Research in Astronomy and Astrophysics, 22:025017 (6pp), 2022 February https://doi.org/10.1088/1674-4527/ac41c5

Application of Random Forest Regressions on Stellar Parameters of A-type Stars and Feature Extraction*

Shu-Xin Chen[1,3], Wei-Min Sun[2], and Ying He[3]
[1] Qiqihar University, Qiqihar 161006, China
[2] Key Lab of In-fiber Integrated Optics, Ministry Education of China, Harbin Engineering University, Harbin 150009, China; sunweimin@hrbeu.edu.cn
[3] Department of Computer Science and Technology, Tianjin Ren'ai College, Tianjin 301636, China
Received 2020 September 16; revised 2021 November 28; accepted 2021 November 29; published 2022 February 2

Abstract

Measuring the stellar parameters of A-type stars is more difficult than FGK stars because of the sparse features in their spectra and the degeneracy between effective temperature ($T_{\rm eff}$) and gravity ($\log g$). Modeling the relationship between fundamental stellar parameters and features through machine learning is possible because we can employ the advantage of big data rather than sparse known features. As soon as the model is successfully trained, it can be an efficient approach for predicting $T_{\rm eff}$ and $\log g$ for A-type stars especially when there is large uncertainty in the continuum caused by flux calibration or extinction. In this paper, A-type stars are selected from LAMOST DR7 with a signal-to-noise ratio greater than 50 and the $T_{\rm eff}$ ranging within 7000 to 10,000 K. We perform the Random Forest (RF) algorithm, one of the most widely used machine learning algorithms to establish the regression relationship between the flux of all wavelengths and their corresponding stellar parameters ($T_{\rm eff}$) and ($\log g$) respectively. The trained RF model not only can regress the stellar parameters but also can obtain the rank of the wavelength based on their sensibility to parameters. According to the rankings, we define line indices by merging adjacent wavelengths. The objectively defined line indices in this work are amendments to Lick indices including some weak lines. We use the Support Vector Regression algorithm based on our new defined line indices to measure the temperature and gravity and use some common stars from Simbad to evaluate our result. In addition, the Gaia Hertzsprung-Russell diagram is used for checking the accuracy of $T_{\rm eff}$ and $\log g$.

Key words: methods: data analysis – surveys – stars: early-type – stars: abundances

1. Introduction

The A-type stars encompass a bewildering array of stellar types, and many horizontal-branch stars shown in the A-type star region on the Hertzsprung-Russell (HR) diagram suggest their evolutionary states. The fundamental stellar atmospheric parameters ($T_{\rm eff}$ and $\log g$) are the basis for astrophysics study of A-type stars, and estimation of these parameters are often from strong Balmer spectral lines. For low-resolution spectra, line index is an effective method to extract spectral features and has been widely used in astronomical research. Cenarro (2001) used the line index to calculate the Ca II flux and measured stellar atmospheric parameters to determine the effective temperature. Covery et al. (2007) wrote IDL programs to use Hammer line index to automatically classify stellar spectra. Yi et al. (2014) also added the features extracted from the spectrum using the Random Forest (RF) algorithm on the basis of Covery's program as a new feature index and applied it to the spectral classification of M dwarfs, and proving that the improved feature index has a better performance in the classification of M dwarfs. Inspired by the work of Yi et al. (2014), we apply RF in A-type stars to define new spectral line indices representing features for low-resolution spectra, and this specific definition of line indices of A-type stars is sensitive to their stellar parameters.

Among all definitions of line index systems, the Lick index is one of the most widely used line index systems applied in many spectral analysis fields. The line indices for A-type stars released by LAMOST were calculated following the definition of the Lick system, which includes most of the prominent absorption lines. Hou et al. (2014) described the details of lines of A-type stars for low-resolution spectra. The advance of using Lick indices is that the error of flux calibration and radial velocity measurement can be ignored and the noise has little effect on the line indices. Tan et al. (2013) used line index as the training feature of sky survey data in the measurement process of stellar atmospheric physical parameters, and obtained the best regression model in the training of linear regression. Wang et al. (2014) used the Lick line index and

* Supported by the National Science Foundation for Young Scientists of China Grant No. 11800313. and the Joint Research Fund in Astronomy (U2031142) under cooperative agreement between the National Natural Science Foundation of China (NSFC) and Chinese Academy of Sciences (CAS). Technology Innovation Center of Agricultural Multi-Dimensional Sensor Information Perception, Heilongjiang Province.

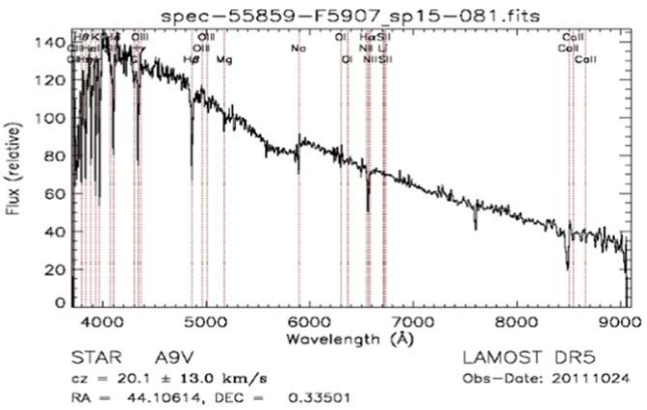

Figure 1. A pipeline classified A9V-type star "spec-55859-f5907_sp15-081.fits", whose $T_{\rm eff}$ is 6833.16 K.

applied the partial least-squares regression method for the measurement of the atmospheric physical parameters. The result of the partial least-squares regression model is not only consistent with the parameters of Sloan Stellar Parameter Pipeline (SSPP) released but also the partial least-squares regression can reduce the computational complexity, speed up the training process. Pan et al. (2015) pointed out different sensitivities of spectral lines to the effective temperature of main-sequence stars. They used line index as input of Support Vector Machines (SVM) to do the classification of stars.

However, there are only strong lines in the Lick system that are not enough for the correct parameterization of A-type stars. Thus, we are motivated to accurately estimate the $T_{\rm eff}$ and log g for A-type stars and get relatively weak features that are sensitive to the stellar parameters. To obtain the possible additional features, we choose to use the decision tree based RF algorithm to extract more features other than Balmer lines and Calcium HK, etc. RF is a regression method that has been used in several astronomical research. For example, Bai et al. (2019) applied RF to the stellar effective temperature regression for the second Gaia data release with the precision of about 191 K, based on the combination of the stars in four spectroscopic surveys.

In this work, we use LAMOST DR7 released A-type spectra with full wavelength as input of RF algorithm to establish the regression model for stellar parameters. Then we rank the wavelength according to the sensitivity to the parameters and obtain the most sensible lines finally. We then define the line indices for these lines and compare them to Lick indices. Using the newly defined indices, we employ Support Vector Regression (SVR) to estimate the stellar parameters for A-type stars. The result of temperature and gravity from our method agrees with those from LAMOST. Cross-matching with Simbad, we get around 200 common stars with published parameters. A comparison of parameters is conducted to the common star. In addition, we calculate the absolute magnitude

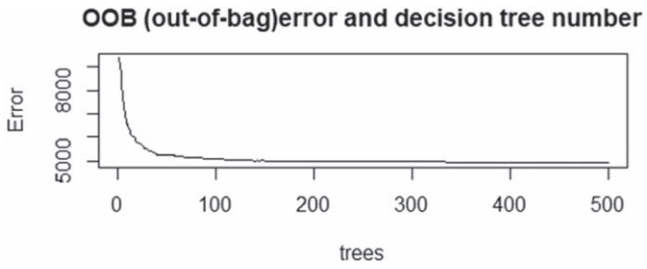

Figure 2. OOB(Out-Of-Bag) error and decision tree number in the random forest.

for the star with Gaia parallax and use the HR diagram to check our result.

The article is organized as follows. In Section 2, we introduce the LAMOST data we used. In Section 3, we present the application of RF regression in deriving $T_{\rm eff}$, log g and [Fe/H] of A-type stars from full spectra and definition of specific line indices for parameter determination of A-type stars. Section 4 introduces the application of SVR to estimate stellar parameters using our defined indices, and also presents HR and Keil diagrams to check the parameters we compute, and Section 5 summary the work in this paper.

2. Data

2.1. LAMOST Released Spectra of A-type Stars

The published LAMOST DR7 catalog includes 599,762 A-type star spectra, which were obtained during the pilot survey and 7 yr regular surveys. There are two formats for the A-type star catalog: i.e., FITS and CSV. The full spectra ranging from 3700 to 8800 Å are used as input of the RF algorithm in the first run. The class of these stars contains both spectral type and luminosity class provided by the LAMOST analysis pipeline. We also compare our defined index system with the line indices published in the LAMOST LRS Line-Index Catalog of A-Type Stars. The comparison includes kp12, Halpha12, and Hgamma12 are the Ca II-K, Halpha, and Hgamma. $T_{\rm eff}$ and log g are from the catalog LAMOST LRS Stellar Parameter Catalog of A, F, G, and K Stars, in which parameters of 114,208 A type spectra are included. Cross-matching with Gaia EDR3, we obtained 108,581 stars with good parallax. We also remove some spectra classified as A-type but with a temperature lower than 7000 K. An example is shown in Figure 1 titled "spec-55859-f5907_sp15-081.fits", of which the effective temperature is 6833 K and class is A9V-type. Thus, we selected A-type stellar data with temperatures from 7000 to 12,500 K and S/N greater than 50.

2.2. Removing Contamination of Negative Index Values

To obtain a robust relationship between stellar atmospheric parameters and spectral features for A-type stars, a clear sample without affection emission lines from stellar disks or exchange

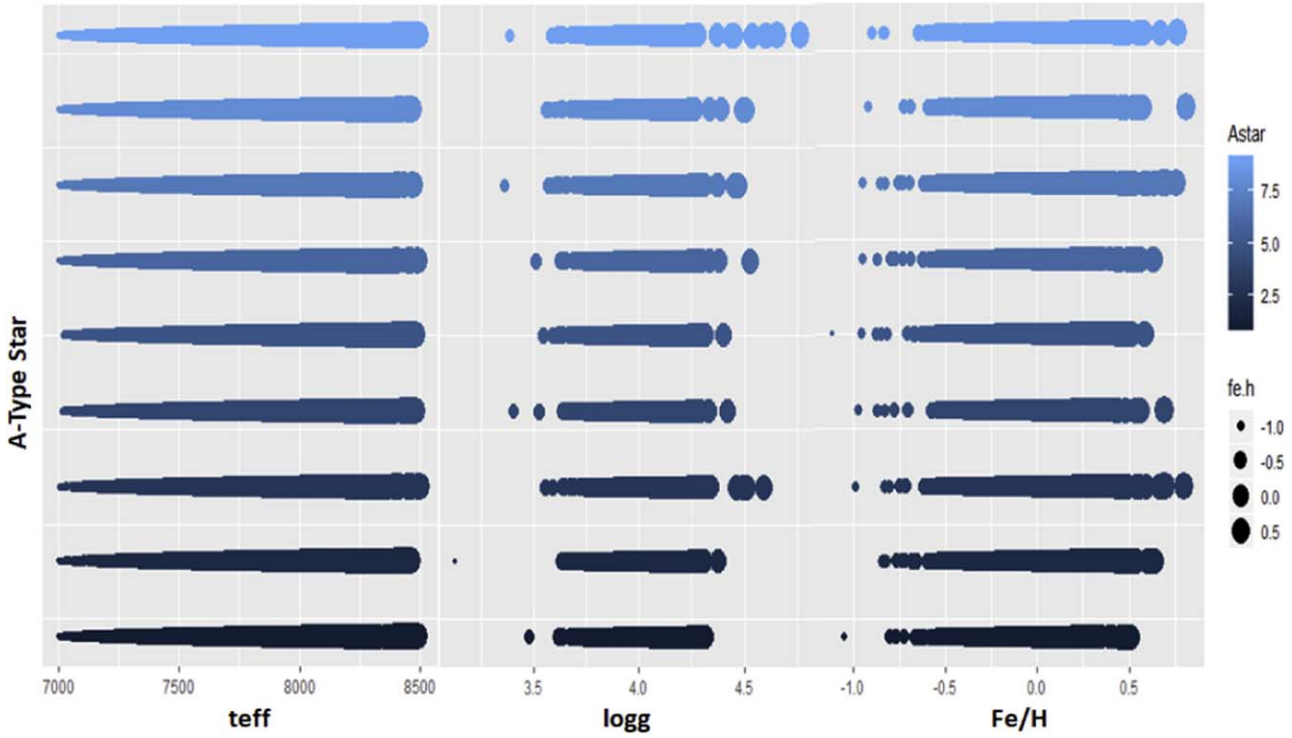

Figure 3. Distribution of the three physical parameters $T_{\rm eff}$ (effective temperature), and log g (surface gravity), and Fe/H (chemical abundance) from A-type stellar spectra published by LAMOST.

of material between binaries is necessary. We checked the line indices of A-type stars released by LAMOST and remove those spectra having negative index values.

3. Random Forest Prediction Analysis

The random forest (RF) algorithm, which belongs to the ensemble learning method in machine learning, is a combination of supervised prediction models. It can handle high-dimensional data sets with good advantages and hold thousands of input variables. The model can output the importance of variables and establish a model for setting the variables of the data set. All decision trees depend on the corresponding random vectors. All the vectors are independent and identically distributed, and the most important variables are determined by reducing the dimensionality. Finally, the results of the classification tree are summed, and the accuracy of the prediction model is improved. Even with a large number of missing data, RFs can also maintain accuracy.

3.1. Random Sampling in the Whole Dataset

From the total A-type data set of around 80 thousand spectra described in Section 2.1, we randomly sample the data set to train the model. Section 3.3 will introduce the method for calculating the distance between different data points through an RF, thus realizing the regression. When the data set is not verified, the outside prediction error can be calculated, the category corresponding to the sample points that are not used when the tree is generated can be estimated by the spanning tree, and the outside prediction can be obtained by comparing with the real category.

3.2. Normalization

Before establishing the RF model, we remove the pseudo-continuum of each spectrum to keep spectral lines. We use a ninth-order polynomial to fit each spectrum, removing those points outside 3σ from the fitted curve, and iteratively repeat the fitting four times. Then the intensity of each spectrum is rectified by dividing the observed spectrum by the pseudo-continuum.

3.3. Random Forest Algorithm

All vectors in the RF are independent and identically distributed. Random forests are randomizations of column variables and row observations of data sets, generating multiple classification numbers. Finally, the results of classification trees are aggregated. Compared to neural networks, RFs reduce computation and improve prediction accuracy. Moreover, this algorithm is not sensitive to multicollinearity, and it is

Table 1
Identification of Elements Sensitive to Parameters based on the Location of the First 30 Feature Points

ID	Name	Vacuum Wavelength	Importance for Teff
1, 3, 4, 9,18, 20	Ca II K	3933, 3935, 3934, 3936, 3932, 3931	0.495, 0.090, 0.054, 0.005, 0.002, 0.002
2, 5, 6,13	Co I Mn I Cr I V I	4110, 4109, 4111, 4112	0.128, 0.037, 0.025, 0.004
7, 11, 14, 15, 24	Fe I Mn I Ti I Cr I	3980, 3978, 3982, 3979, 3981	0.007, 0.005, 0.004, 0.003, 0.002
8, 22	CN	3879, 3878	0.006, 0.002
10, 21, 25, 29	Fe I	3898, 3899, 3906, 3900	0.005, 0.002, 0.002, 0.002
12	Ca I	4108	0.005
16	Heta	3970	0.003
17	CH	3963	0.002
19	Fe II	4097	0.002
23, 26	Fe I	4107, 4106	0.002, 0.002
27	Mg I	3903	0.002
28	Ca I	4098	0.002
28	Fe I	3977	0.002
ID	Name	Vacuum Wavelength	Importance for logg
1, 2, 8	Co I Mn I Cr I V I	4110, 4111, 4112	0.065, 0.055, 0.010
3, 10, 11	Eu I Ba II Si II	4130, 4129, 4131	0.025, 0.008, 0.008
4, 23	Fe I CH	4181, 4180	0.022, 0.005
5, 6,18	Fe I	3966, 3967, 3960	0.013, 0.011, 0.007
7, 15, 24, 25, 27, 28, 30	Ca II K	3933, 3835, 3937, 3932, 3934, 3936,3931	0.011, 0.007, 0.005, 0.005, 0.005, 0.004
9	CH	4345	0.008
12, 22	Ti I Fe I Ca I	3956, 3957	0.008, 0.006
13	CN	3860	0.007
14,20	La II, Fe I, Ti I	3989 3988	0.007,0.006
16	Cr I Fe I	4142	0.007
17	Fe I	3909	0.007
19	CH Co I	3873	0.007
21	CH Cr I	4339	0.006
26	La II Fe I Cr I	3949	0.005
29	Mn I Ti I Fe I	4026	0.005
ID	Name	Vacuum Wavelength	Importance for [Fe/H]
1	Fe I Ti I	4078	0.176
2	Sr,II	4077	0.106
3	Mn I	4131	0.051
4, 20, 30	Fe I	4032, 4037, 4069	0.048, 0.008, 0.006
5, 6,10	Fe I Fe II	3969, 3965, 3966	0.016, 0.014, 0.011
7, 16	CH Mn I	4033, 4034	0.014,0.009
8, 21, 25	Ni,I Fe I	4142, 4140, 4141	0.014,0.009,0.007
9, 15, 18, 19,23,27	Fe I	3920, 3918, 3939, 3937, 3940, 3936	0.011, 0.009, 0.009, 0.008, 0.007, 0.007
11	Eu,II	4129	0.010
12,13,14	Fe I	3954, 3952, 3953	0.010, 0.010,0.010
17	Y,II	3950	0.009
22	Mn I Ti I Fe I	4132	0.007
24	HBeta	4861	0.007
26	Fe I Cr I	4337	0.007
28	Ca II K	3933	0.006
29	CH	4345	0.006

sufficiently robust to process missing data and non-balanced data.

The RF algorithm for prediction and regression mainly includes N randomly selected sample units from the original data to generate decision or regression trees, and $m < M$ randomly selected variables at each node as the candidate variables of the segmentation node. The number of variables at each node should be consistent. The full wavelength spectra as input of the RF and the results of each decision or regression tree are integrated to generate predicted values. In the training

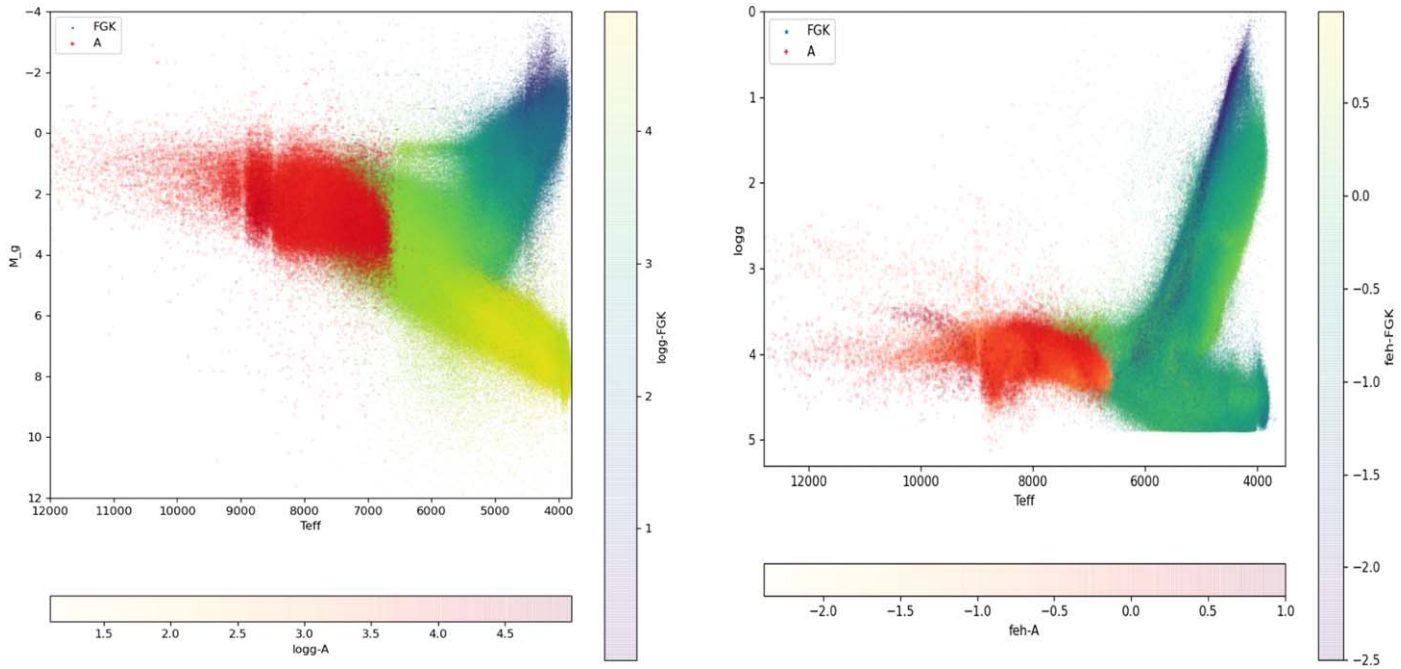

Figure 4. Checking the $T_{\rm eff}$ and $\log g$ with both on the HR (left panel) and Keil (right panel). Red dots in both panels represent A-type stars with the parameter estimated through line indices.

Table 2
List of Three New Defined Line Indices for Parameter Regression

Name	Index Band-pass (Å)	Left Band (Å)	Right Band (Å)
Ca II K	3929–3937	3920–3922	4006–4010
Blend(Co I Mn I Cr I V I)	4109–4113	4103–4106	4115–4120
Sr II	4076–4078	4073–4075	4080–4082

Table 3
Effective Temperature $T_{\rm eff}$ as Predicted by Random Forest Algorithm with Three New Indices as Input

Input			$T_{\rm eff}$ Prediction		
Ca II K	Blended	Sr II	fit	lwr	upr
2.88	5.88	5.88	8000.029	7977.766	8022.293
2.58	5.58	5.58	7887.576	7857.422	7917.731

process, multiple decision trees will be generated, and each decision tree will produce a corresponding prediction output according to the input data set. The number of decision trees is a key parameter in the RF algorithm, the larger the number of decision trees, the better the regression results, the longer time consumption. In this work, we used 3800 decision trees as well as the number of input spectral data points. The remaining parameters were set to the default values.

The out-of-bag (OOB) error—which is an unbiased estimate of the generalization error whose result approximates the K-th tree fold cross-validation which requires additional computation —and the decision tree number in the RF are shown in Figure 2. The number of trees is about 500 to realize the regression. The difference for each split is less than 1. Mean of squared residuals is 4926.627, in addition, Var value is 96.57, which comply with the requirements of Section 3.

We rank the wavelength according to the importance of the parameters and then identify the spectral lines where the first 30 feature points for $T_{\rm eff}$, $\log g$ and [Fe/H] are located by searching for the line table from Moore et al. (1966). The details are listed in Table 1. We only listed the main elements contained in spectral lines with low-resolution. The first column lists the feature ID. In order to make the table more concise, features that fall on the same absorption line are placed in the same entry. The second column shows the name of the line in which feature points are located. The third column lists the vacuum wavelength corresponding to each spectral line. The fourth column shows the importance of the corresponding feature determined with the RF algorithm.

As listed in Table 1, we group the conjuncted wavelengths as spectral features. To obtain the most sensitive lines to three parameters, we consider top one or two features for each parameter. Then, we defined three most important features, Ca II K at 3933 Å, blended feature of Co I, Mn I, Cr I, and V I

lines ranging from 4109 to 4112 Å, and Sr II at 4077 Å. The detailed definitions are listed in Table 2, including the feature name, index bandpass, and two sidebands.

3.4. Random Forest with New Defined Line Indices

In the RF algorithm, each tree grows to its maximum extent, and there is no branch-pruning process. Using training data that perform better in regression analysis can result in improved learning model characteristics. In this step, we made an RF temperature model using Ca II K, Blended Co I Mn I Cr I V I, and Sr II as input rather than using full spectra, the effective temperature is predicted as shown in Table 3.

4. Verification with SVR Algorithm

SVR is one of the best regression algorithms that focuses on handling overall error and tries to avoid outlier issues better than algorithms like linear regression. SVR builds a hyperplane in an N-Dimensional vector space, where we aim to keep data points inside the hyperplane for regression. We tried the SVR algorithm using the software package Sklearn with the newly defined line indices as input. Comparing with LAMOST stellar parameter catalog, the precision is 123 K for $T_{\rm eff}$, 0.32 dex $\log g$, and 0.28 dex for [Fe/H] respectively.

4.1. Verification with Gaia Data

We cross match our sample with Gaia using Topcat to obtain parallax of these A type stars, and then calculate their absolute magnitudes. We plot them on both the HR and Keil diagrams to verify the regression results shown in Figure 4.

4.2. HR Diagram of A-type Stars

A schematic representation of how rotation affects the position of a star in the HR diagram, shown as Figure 4. In any case, a rotating star generally appears to be above the main sequence. Rotation displaces a star in the HR diagram. Consider a star seen in the equatorial plane. If it were possible to increase this star's rotational velocity, we would see it move to the right and down, which toward cooler temperatures and lower luminosities. On the other hand, a star seen pole-on toward higher luminosities would move generally upwards in the HR diagram. Neither of these paths is necessarily parallel to the main sequence, and so a rapidly rotating main-sequence star, no matter the orientation, tends to lie above the main sequence. The A-type and early F-type stars have detected subtle differential effects in the spectra and photometry of rapid rotators, even those that are seen pole-on.

5. Discussion

Because line index would not be seriously affected by noise, it is a good feature representation of stellar spectra especially with low S/N ratio. In this work, we re-define a line index system using the RF algorithm. We apply the system in the LAMOST DR7 and get very good prediction performance. The indices are verified with SVR, and the correctness is verified by using Gaia data. The result shows that the RFs are a very useful tool for feature extraction dealing with high-dimensional data. For unbalanced data sets, RFs provide an effective way to balance data set errors to achieve balanced errors. Using our newly defined line index system for A type stars to predict the stellar parameters of A-type stars, we can avoid the effect of interstellar extinction and degeneration of parameters.

Acknowledgments

We are very grateful to the anonymous referee for many useful comments and suggestions. This work was funded by the Joint Research Fund in Astronomy (Grant No. U2031142) under cooperative agreement between the National Natural Science Foundation of China (NSFC) and Chinese Academy of Sciences (CAS); the National Science Foundation for Young Scientists of China (Grant No. 11803013) and Technology Innovation Center of Agricultural Multi-Dimensional Sensor Information Perception, Heilongjiang Province.

This research uses data obtained through the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST), which is funded by the National Astronomical Observatories, Chinese Academy of Sciences.

References

Bai, Y., Liu, J., Bai, Z., Wang, S., & Fan, D. 2019, AJ, 158, 93
Cenarro, A. J. 2001, MNRAS, 32, 959
Covery, K. R., Ivezić, Ž., Schlegel, D., et al. 2007, AJ, 134, 2398
Hou, W., Luo, A. L., Ren, J. J., Wei, P., & Li, Y. B. 2014, in Proc. Int. Conf. Putting A Stars into Context: Evolution, Environment, and Related Stars, Moscow, 3–7 June 2013, 145
Moore, C. E., Minnaert, M. G. J., & Houtgast, J. 1966, The Solar Spectrum 2935 Å to 8770 Å (Washington, DC: US Govt Printing Office)
Pan, J., Wang, J., Tan, X., Yu, J., & An, R. 2015, JICS, 12, 5339
Tan, X., Pan, J. C., Wang, J., Luo, A., & Tu, L. 2013, Spectrosc. Spectral Anal., 33, 1397
Wang, J., Pan, J., & Tan, X. 2014, Spectrosc. Spectral Anal., 34, 833
Yi, Z., Luo, A., Song, Y., et al. 2014, AJ, 147, 33