ELSEVIER

CrossMark

Chinese Astronomy and Astrophysics 41 (2017) 331–346

CHINESE
ASTRONOMY
AND ASTROPHYSICS

Application of Multi-task Sparse Lasso Feature Extraction and Support Vector Machine Regression in the Stellar Atmospheric Parameterization† ⋆

GAO Wei LI Xiang-ru△

School of Mathematical Sciences, South China Normal University, Guangzhou 510631

Abstract The multi-task learning takes the multiple tasks together to make analysis and calculation, so as to dig out the correlations among them, and therefore to improve the accuracy of the analyzed results. This kind of methods have been widely applied to the machine learning, pattern recognition, computer vision, and other related fields. This paper investigates the application of multi-task learning in estimating the stellar atmospheric parameters, including the surface temperature (T_{eff}), surface gravitational acceleration (lg g), and chemical abundance ([Fe/H]). Firstly, the spectral features of the three stellar atmospheric parameters are extracted by using the multi-task sparse group Lasso algorithm, then the support vector machine is used to estimate the atmospheric physical parameters. The proposed scheme is evaluated on both the Sloan stellar spectra and the theoretical spectra computed from the Kurucz's New Opacity Distribution Function (NEWODF) model. The mean absolute errors (MAEs) on the Sloan spectra are: 0.0064 for lg $(T_{\rm eff}/{\rm K})$, 0.1622 for lg $(g/({\rm cm}\cdot{\rm s}^{-2}))$, and 0.1221 dex for [Fe/H]; the MAEs on the synthetic spectra are 0.0006 for lg $(T_{\rm eff}/{\rm K})$, 0.0098 for lg $(g/({\rm cm}\cdot{\rm s}^{-2}))$, and 0.0082 dex for [Fe/H]. Experimental results show that the proposed scheme has a rather high accuracy for the estimation of stellar atmospheric parameters.

Key words stars: fundamental parameters, methods: data analysis, methods: statistical, methods: miscellaneous

---

† Supported by National Natural Science Foundation (61273248, 61075033), NNSF-CAS Joint Foundation (U1531242), Guangdong Natural Science Foundation (2014A030313425, S2011010003348)

Received 2015–12–10; revised version 2016–01–26

⋆ A translation of *Acta Astron. Sin.* Vol. 57, No. 4, pp. 389–401, 2016

△ xiangru.li@gmail.com

doi:10.1016/j.chinastron.2017.08.004

1. INTRODUCTION

Thanks to the rapid development of modern science and technology, the Sloan digital sky survey telescope of the United States[1] has obtained a large number of spectral data, and the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST)[2] in China is now the telescope with the highest acquisition rate of spectral data in the world, one observation can obtain simultaneously the spectra of about 4000 celestial objects, making the number of celestial spectra observed by humans increase to the order of magnitude of ten millions.

The acquirement of the huge number of stellar spectral data makes the problem how to estimate accurately and rapidly the 3 atmospheric physical parameters (surface temperature T_{eff}, surface gravitational acceleration $\lg g$, and chemical abundance [Fe/H]) of stars become a subject worth to be investigated. As we known, the number of spectra from remote celestial objects is huge, and in the propagation they will suffer a large number of noise interferences, such as the atmospheric environment, scattered light, cosmic ray, etc., as well as the influences of the instrumental instability and systematic error. These factors will affect seriously the accuracy and speed of our estimation on the stellar atmospheric parameters, hence, we have to make preprocessing on the spectral data to reduce the data number, raise the speed, and reduce the noise interference, then to estimate the atmospheric physical parameters from the extracted spectral features.

In the fields of pattern recognition, data digging, and so on, the data preprocessing mentioned above is called the feature extraction. The typical feature extraction methods include the stacked autoencoder based on the neural network, the Principle Component Analysis (PCA)[3,4], the Least Absolute Shrinkage Selection and Operator (Lasso)[5,6], etc. Especially for the Lasso algorithm proposed by Tibshirani in 1996, by imposing an l_1-norm constraint on the unknown coefficient vector, this method makes the coefficients with a s-maller magnitude shrink automatically into zeroes, so as to realize the purposes of variable selection and feature extraction. However, the Lasso method has two limitations: firstly, one piece of spectral data contains the information of all the spectroscopic physical parameters, and between the different spectroscopic physical parameters some potential relations may exist, the above method takes considerations of the 3 stellar atmospheric parameters separately, and makes the feature extraction and dimension reduction on the spectral data, it will cause the loss of the imbedded information between the different physical parameters, and therefore the reduction of prediction accuracy; secondly, the separated treatments of the 3 physical parameters are overelaborate, time-consuming, and low-efficient.

In fact, to make the feature extraction from the spectra data by taking consideration of the 3 stellar atmospheric parameters altogether is a multi-task learning (MTL) problem. Namely, this is the method that multiple tasks are learning together to dig out the relationship information among the different tasks, and to recognize simultaneously the differences between the different tasks, so that to improve the prediction accuracy and generalization ability of the prediction model. The multi-task sparse group Lasso[7−11] used by this paper