Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 5, No. 1, 2023

Regression Prediction of Effective Temperature of Stellar Spectrum Based on Kernel Method

Xiang Ji, Yadong Wu*, Wei Zhang, Yi Li

Sichuan University of Science and Engineering, China
* Corresponding author: Yadong Wu (Email: wyd028@163.com)

Abstract: There are many factors that lead to the huge difference in stellar spectra, the most important of which are atmospheric physical parameters, namely, effective temperature, surface gravity and chemical abundance. This study is based on the classified star data in LAMOST DR6 with a temperature value of 7500-9000K and a signal-to-noise ratio of S/N greater than 50. The results of the two methods are compared through the regression verification of nuclear least squares regression (KLSR) and nuclear PCA regression (KPCR), The scope of application of the two methods is discussed, and the two most important lick parameters affecting the effective temperature are regressed. By regressing the effective temperature of such stars, a method for predicting the effective temperature of measured spectral data under large samples is given, which provides a certain reference for predicting the trend of star evolution and studying the evolution law of such stars.

Keywords: Stellar spectra, Effective temperature, KLSR, KPCR, Lick parameters.

1. Introduction

In 2000, the Sloan Digital Sky Survey project was officially launched, which is a major milestone in astronomical observation. The data obtained by the project in just a few weeks has exceeded the total data in the history of human astronomy. In exploring the scientific research of astronomical big data, the massive astronomical spectrum contains rich information. The spectral line data includes the chemical composition, brightness, temperature, spatial distribution, composition and evolution history of celestial bodies. The large-area multi-target optical fiber spectral telescope [1] independently designed by China also officially launched the sky survey in September 2012. As of December 2017, after six years of sky survey by LAMOST, the spectral data published by the virtual observatory reached 9017844 spectra, including 8171443 stars, most of which are the spectra of main sequence stars.

The spectral data used in this study is from LAMOST DR6. There are seven types of stars including O, B, A, F, G, K and M. Each data has been normalized, and the logarithm of temperature is used in the regression model_ 10 Teff replaces Teff to reduce the fluctuation of the temperature reference value and better describe the change of the model output. For the error of the experiment, the average of the absolute value of the difference between the model output value of the test set and the actual value is used as a measure to test the regression effect of the model.

A total of 11586 stellar spectral data were used in this experiment, including 1597 Class B stars, 1987 Class A stars, 2009 Class F stars, 2027 Class G stars, 2022 Class K stars, and 1944 Class M stars, of which the sample of Class O stars is too small to be included in the experiment.

In order to test the effect of data noise on the two models, the experiment also used a part of spectral data of stars and galaxies to interpolate the star data.The Figure 1 shows the spectral data of a galaxy in the data used in this study. The blue line is the emission band and the green line is the absorption band.

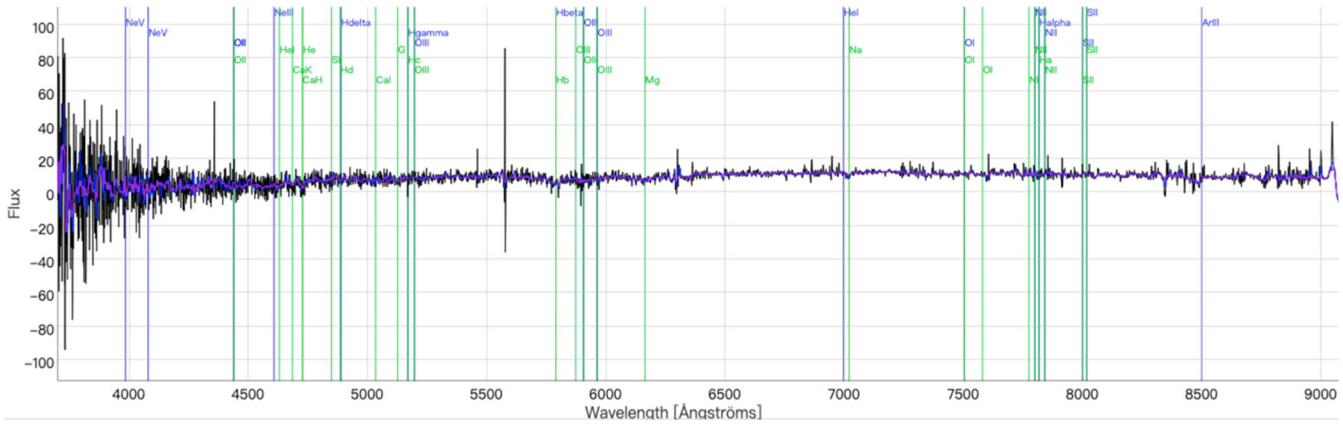


Figure 1. LAMOST DR6 V2 LowResolution Spectral Map of Stellar

2. Nuclear Regression

Kernel method [2] is a classic method to solve nonlinear problems. Its core idea is to use nonlinear mapping to project data into the feature space, and then use linear learners to regress this feature space, as shown in formula (1):

$$f(x) = \sum_{i=1}^{l} w_i \varphi_i(x) + b \quad (1)$$

At this time, the kernel function K satisfies: for all x, z ∈ X, there is: K (x, z)=< φ (x), φ (z)>. Therefore, the form of nonlinear kernel regression model is shown in formula (2):

$$f(x) = \sum_{i=1}^{n} a_i y_i K(x, x_i) + b \quad (2)$$

The nuclear method can be realized in many ways. According to the characteristics of different data, the corresponding expansion method should be adopted. This paper uses two models for stellar parameter regression, namely KLSR [3] model and KPCR [4] model

2.1. Kernel least squares regression

The kernel least square regression (KLSR) model is shown in formula (3), where $K(x, x_i)$is the kernel vector, x_i is the training point data, and the linear regression coefficient of the regression model is calculated by the least square method β:

$$f(x) = \sum_{i=1}^{n} \beta_i K(x, x_i) + \beta_0 \quad (3)$$

Because KLSR is a very common and simple nonlinear kernel regression model, it does not need to do too much data processing compared with KPCR model, and only needs to use a fixed method to calculate the regression coefficient β that will do.

2.2. Kernel principal component regression

Compared with KLSR algorithm, KPCR algorithm is relatively complex, and its main steps are as follows: (1) Star spectrum preprocessing: normalize the data using the maximum and minimum method:

$$y = newMin + (newMax - newMin)\left(\frac{x - Min}{Max - Min}\right) \quad (4)$$

(2) Build KPCR regression model:

a. Calculate the normalized p-dimensional spectral data x through step 1 The kernel matrix K of p, where $K_{ij} =< \varphi(x_i), \varphi(x_j) >= [k(x_i, x_j)]$

b. Centralize the feature space, and the core matrix after centring is:

$$\widehat{K} = K - 1_n K - K 1_n + 1_n K 1_n \quad (5)$$

Where 1_n is the centralization matrix of unit 1.

c. Solving eigenvalue equation n of kernel matrix $n\lambda_k a^k = \widehat{K} a^k$, and standardize a^k as $< a^k, a^k >= 1/\lambda_k$

d. The core principal components of model sample x are as follows:

$$\beta(x)_k = \left(V^k \varphi(x)\right) = \sum_{i=1}^{n} a_i^k \widehat{K}(x_i, x) \quad (6)$$

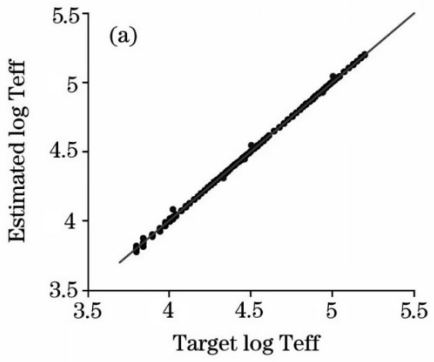
e. Select the first 1 principal components and their corresponding physical parameters to construct the KPCR regression model, and the regression coefficient $w_k, k = 1, \ldots, l$, where l is determined by the least squares.

3. Experiment and Analysis

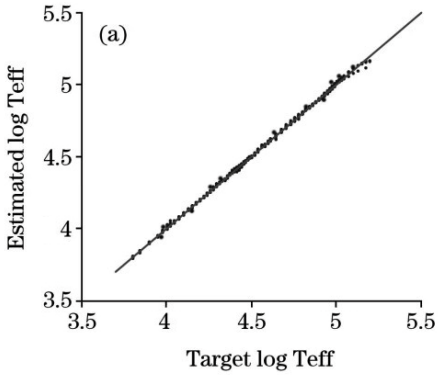
3.1. Kernel regression model

3.1.1. Experiment of kernel regression model without noise

In this study, Gauss kernel function[5] is used to construct KLSR model and KPCR model in noise-free experiment. The final value of nuclear width h is 0.5 by means of comparative experiment. Figure 2 shows the comparison between the calculated parameters and actual parameters of the two models without noise. It is worth noting that when the amount of experimental data is large, the plotted data points will be relatively concentrated. When many data points overlap and overlap, it is not conducive to directly display the local laws and trends of the data and the correlation characteristics between the characteristic values of the line index. Therefore, this paper selects the corresponding proportion of local data sets to fit the regression.



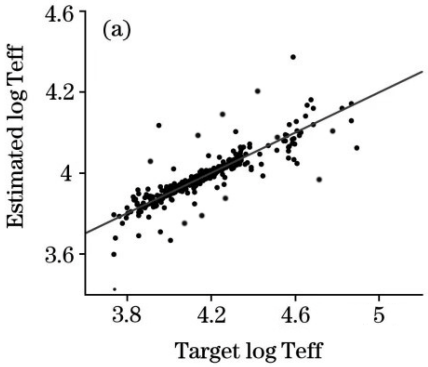
(a) KLSR



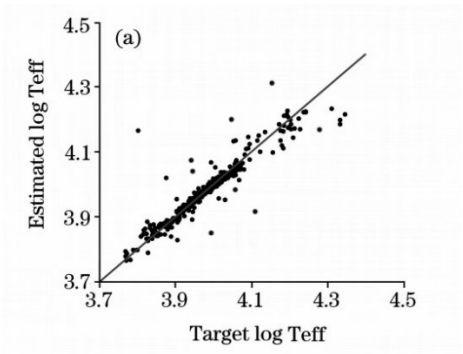
(b) KPCR

Figure 2. Relationship between calculated and actual effective temperature values under KLSR and KPCR models

3.1.2. Experiment of kernel regression model with noise



(a) KLSR



(b) KPCR

Figure 3. Relationship between calculated and actual effective temperature values under KLSR and KPCR models with data noise

3.1.3. Experimental result

Combined with Figure 2 and Figure 3, it can be seen that the KLSR model and KPCR model can better predict the effective temperature of the stellar spectrum in the absence of data noise interference, but the robustness of KPCR is relatively good after adding noise. In addition, due to the nonlinear dimensionality reduction, the training speed of KPCR model will be faster, and this advantage will become more obvious with the increase of spectral data dimension and volume. The Table 1 lists the training indicators of the two methods in both cases.

Table 1. Training indexes of KLSR model and KPCR model in two modes

	absolute error without noise (log Teff)	absolute error with noise (log Teff)	Average training time (h)
KLSR	0.0008	0.0319	6.5
KPCR	0.0056	0.0107	3

It is worth noting that the kernel function of this study uses Gauss kernel, such as support vector (SVM) regression or neural network method to construct the regression function, and the results will show similar results, which has been proved by experiments, and will not be repeated. For the regression of effective temperature, parameters are also very important influencing factors in addition to the model. If the relationship between important parameters and effective temperature can be found, the regression accuracy can be greatly improved. The following will introduce the impact of two key characteristic values on the regression effect.

3.2. Eigenvalue fitting

The characteristics of spectrum can be described by spectral line index, of which the Lick line index is the most extensive. In 1994 Guy Worthy [6] and others gave the complete definition and description of the Lick line index, which gave 26 kinds of Lick line index. This paper selected the most representative kp12 and halpha12 fields to study, and used KPCR model to regress the two fields respectively, and then analyzed the relationship between the Lick line index and Teff, The regression results are shown in Figure 4:

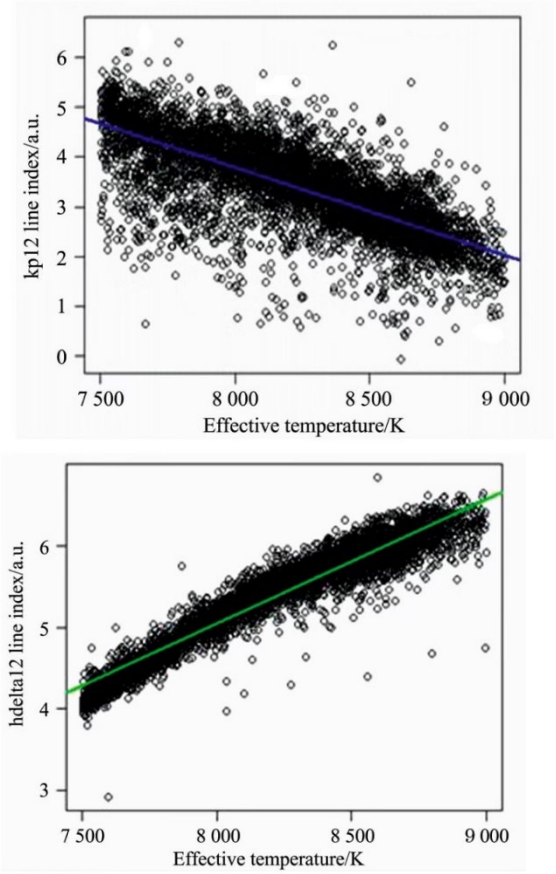


Figure 4. Fitting analysis of correlation between two kinds of Lick line index and effective temperature

From the experimental results, it can be seen that under the condition of large samples, the effective temperature of the star spectrum has a certain correlation with the kp12 field and the halpha12 field. Among them, the outliers of the halpha12 field are less, and the correlation is more obvious. After linear fitting, their respective fitting curves are drawn. The fitting curves of the kp12 field and the effective temperature are shown in the blue line in the figure, and the fitting curves of the halpha12 field and the effective temperature are shown in the green line in the figure, The specific correlation values are shown in the table below. By fitting the curve, the effective temperature value of the spectrum can be roughly calculated from the two field values of the spectrum. The error value depends on the range of the specific value to be corrected, and the specific correction parameters can be calculated by ridge regression and other methods.

Table 2. Correlation value of two kinds of Lick line index and effective temperature

correlation coefficient	Teff	kp12	hdelta12
Teff	1	0.698	0.897
kp12	0.698	1	-0.745
Hdelta12	0.897	-0.745	1

4. Conclusion

In this study, two different nuclear regression models were used to regress the stellar spectral data under different conditions. Through the performance of the model under different conditions, the model is evaluated from the point of view of model fitting degree, training time and other indicators, and regression experiments are carried out on the effective temperature of stars using the appropriate model and the Lick line index of star spectral data as the indicator, and the fitting relationship between the two parameters and the effective temperature of stars is obtained, and the method for predicting the effective temperature of the measured spectral data under large samples is given, Then correctly predict the development curve of future star evolution, and provide a reliable demonstration model for subsequent research and analysis of star evolution laws.

5. Future Work

This research can be improved in the following aspects in the future: first, use more kernel functions to repeat experiments on the model until a kernel function with the best accuracy and efficiency is determined; second, different types of stars may have different applicability to the model, so different types of stars can be used to train the model separately, so as to explore the impact of star types on the model to find the most applicable model; finally, The correlation fitting conducted in this study has a certain error range, which can be reduced by mathematical means such as ridge regression [7] to improve the correlation between Lick line index and effective temperature, so as to better predict the effective temperature of stars.

References

[1] Luo A L, Zhao Y H, Zhao G, et al. The first data release (DR1) of the LAMOST regular survey[J]. Research in Astronomy and Astrophysics, 2015, 15(8): 1095.

[2] Bishop C M. Neural networks and their applications[J]. Review of scientific instruments, 1994, 65(6): 1803-1832.

[3] Rosipal R, Trejo L J. Kernel partial least squares regression in reproducing kernel hilbert space[J]. Journal of machine learning research, 2001, 2(Dec): 97-123.

[4] Rosipal R, Girolami M, Trejo L J, et al. Kernel PCA for feature extraction and de-noising in nonlinear regression[J]. Neural Computing & Applications, 2001, 10: 231-243.

[5] Scholkopf B, Sung K K, Burges C J C, et al. Comparing support vector machines with Gaussian kernels to radial basis function classifiers[J]. IEEE transactions on Signal Processing, 1997, 45(11): 2758-2765.

[6] Worthey G, Faber S M, Gonzalez J J, et al. Old stellar populations. 5: Absorption feature indices for the complete LICK/IDS sample of stars[J]. The Astrophysical Journal Supplement Series, 1994, 94: 687-722.

[7] McDonald G C. Ridge regression[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2009, 1(1): 93-100.

[8] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. science, 2000, 290(5500): 2323-2326.

[9] Madgwick D S, Coil A L, Conselice C J, et al. The DEEP2 galaxy redshift survey: Spectral classification of galaxies at z∼ 1[J]. The Astrophysical Journal, 2003, 599(2): 997.

[10] Singh H P , Gupta R , Gulati R K . Stellar spectral classification based on principle component analysis and artificial neural networks.[J]. Monthly Notices of the Royal Astronomical Society, 1996, 138(2):309.