Southern Horizons in Time-Domain Astronomy
Proceedings IAU Symposium No. 339, 2019
E. Griffin, ed.

doi:10.1017/S1743921318002983

Stellar Parameterisation using KPCA and SVM

H. Yuan, Y. Zhang, Y. Lei, Y. Dong, Z. Bai, G. Li, W. Zhang, H. Zhang and Y. Zhao

Key Laboratory of Optical Astronomy, NAO, CAS, Beijing, China
email: yuanhl@bao.ac.cn

Abstract. With so many spectroscopic surveys, both past and upcoming, such as SDSS and LAMOST, the number of accessible stellar spectra is continuously increasing. There is therefore a great need for automated procedures that will derive estimates of stellar parameters. Working with spectra from SDSS and LAMOST, we put forward a hybrid approach of Kernel Principal Component Analysis (KPCA) and Support Vector Machine (SVM) to determine the stellar atmospheric parameters effective temperature, surface gravity and metallicity. For stars with both APOGEE and LAMOST spectra, we adopt the LAMOST spectra and APOGEE parameters, and then use KPCA to reduce dimensionality and SVM to measure parameters. Our method provides reliable and precise results; for example, the standard deviation of effective temperature, surface gravity and metallicity for the test sample come to approximately 47–75 K, 0.11–0.15 dex and 0.06–0.075 dex, respectively. The impact of the signal:noise ratio of the observations upon the accuracy of the results is also investigated.

Keywords. Techniques: spectroscopic, methods: data analysis, stars: atmospheres

1. Introduction

Massive databases of spectra of celestial objects are being obtained nowadays by surveys of large areas of the sky, such as SDSS (York *et al.* 2000; http://www.sdss.org/), 2dF (Colless *et al.* 2001; http://www.2dfgrs.net/), and LAMOST (Wang *et al.* 1996; http://www.lamost.org/). Analyses of those data yield radial velocities and atmospheric parameters of several million objects. Many novel parameterisation methods have been developed and adapted to different applications. Nevertheless, since most of the spectra from the large sky surveys are of medium or low resolution, the efficiency and precision of applying existing methods to those big data sets should be validated and tuned in detail.

The APO Galactic Evolution Experiment (APOGEE) (Holtzman *et al.* 2015), which is part of the SDSS project, has published several hundred thousand high-resolution, high signal-to-noise infrared spectroscopy data together with well-estimated stellar parameters. The combination of the APOGEE stellar parameters with the large volumes of LAMOST spectra has a high potential. The Kernel Principal Component Analysis (KPCA), an extension of Principal Component Analysis (PCA), is commonly used to reduce dimensionality. In the case of astronomical spectra, the pixel number (the dimensionality) is about several thousand and is thus too large for most machine-learning algorithms. Instead of extracting complicated spectral indices by applying astronomical knowledge, the KPCA dimensionality reduction is data oriented and easy to use. The processed data are then trained and tested using Support Vector Machine (SVM). When we select objects common to both LAMOST and APOGEE as the training set, the stellar parameters of other LAMOST spectra can be predicted. In this work, the training

Table 1. Performance of KPCA and SVM methods using APOGEE and LAMOST data.

g-band SNR	Parameter	Sample Size	Gaussian fit σ	Gaussian offset	N-fold σ	N-fold offset
100–500	$T_{\rm eff}$	3617	23.58 K	–0.076 K	47.76 K	–0.62 K
50–100	$T_{\rm eff}$	5820	33.79 K	–0.39 K	50.94 K	0.82 K
20–50	$T_{\rm eff}$	9551	43.55 K	–1.375 K	75.22 K	2.30 K
100–500	log g	3530	0.052 dex	–0.00034 dex	0.11 dex	–0.0059 dex
50–100	log g	5650	0.070 dex	–0.00084 dex	0.11 dex	–0.0043 dex
20–50	log g	9544	0.093 dex	0.00093 dex	0.15 dex	–0.0063 dex
100–500	[Fe/H]	3509	0.019 dex	0.00049 dex	0.056 dex	–0.0065 dex
50–100	[Fe/H]	5629	0.024 dex	–0.00022 dex	0.066 dex	–0.0077 dex
20–50	[Fe/H]	9346	0.035 dex	0.00019 dex	0.075 dex	–0.0060 dex

precision is validated by applying a 10-fold cross-validation method. The result should provide a valuable reference for all big-volume machine-learning algorithms.

2. Data and Method

Targets that are common to both APOGEE and LAMOST were chosen and divided into different groups according to their signal-to-noise ratios (SNR) in the g-band. We used normalised spectra from LAMOST and the parameters from APOGEE. The cross-match distance of APOGEE and LAMOST was 3″. Individual pixels with abnormal masks were removed. Wavelengths were corrected to the rest-frame using radial velocities from APOGEE, and then re-sampled between 4000–8800 Å with a step length of 1 Å; fluxes were normalised. We limited the method to data with $3500 < T_{\rm eff} < 5500$ K. The *scikit-learn* python package (See http://scikit-learn.org) was used. First, the spectral dimensionality reduction was performed using KPCA with the radial basis function (RBF) kernel ($\gamma = 10$). An output dimension of 500 was selected. The parameter regression was then carried out using SVM with the RBF kernel. A small number of outliers was removed in order to keep the sample clean. Two classes for cross-validation from the *scikit-learn* package, GridSearchCV and RandomizedSearchCV, were applied to optimise the SVM parameters.
The final parameters for $T_{\rm eff}$ were $[\gamma = 4096, C = 16, \epsilon = 0.02]$;
the final parameters for log g were $[\gamma = 8, C = 16, \epsilon = 0.01]$;
the final parameters for [Fe/H] were $[\gamma = 16, C = 4, \epsilon = 0.01]$. A 10-fold cross-validation was performed to estimate the performance. Each time 90% of the sample was reduced dimensionally and trained; the remaining 10% was processed with the trained reduction model and the regression model.

3. Results

To assess the performance, we compared the estimated values with the ones from APOGEE using the 10-fold cross-validation method. The results are shown in Table 1. Our results for the high-SNR case are also illustrated in Fig. 1. The cross-validation standard deviation in $T_{\rm eff}$ was 47.76 K, 50.94 K and 75.22 K, for the 100–500, 50–100 and 20–50 SNR groups, respectively. The deviation was obviously smaller for higher SNR data. However, the difference between the first two SNR groups was not distinct. The SNR suggested for applying this method is above 50 in the g-band. It was also apparent that the Gaussian fit deviation was much smaller than the N-fold cross-check one, by approximately 50%. Further analysis showed that more than 95% of the sample

https://doi.org/10.1017/S1743921318002983 Published online by Cambridge University Press

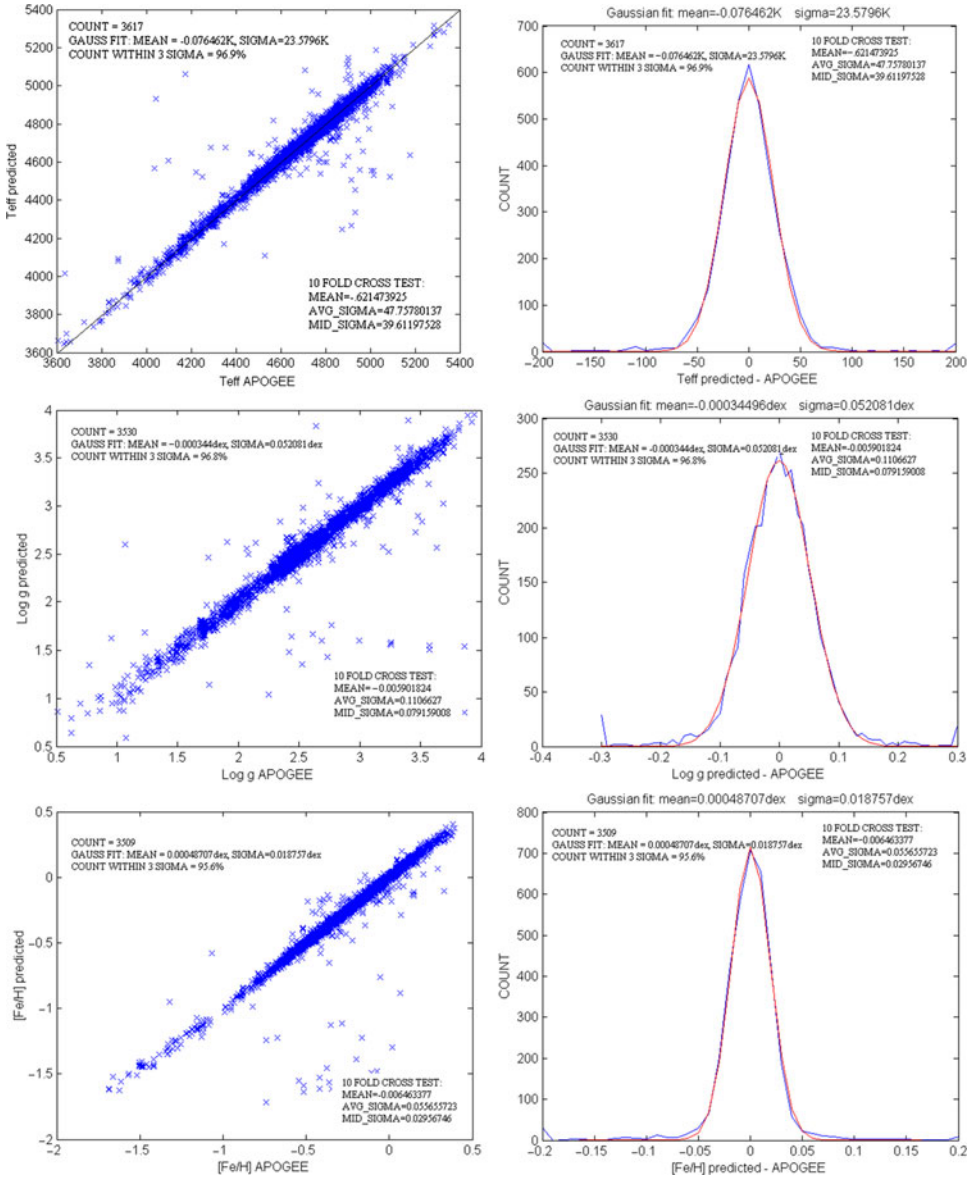

Figure 1. Performance of KPCA and SVM method using APOGEE and LAMOST data in the g-band; SNR between 100 and 500.

had deviations less than 3 × the Gaussian fit σ. The mean offset between the predicted values and the values from APOGEE was almost negligible.

Deviation in log g varied from 0.11–0.15 dex, depending on the SNR. Compared to log g, the deviation for [Fe/H] was smaller by about 50%. The SNR had a similar effect on the estimate of log g and [Fe/H], as seen in $T_{\rm eff}$. The Gaussian fit deviation was smaller than the N-fold cross-validation by approximately 60%.

The larger deviation in log g, compared to that of [Fe/H], may be caused by the small range of log g covered by APOGEE, which focused on observing Galactic red giants. At present our method of using APOGEE parameters and LAMOST spectra is applicable to giants, and can be expanded to other stars if the training set is changed.

https://doi.org/10.1017/S1743921318002983 Published online by Cambridge University Press

4. Conclusions

In this work, the KPCA and SVM methods were used with APOGEE parameters and LAMOST spectra. The training parameters for $T_{\rm eff}$ are $[\gamma = 4096, C = 16, \epsilon = 0.02]$; those for log g are $[\gamma = 8, C = 16, \epsilon = 0.01]$, and those for $[{\rm Fe/H}]$ are $[\gamma = 16, C = 4, \epsilon = 0.01]$. The standard deviation of the effective temperature, surface gravity and metallicity respectively come to approximately 47–75K, 0.1–0.15 dex and 0.06–0.075 dex, respectively. Data with higher SNR can produce much better results. A lower limit to SNR of 50 in the g-band was suggested, but performances with SNR as low as 20 were also acceptable. 95% of the sample have deviations less than 3σ. Our method does appear to be effective for estimating stellar-atmosphere parameters.

Acknowledgements

This work was funded by the National Natural Science Foundation of China under grant nos. 11403046 and U1731109. The LAMOST Fellowship is supported by Special Funding for Advanced Users, budgeted and administered by the Center for Astronomical Mega-Science, Chinese Academic of Sciences.

References

York, D. G., Adelman, J., Anderson, J. E., Jr., *et al.* 2000, *AJ*, 120, 1579
Colless, M., Dalton, G., Maddox, S., *et al.* 2001, *MNRAS*, 328, 1039
Wang, S-G., Su, D-Q., Chu, Y-Q., Cui, X., & Wang, Y-N. 1996, *AO*, 35, 5155
Holtzman, J. A., Shetrone, M., Johnson, J. A., *et al.* 2014, *AJ*, 150, 148