The Research of Visual Classification for LAMOST Low Quality Spectra

Lipeng Bi[1,a], Jingchang Pan[1, b*], Cong Liu[1,c] and Guozhou Ge[1,d]

[1]School of Mechanical Electronic and Information Engineering, Shandong University, Weihai, China

[a]bilipeng0418@126.com, [b]pjc@sdu.edu.cn, [c]onionsheep@gmail.com, [d]geguozhou@139.com

*Corresponding author

Keywords: LAMOST; Interactive platform; Template matching; Spectral classification

Abstract. Massive sky survey data are produced by the LAMOST large-scale survey project. However, the low quality spectra still accounted for about half of total LAMOST observed data at present. For these low quality spectra, spectral classification can not only use of computer technology, but also be combined with artificial view. Therefore, this paper focused on interactive spectral classification design and implementation.

Introduction

Since ancient times, astronomical research has never stopped, and has reached its pinnacle in the support of modern science. At present, there are many large aperture with wide field of view survey telescopes in the world like the Sloan Digital Sky Survey (SDSS) Telescope of America[1,2], the Anglo-Australian Telescope of Australia[3,4] and the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) of China[5,6], etc.

On December 30, 2014, the data set of LAMOST data release two (DR2)[7,8], including spectra of the pilot survey and spectra of the past two years of the regular spectroscopic survey, is published to domestic data users and foreign partners. Although the DR2 release large amounts of spectra, but it contains a lot of low quality spectra which exhibit obvious quality defects, such as strong noise, unobvious spectrum characteristic, low local SNR, abnormal continuum, abnormal connection of two bands, sky subtraction abnormality, etc. According to these low quality spectra, interactive spectral classification research can fully improve the spectra utilization.

By extracting the spectral characteristics and using the template matching based on Euclidean distance algorithm, this paper helps users to complete interactive spectral classification.

Data preparation

Experimental spectra data. The experimental data is from the LAMOST DR2 data set released on December 2014. The DR2 totally contains 4,136,482 spectra, including 909,520 spectra of pilot survey and 3,226,962 spectra of regular survey. In addition, the DR2 contains the atmospheric parameters of 2,207,788 stars, which becomes the largest stellar spectral parameters catalog in the world at present.

Template spectra data. The template library used in the paper is the stellar spectral template library created by Dr. Wei Peng[9] form National Astronomical Observatory of China in 2014 for LAMOST spectra, which includes O, B, A, F, G, K, M, Other eight super classes and 183 detailed subtypes. This template library is by far the most detailed and the most completed stellar template library for LAMOST. The important role for the template library is to provide a referable standard template for spectral classification.

Template spectrum feature extraction

1. Template spectrum normalization

There is only relative flux calibration in LAMOST spectra, so there is no sense to compare absolute flux. For the unification and accuracy of template matching, we normalize the flux of template spectra at first, normalized to [0, 1]. Finding the maximum f_{max} and the minimum f_{min} in all spectral flux points, then processing every flux point f_i to make it become $(f_i - f_{min}) / (f_{max} - f_{min})$ after normalization. After that, all flux points are within the interval [0, 1], which not only ensure that every flux value is not negative, but also unify dimension to provide the basis for template matching.

2. Continuous spectrum feature extraction

We use the method of least square polynomial fit[10] to extract the continuous spectrum characteristics.

For every template spectrum, the fitting polynomial is the continuous spectrum. If the polynomial

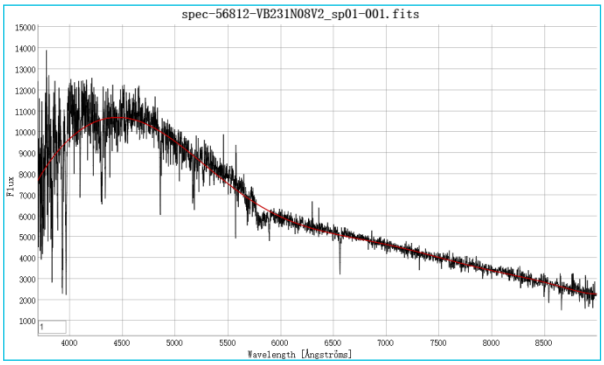

Fig.1. Continuous spectrum from fitting ten-order polynomial

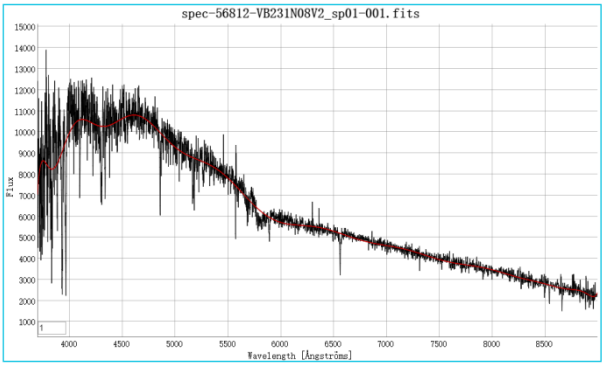

Fig.2. Continuous spectrum from fitting twenty-order polynomial

has fixed number of terms, the coefficients represent its characteristics and shape. The coefficient with higher power item represents the more obvious characteristic. So we can take the coefficients of the polynomial to represent the continuous spectrum characteristics.

From Fig.1 and Fig.2, we can know that the imitative effect is better when $n = 20$. In the paper, we take the 20 coefficients of polynomial as the spectral characteristics of continuous spectrum. We set an array A, each of which represent the spectral characteristics. There are 20 elements in array A so far.

3. Spectral lines feature extraction

Every astronomical spectrum contains many spectral lines, which are frequently used to measure the atmospheric physical parameters (effective temperature, surface gravity, and chemical abundances). These spectral lines containing important characteristics, can provide the basis for spectral classification. For LAMOST spectrum, we select total 40 lines which range from 3869.78Å to 8664.52Å and contain many special element lines, such as H line, He line, Na line, Ca line and Mg line, et al., to represent spectral features.

Take a Ca line as an example, its center wavelength λ is 8544.44Å. Due to the reasons of radial velocity and spectral redshift in LAMOST spectra, we can not simply take the flux corresponding to center wavelength as characteristics of the Ca line. In order to extract the spectral characteristics, we select a certain wavelength range near the center of [λ-3, λ+3] as the spectral bandpass range. All flux values within the spectral bandpass range together represent characteristics of the Ca line. We tried to use the mean value of these flux values to indicate this spectral line, but found that the result is not ideal. The mean flux value can not accurately represent the shape and peak of these flux

points. So we learn from the method of continuous spectrum feature extraction to use a third-order polynomial to fit points within the bandpass range. The spectral line is characterized by three polynomial coefficients.

For each spectral line, we get three parameters to represent the spectral characteristics. We obtain 120 parameters from the selected 40 spectral lines according to the size of the wavelength selection, which abundantly and vividly characterize the shape and peak of template spectrum in the vicinity of the spectral line. Thus far, the array A showing the characteristics of template spectrum include 140 elements. In order to better display, the spectral name is pushed in the array A as the 141 element.

In order to facilitate reading, the result of 183 template spectra feature extraction is stored on the server in JSON format. The JSON data is a two-dimensional array of 183 elements, each element is an array of 141 elements.

Experimental spectrum feature extraction and template matching

Experimental spectrum also needs to feature extraction, and feature extraction process is the same as the template spectrum feature extraction process. We use min-max normalization, and extract continuous spectrum features and 40 spectral lines features. Experimental spectrum is represented by an array of 140 elements.

Matching experimental spectrum and template spectrum is completed by Euclidean distance metric algorithm, which result is the similarity of experimental spectrum and template spectrum. The smaller resulting value represents the more similar experimental spectrum and template spectrum.

For the experimental spectrum and all template spectra we want to measure, we calculate Euclidean distance with their array of features. The obtained 183 results can be considered as the degree of similarity with the experimental spectrum and 183 template spectra. These results are stored in an array *resultArr*. Each item of array *resultArr* is an object. Each object has three attributes, namely *resultArr[i].index*, *resultArr[i].name* and *resultArr[i].distance*. Wherein, the *resultArr[i].index* represents the position of the spectral template in the template library, the *resultArr[i].name* indicates the template spectral category name, the *resultArr[i].distance* shows the result of distance measurement. We select two elements with minimum distance in the resulting array. They are spectral type most similar to experimental spectrum. Reading the two matching template spectra to get the original spectral wavelength and flux, display the two template spectra on the page, to provide intuitive basis of visual classification for users.

For example, the experimental spectrum *spec-56812-VB231N08V1_sp01-053.fits*, its header information displays that *CLASS = 'STAR'* and *SUBCLASS = 'G0'*, so this spectrum is likely to the ‘G0’ type star. The result of our template matching is shown in figure 3.

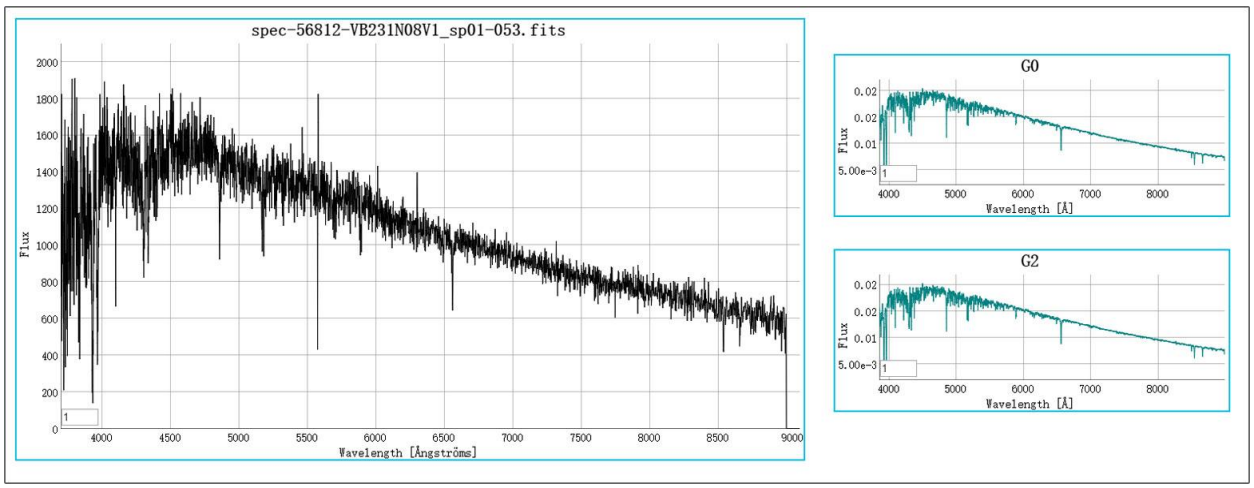

Fig.3. The result of template matching

There are experimental spectrum (left) and matching two template spectra (right) in Fig.3. The similar matching result is ‘G0’ type and ‘G2’ type, its shows that the initial classification result is correct and the spectrum is the ‘G0’ type star. For these spectra which initial classification is not correct, we can revise the classification result by the method of template matching. The result of artificial classification is saved in database, for convenient user view.

Summary

In view of the LAMOST low quality spectra, this paper design an interactive platform to improve the low quality artificial spectral classification. We normalize template spectra and experimental spectra to unify dimension at first. Then we extract the continuous spectrum characteristics by polynomial fitting, extract characteristics of the specific spectral lines, and we take the continuous spectrum characteristics and the spectral lines characteristics as the spectral characteristics. At last, we measure distance through Euclidean distance metric algorithm, select and display the two most similar template spectra, to provide a reference for users with interactive artificial spectral classification.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (U1431102).

References

[1] G. Bruzual, S. Chariot. Stellar population synthesis at the resolution of 2003[J]. Monthly Notices of the Royal Astronomical Society, 2003,344: 1000-1028.

[2] Donald G. York, et al. The Sloan Digital Sky Survey: Technical Summary[J]. The Astrophysical Journal, 2000,120: 1579-1587.

[3] S. C. Ellis, et al. Suppression of the near-infrared OH night-sky lines with fibre Bragg gratings - first results[J]. Monthly Notices of the Royal Astronomical Society, 2012, 425: 1682-1695.

[4] D. A. Fischer, Jeff Valenti. The Planet-Metallicity Correlation[J]. The Astrophysical Journal, 2005, 622: 1102-1117.

[5] Xiang-Qun Cui, et al. The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST)[J]. Research in Astronomy and Astrophysics, 2012,12: 1197-1242.

[6] Ali Luo, et al. Data release of the LAMOST pilot survey[J]. Research in Astronomy and Astrophysics, 2012,12: 1243-1246.

[7] Zhang B, Chen X Y, Liu C, et al. Member candidates of the star clusters from LAMOST DR2 data[J]. arXiv preprint arXiv:1506.04222, 2015.

[8] Zhao J K, Zhao G, Chen Y Q, et al. Halo stream candidates in the LAMOST DR2[J]. Research in Astronomy and Astrophysics, 2015, 15(8): 1378.

[9] Peng Wei, Ali Luo, et al. On the construction of a new stellar classification template library for the LAMOST spectral analysis pipeline[J]. The Astronomical Journal, 2014, 147(5): 101.

[10] Steinier J, Termonia Y, Deltour J. Smoothing and differentiation of data by simplified least square procedure[J]. Analytical Chemistry, 1972, 44(11): 1906-1909.