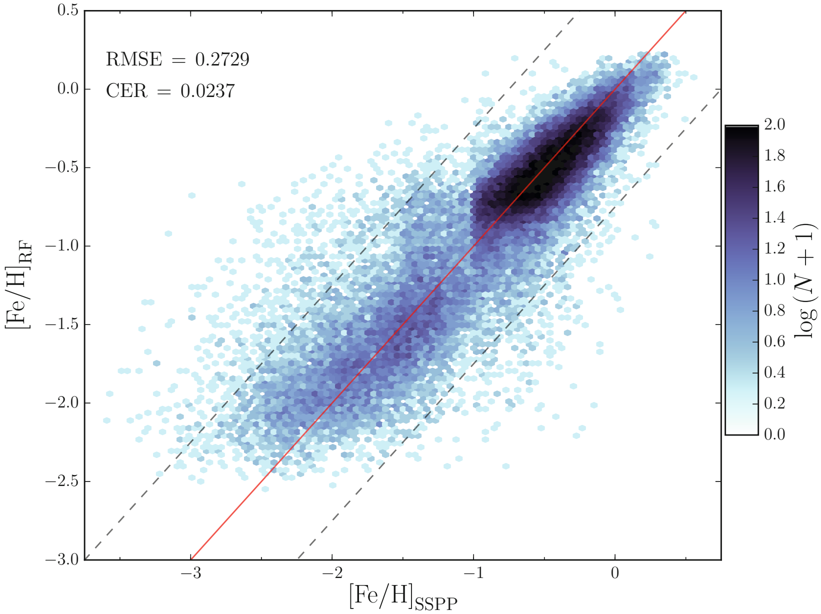

Fig. 1. Final results from the optimized random forest regression model to determine [Fe/H] from SDSS photometric colors. The spectroscopically measured values of [Fe/H] are shown on the abscissa, while the cross-validated random forest predictions are shown on the ordinate. Individual points show the density of sources in a given pixel, as color-coded according to the legend on the right. The overall performance of the model is good, with a cross-validated root-mean-square-error of ∼0.273 dex. The catastrophic error rate is small, with only 2.4% of sources having a predicted metallicity that differs from the spectroscopically measured value by more than 0.75 dex. The solid red line shows the location of a perfect one-to-one regression, while the dashed grey lines show the boundaries for catastrophic prediction errors.

4 Conclusions

Metallicity is a fundamental parameter of all stars. I have demonstrated that for stars with $4500\ \mathrm{K} \leq T_{\mathrm{eff}} \leq 7000\ \mathrm{K}$ and reliable $u'g'r'i'z'$ photometry it is possible to measure [Fe/H] with a typical scatter of ∼0.27 dex. In addition to being reliable, this method is fast and can readily be applied to billions of stars. Thus, it is possible to provide metallicity measurements for a few orders of magnitude more stars than current spectroscopic surveys. The potential applications of the method are numerous, including: the search for stellar structures in the Milky Way halo (e.g., [6]) or the discovery of the rare class of extremely metal poor stars (e.g., [12]). As additional wide-field photometric surveys come online, machine-learning techniques, such as the one described here, promise to shed light on several mysteries concerning the formation of the Milky Way.

Acknowledgement. I am thankful to Brian Bue and Umaa Rebbapragada for fruitful conversations on model selection. I am grateful for support from a NASA

Hubble Fellowship grant: HST-HF-51325.01, awarded by STScI, operated by AURA, Inc., for NASA, under contract NAS 5-26555. This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

1. Ahn, C.P., Alexandroff, R., Allende Prieto, C., Anders, F., Anderson, S.F., Anderton, T., Andrews, B.H., Aubourg, É., Bailey, S., Bastien, F.A. et al.: The Tenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Apache Point Observatory Galactic Evolution Experiment. 211, 17 (April 2014)
2. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001), `http://dx.doi.org/10.1023/A`
3. Brown, T.M., Latham, D.W., Everett, M.E., Esquerdo, G.A.: Kepler Input Catalog: Photometric Calibration and Stellar Classification. 142, 112 (October 2011)
4. Debosscher, J., Sarro, L.M., Aerts, C., Cuypers, J., Vandenbussche, B., Garrido, R., Solano, E.: Automated supervised classification of variable stars. I. Methodology. 475, 1159–1183 (2007)
5. Dubath, P., Rimoldini, L., Süveges, M., Blomme, J., López, M., Sarro, L.M., De Ridder, J., Cuypers, J., Guy, L., Lecoeur, I., Nienartowicz, K., Jan, A., Beck, M., Mowlavi, N., De Cat, P., Lebzelter, T., Eyer, L.: Random forest automated supervised classification of Hipparcos periodic variable stars 414, 2602–2617 (July 2011)
6. Ivezić, Ž., Sesar, B., Jurić, M., Bond, N., Dalcanton, J., Rockosi, C.M., Yanny, B., Newberg, H.J., Beers, T.C., Allende Prieto, C., Wilhelm, R., Lee, Y.S., Sivarani, T., Norris, J.E., Bailer-Jones, C.A.L., Re Fiorentin, P., Schlegel, D., Uomoto, A., Lupton, R.H., Knapp, G.R., Gunn, J.E., Covey, K.R., Smith, J.A., Miknaitis, G., Doi, M., Tanaka, M., Fukugita, M., Kent, S., Finkbeiner, D., Munn, J.A., Pier, J.R., Quinn, T., Hawley, S., Anderson, S., Kiuchi, F., Chen, A., Bushong, J., Sohi, H., Haggard, D., Kimball, A., Barentine, J., Brewington, H., Harvanek, M., Kleinman, S., Krzesinski, J., Long, D., Nitta, A., Snedden, S., Lee, B., Harris, H., Brinkmann, J., Schneider, D.P., York, D.G.: The Milky Way Tomography with SDSS. II. Stellar Metallicity 684, 287–325 (2008)
7. Ivezić, Ž., Tyson, J.A., Acosta, E., Allsman, R., Anderson, S.F., Andrew, J., Angel, R., Axelrod, T., Barr, J.D., Becker, A.C., Becla, J., Beldica, C., Blandford, R.D., Bloom, J.S., Borne, K., Brandt, W.N., Brown, M.E., Bullock, J.S., Burke, D.L., Chandrasekharan, S., Chesley, S., Claver, C.F., Connolly, A., Cook, K.H., Cooray, A., Covey, K.R., Cribbs, C., Cutri, R., Daues, G., Delgado, F., Ferguson, H., Gawiser, E., Geary, J.C., Gee, P., Geha, M., Gibson, R.R., Gilmore, D.K., Gressler, W.J., Hogan, C., Huffer, M.E., Jacoby, S.H., Jain, B., Jernigan, J.G., Jones, R.L., Juric, M., Kahn, S.M., Kalirai, J.S., Kantor, J.P., Kessler, R., Kirkby, D., Knox, L., Krabbendam, V.L., Krughoff, S., Kulkarni, S., Lambert, R., Levine, D., Liang, M., Lim, K., Lupton, R.H., Marshall, P., Marshall, S., May, M., Miller, M., Mills, D.J., Monet, D.G., Neill, D.R., Nordby, M., O'Connor, P., Oliver, J., Olivier, S.S., Olsen, K., Owen, R.E., Peterson, J.R., Petry, C.E., Pierfederici, F., Pietrowicz, S., Pike, R., Pinto, P.A., Plante, R., Radeka, V., Rasmussen, A., Ridgway, S.T., Rosing, W., Saha, A., Schalk, T.L., Schindler, R.H., Schneider, D.P., Schumacher, G., Sebag, J., Seppala, L.G., Shipsey, I., Silvestri, N., Smith, J.A., Smith, R.C., Strauss, M.A., Stubbs, C.W., Sweeney, D., Szalay, A., Thaler, J.J., Vanden Berk, D., Walkowicz, L., Warner, M., Willman, B., Wittman, D., Wolff, S.C., Wood-Vasey, W.M., Yoachim, P., Zhan, H.: for the LSST Collaboration: LSST: from Science Drivers to Reference Design and Anticipated Data Products. ArXiv e-prints (May 2008)

8. Kerekes, G., Csabai, I., Dobos, L., Trencséni, M.: Photo-Met: A non-parametric method for estimating stellar metallicity from photometric observations. Astronomische Nachrichten 334, 1012 (2013)
9. Lee, Y.S., Beers, T.C., Sivarani, T., Allende Prieto, C., Koesterke, L., Wilhelm, R., Re Fiorentin, P., Bailer-Jones, C.A.L., Norris, J.E., Rockosi, C.M., Yanny, B., Newberg, H.J., Covey, K.R., Zhang, H.T., Luo, A.L.: The SEGUE Stellar Parameter Pipeline. I. Description and Comparison of Individual Methods. 136, 2022–2049 (2008)
10. Miller, A.A., Bloom, J.S., Richards, J.W., Lee, Y.S., Starr, D.L., Butler, N.R., Tokarz, S., Smith, N., Eisner, J.A.: A Machine-learning Method to Infer Fundamental Stellar Parameters from Photometric Light Curves. 798, 122 (2015)
11. Richards, J.W., Starr, D.L., Butler, N.R., Bloom, J.S., Brewer, J.M., Crellin-Quick, A., Higgins, J., Kennedy, R., Rischard, M.: On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data 733, 10 (2011)
12. Schlaufman, K.C., Casey, A.R.: The Best and Brightest Metal-poor Stars. 797, 13 (2014)
13. York, D.G., Adelman, J., Anderson, Jr., J.E., Anderson, S.F., Annis, J., Bahcall, N.A., Bakken, J.A., Barkhouser, R., Bastian, S., Berman, E., Boroski, W.N., Bracker, S., Briegel, C., Briggs, J.W., Brinkmann, J., Brunner, R., Burles, S., Carey, L., Carr, M.A., Castander, F.J., Chen, B., Colestock, P.L., Connolly, A.J., Crocker, J.H., Csabai, I., Czarapata, P.C., Davis, J.E., Doi, M., Dombeck, T., Eisenstein, D., Ellman, N., Elms, B.R., Evans, M.L., Fan, X., Federwitz, G.R., Fiscelli, L., Friedman, S., Frieman, J.A., Fukugita, M., Gillespie, B., Gunn, J.E., Gurbani, V.K., de Haas, E., Haldeman, M., Harris, F.H., Hayes, J., Heckman, T.M., Hennessy, G.S., Hindsley, R.B., Holm, S., Holmgren, D.J., Huang, C.h., Hull, C., Husby, D., Ichikawa, S.I., Ichikawa, T., Ivezić, Ž., Kent, S., Kim, R.S.J., Kinney, E., Klaene, M., Kleinman, A.N., Kleinman, S., Knapp, G.R., Korienek, J., Kron, R.G., Kunszt, P.Z., Lamb, D.Q., Lee, B., Leger, R.F., Limmongkol, S., Lindenmeyer, C., Long, D.C., Loomis, C., Loveday, J., Lucinio, R., Lupton, R.H., MacKinnon, B., Mannery, E.J., Mantsch, P.M., Margon, B., McGehee, P., McKay, T.A., Meiksin, A., Merelli, A., Monet, D.G., Munn, J.A., Narayanan, V.K., Nash, T., Neilsen, E., Neswold, R., Newberg, H.J., Nichol, R.C., Nicinski, T., Nonino, M., Okada, N., Okamura, S., Ostriker, J.P., Owen, R., Pauls, A.G., Peoples, J., Peterson, R.L., Petravick, D., Pier, J.R., Pope, A., Pordes, R., Prosapio, A., Rechenmacher, R., Quinn, T.R., Richards, G.T., Richmond, M.W., Rivetta, C.H., Rockosi, C.M., Ruthmansdorfer, K., Sandford, D., Schlegel, D.J., Schneider, D.P., Sekiguchi, M., Sergey, G., Shimasaku, K., Siegmund, W.A., Smee, S., Smith, J.A., Snedden, S., Stone, R., Stoughton, C., Strauss, M.A., Stubbs, C., SubbaRao, M., Szalay, A.S., Szapudi, I., Szokoly, G.P., Thakar, A.R., Tremonti, C., Tucker, D.L., Uomoto, A., Vanden Berk, D., Vogeley, M.S., Waddell, P., Wang, S.i., Watanabe, M., Weinberg, D.H., Yanny, B., Yasuda, N., SDSS Collaboration: The Sloan Digital Sky Survey: Technical Summary. 120, 1579–1587 (2000)