Automatic Identification of LAMOST Spectra with Continuum Problem Based on High Performance Computing

Jingjing Yu, Jingchang Pan* and Fanlong Meng
School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China
Email: yujingjing2018@163.com, *pjc@sdu.edu.cn, morphy8@163.com

Abstract. Continuum problem is a phenomenon that the continuum of spectra get off their actual continuum even break off due to interstellar extinction and flux calibration, and it will have negative impact on the subsequent process such as spectral line extraction and so on. Based on this problem and fully considering of the continuum features of the stellar spectra, a method of automatic detection and recognition of the continuum problem in the stellar spectra based on High Performance Computing (HPC) is proposed, which will improve work efficiency greatly compared with the traditional human eyes examination under the condition of maintaining high accuracy. Continuum template matching is used to identify the continuum problem spectra in this paper. The first step goes to fit continuum of the test stellar spectra and the template spectra, then the flux differences of the two continuum we fitted before at every point of wavelength in the continuum spectra will be calculated based on full spectra matching to analyze the features of its distribution. The features we count are average (called β) value and standard deviation (called δ). The percentage of points distributed in range β±α*δ will be detected to confirm that if there is continuum problem. Experiments was taken to identify the continuum problem spectra on HPC, which demonstrate the validity and efficiency of the method, and this has important reference value to resolve similar massive spectra data processing problems in the future.

1. Introduction
The Large Sky Area Multi-Object Fiber Spectroscopic Telescope [1] (LAMOST) is a National Major Scientific Project built by the Chinese Academy of Science. Since the launch of the formal pilot survey, LAMOST has produced tens of thousands of spectra with each observation, and there are a lot of continuum problem spectra. Continuum problem spectra have negative impact on the subsequent process such as spectral line extraction and so on. The continuum problem spectra were selected by human eyes before, which was time-consuming and labor-intensive and reduced the efficiency and accuracy of spectra analysis. However, it is almost impossible to find the automatic detection of the continuum problem spectra. A method of automatic identification of LAMOST continuum problem spectra based on High Performance Computing (HPC) is proposed after studied the continuum of types of stellar. The flux differences of the two continuum we fitted before at every point of wavelength in the continuum spectra will be calculated based on full spectra matching to analyze the features of its distribution. The features we count are average (called β) value and standard deviation (called δ). The percentage of points distributed in range β±α*δ will be detected to confirm that if there is continuum problem.

Fast computation and accurate results are the most outstanding advantages of high performance computing platform compared to traditional computing, and the efficient computing power relies on multi-core computing. Python is used in this paper to realize the parallel computing on HPC by calling...
MPI, a plurality of FITS files is divided into subtask that can be executed in parallel across arbitrary numbers of processor cores. And then each processor core is executed in a separate order, the root process is specified to collect the result of each core after the last one is completed, thereby improving the computational efficiency. The experiments on the HPC show that the method we proposed in this paper can achieve high efficiency and accuracy on the identification of continuum problem spectra.

2. The Data
LAMOST has published these spectra online and perform its second data release (DR2) online. In this data release, over 4 million FITS (Flexible Image Transport System) file are published on-line. FITS is a data format for data exchange between different platforms which is used to store images. Each spectrum corresponds to a FITS file. The flux and wavelength information are exist in the data unit of FITS, and some information can be read from FITS directly such as flux, COEFF0 (central wavelength of first pixel), COEFF1 (Log10 dispersion per pixel), NAXIS1 (the number of wavelength array), while the wavelength should be calculated by formula (1), In the formula, “i” is the numbers of sampling points of wavelength, which is in the region of [0, NAXIS1].

\[
\text{wavelength} = 10^{\text{COEFF0} + i \times \text{COEFF1}}
\]  

(1)

Until July 2014, LAMOST has completed its pilot survey which was launched in October 2011 and ended in June 2012, and the first two years of its regular survey which was initiated on September 2012. LAMOST is a special reflecting Schmidt telescope with 4000 fibers in a field of view of 20 deg2 in the sky. After three-years-survey, it totally obtains 4,136,482 spectra, which consist of stars, galaxies, quasars and other unknown type.

3. Method Introduction

3.1 Stellar Spectra Classification Based on Lick Indices
Lick \cite{2} index is a relatively wide spectral characteristics. Each index in the absorption line with the most prominent, it is named. Lick is an indices that used by astronomers as a line feature. It is simple and feasible to classify stellar spectra by calculating the Euclidean distance of lick indices: on the one hand, 25 dimensional feature matrix can be got by lick indices, which greatly reduces the dimensionality of data and reduces the amount of calculation. On the other hand, lick indices is calculated by the average value of flux so that they have a higher signal-to-noise ratio(S/N) \cite{3} than the original spectra.

3.2 Zero Flux Spectra Detection

![Figure 1](spec-55916-B5591606_sp06-038.fits)

Figure 1 The spectrum which has zero flux at continuity points.
There are many zero flux spectra in continuum problem spectra, as shown in Figure 1 which is the most obvious problem. In this case, not only the continuum fitted inaccurate but the detection efficiency is reduced. To this end, we eliminate the zero flux spectra before all the work. The definite means are as follows:

1. Set a window size of \( \text{window} \) = 10;
2. Traverse flux values at all wavelength points of the entire spectrum with the window. If the flow at a certain point is zero, then the spectrum is judged to be zero flux spectrum if all the points in the whole window are 0.

### 3.3 Fitting the Continuum

There are many methods to fit continuum such as polynomial fitting, wavelet filtering \(^4\), the median filtering \(^5\) and morphological filter and so on. In this paper, we use the method of fitting continuum from Pan Jingchang and Wei Peng, which is mentioned in “A new method for automatic identification of stellar spectra of emission lines \(^6\). The method is improved on the basis of the polynomial approximation method, when some of the points are selected not all of the points when fitting the continuum, which makes the influence of the wide line and the strong line smaller. To prepare for the continuum fitting, something has been done to deal with the original spectral flux array of stellar spectra in this paper, such as normalizing \(^7\) the flux to region of \([0, 1]\) and flux interpolation. The all steps to fit continuum are as follows:

1. Take \( w(n) \) and \( f(n) \) as the original spectral wave and flux of a stellar spectrum respectively;
2. For test spectra, the flux are interpolated to the wavelength of the template spectra, and the flow value of the test spectra at each wavelength point of the template spectra is obtained, which is \( f_x(n) \);
3. Normalize the flux of \( f_x(n) \), and \( f_x(n) \) as the normalized spectrum;
4. Compute the mean flux \( f_x(n) \) and standard flux \( f_x(n) \) of the normalized flux \( f_x(n) \) with the window size of 10;
5. Take \( f_x(n) \) as the points which in the region \([-0.5, f_x(n) + f_x(n), 0.5, f_x(n) + f_x(n)]\);
6. The continuum of the original spectrum is got by six order polynomial fitting to \( w(n) \) and \( f_x(n) \);

To make the test spectra and template spectra are in the same measurement and preserve the uniqueness of variables, we interpolate the flux of test spectra to the wavelength of the template spectra and normalize the flux of spectra. The comparison of flux between test spectrum and its template spectrum of G5 without flux normalization is shown in Figure 2(a). The measurement of flux is vary greatly in this case. The flux of the two spectrum after flux normalization are shown in Figure 2(b). From these figure, we can know that flux normalization is a crucial step for the continuum template matching.

### 3.4 Continuum Template Matching

The template spectra we used in this paper is 183 template spectra which is from mentioned in “On the Construction of a New Stellar Classification Template Library for the LAMOST Spectral Analysis Pipeline\(^8\)” by Wei Peng. The continuum of test spectra and template spectra are obtained after steps above. The method based on distance measure to detect continuum problem spectra. The steps of continuum template match method are as follows:

1. Take \( f(1) \) , \( f(2) \) as the continuum of test spectra and template spectra after steps in part C;
2. Calculate the distance between \( f(1) \) and \( f(2) \), we define the distance as an array \( Fr = f_1 - f_2, x = 1,2, \ldots, n \) which represents the flux differences at each wavelength point, \( n \) is the dimension of fits files which means the numbers of wavelength points;
3. Calculate the mean value \( M \) and variance \( S \) of the array \( Fr \), \( \beta \) is defined as the threshold of flux differences;
(4) The test spectrum is defined as continuum problem spectrum if one of the following three conditions is satisfied:

a) Get the number $N$ of the flux differences at a certain wavelength not in the region $[\mu - \alpha \sigma, \mu + \alpha \sigma]$, if $N$ is greater than 1 and $N$ accounts for 5% or more of the total wavelength points, then this spectrum has continuum problem;

b) If one of the value in the array $F_r$ is greater than $\beta$, then this spectrum has continuum problem;

c) Get the number $I$ of the intersections of the two continuum, if $I$ is greater than or equal to 3, then this spectrum has continuum problem.

Figure 2 The comparison of flux. (a) this is the flux comparison of test spectrum and template spectrum without flux normalization. (b) this is the flux comparison of test spectrum and template spectrum after flux normalization.
In order to determine the continuum problem spectra, the selection of $\alpha$ and $\beta$ determines the accuracy of the continuum spectra detection. After the experiments, the value of $\alpha$ is 1.5, and the value of $\beta$ is 0.0025 after normalizing the flux to the same measure. There are some spectra with low signal-to-noise ratio which have the smaller measure than the others that leads to the spectra with continuum problem cannot be detected by the two thresholds previous. In order to ensure that all the continuum problem spectra can be detected, step c came into being.

4. Experiments Based on HPC and results analysis

Python is used in this paper to realize the parallel computing on HPC by calling MPI, a plurality of FITS files is divided into subtask that can be executed in parallel across arbitrary numbers of processor cores. And then each processor core is executed in a separate order, the root process is specified to collect the result of each core after the last one is completed, thereby improving the computational efficiency.

A file with 33,242 FITS files’ data is used as the sample file of the test experiments on HPC, 60 processing cores are used to identify the continuum problem spectra. Processing core number increase from 8 to 60 linearly, and the changes of processing time was shown in TABLE.1 to TABLE.4. And the trend of processing time on HPC is shown in Figure.3.

![Figure 3 The trend of processing time.](image)

| Table 1 | The changes of processing time on HPC |
|---------|-------------------------------------|
| cores  | 8        | 10        | 12        | 14        | 16        | 18        |
| time(s) | 3238     | 2787      | 2490      | 2015      | 1816      | 1538      |

| Table 2 | The changes of processing time on HPC |
|---------|-------------------------------------|
| cores  | 20        | 22        | 24        | 26        | 28        | 30        |
| time(s) | 1446      | 1321      | 1234      | 1082      | 973       | 944       |

| Table 3 | The changes of processing time on HPC |
|---------|-------------------------------------|
| cores  | 32        | 34        | 36        | 38        | 40        | 42        | 44        | 46        |
| time(s) | 908       | 856       | 792       | 768       | 720       | 702       | 660       | 647       |
Table 4 The changes of processing time on HPC

| cores | 48  | 50  | 52  | 54  | 56  | 58  | 60  |
|-------|-----|-----|-----|-----|-----|-----|-----|
| time(s) | 616 | 590 | 558 | 529 | 521 | 498 | 486 |

Some experiments have been done on the tow platform to identify continuum problem spectra from 33,242 FITS files, the single-machine environment and HPC. The experiments are carried out on the two platforms, and the experimental results show that the continuum problem spectra identified are the same, however, the processing time quite different between the two platforms. In the single-machine environment, it is take 7 hours, 51 mins for the experiment, and the processing time on HPC is 8 mins, 6 secs, which greatly improves the processing efficiency.

Figure 4 The original spectrum of the continuum problem spectrum, and the distribution of flux differences. The blue curve shows the differences between the two continuum spectra of the original spectra.

A continuum problem spectrum of F5 type which is identified by the method we proposed in this paper is shown as Figure 4. The comparison of the continuum of test spectra and template spectra is shown in the subgraph above of Figure 4. The distribution of the flux differences is shown in the following subgraph of Figure 4. The region \([M - \alpha S, M + \alpha S]\) is identified by three red solid lines, and the thresholds \(\beta\) is identified by two green dashed lines.

The type of the spectrum shown in Figure 4 above is F9, which is the late-type star. The feature of the F9 type spectrum is obvious, an absorption bands \([9]\) exist at 4300Å, and the peak value appears at about 4500Å. The red end of the spectrum is smooth, and the metal line begin to strengthen. However, the feature of the continuum problem spectrum above is different from F9’s. The spectrum of the red end is obviously deviated from the end of the spectrum, and the blue end is raised, which makes the whole continuous spectrum appear too high in the blue end and too low in the red end. According to the Figure 4 of distribution of flux differences, it is obvious that the distribution of the test spectrum and template spectrum is different. And a numbers of flux differences are found greater than \(\beta\), which causes the continuum problem.

The experimental results show that the method we proposed in this paper to identify the continuum problem spectra is simple and effective. And the parallel computing on HPC has performed a number of unique advantages in dealing with large amounts of FITS files, which has saved a substantial amount of time.
5. Conclusions
A method of automatic detection and recognition of the continuum problem in the stellar spectra based on HPC is proposed in this paper. The main work is divided into the following three parts:

1. Fit continuum. The continuum spectra of the test spectrum and the template spectrum with corresponding type are extracted and then we do the process of flux interpolation and normalization in order to analyze data by means of the non-dimensional contrast.

2. Continuum template matching. In order to make continuum template matching for the test spectrum, the flux differences of the test stellar spectra and template spectra at every point of wavelength in the continuum spectra will be calculated to analyze the features of its distribution. The features we count are average (called $\beta$) value and standard deviation (called $\delta$). The percentage of points distributed in range $\beta \pm \alpha*\delta$ will be detected to confirm that if there is continuum problem.

3. Confirm the subclass of the test stellar spectra. The subclasses of most continuum problem spectra are not defined. The prerequisite of continuum template matching is to determine the subclass of the test spectrum. In this paper, we assign the stellar type of the template spectrum with the minimum distance to the test spectrum by calculating the Euclidean distance between the two lick indices of the template spectrum and the test spectrum.

4. Automatic identification of LAMOST continuum problem spectra based on high performance computing. We studied parallel computing method based on HPC using python, which by calling MPI to divide many FITS files into a number of processes. The result of each processes will be collected by root when the last processes was completed.

To sum up, the parallel computation method of the automatic recognition of the continuum problem spectrum is realized on the high performance computing platform. It achieves more rapid and efficient recognition effect compared to the single computer.

6. Acknowledgment
This work was supported by the National Natural Science Foundation of China (U1431102).

7. References
[1] BAI Zhongrui. LAMOST 2D pipeline. InteNational Astronomical Union. 2012.
[2] M Franchini, C Morossi, PD Marcantonio, ML Malagnini, M Chavez. The Lick/SDSS Library. II. [Ca/Fe] and [Mg/Fe] in F, G, and K Stars from SDSS-DR7. Astrophysical Journal, 2011, 730(2):117
[3] Franchini, M., Morossi, C., Di Marcantonio, P., Malagnini, M. L., & Chavez, M. 2011, ApJ, 730, 117
[4] [Starck J.L., R. Siebenmorgen, R. Gredel. The astrophysical journal, 1997, 482: 1011.
[5] J.P.Fitch, E.J.Colye and N.C.Gallagher. Median Filtering by Threshold Decomposition. IEEE Trans. ASSP. Vol.ASSP-32, 1183-1188, 1984
[6] Jingchang Pan, Caiming Zhang, Peng Wei, et al. A New Automated Method to Identify Emission Line Star from Massive Spectra (in Chinese), Spectroscopy and Spectral Analysis, 2012, 32(6): 1689-1693.
[7] JIANG Bin, PAN Jing-chang, GUO Qiang. Journal of Electronic Science and Technology of China, 2007, 36(006): 1292.
[8] P Wei, A Luo, Y Li, L Tu, F Wang. On the Construction of a New Stellar Classification Template Library for the LAMOST Spectral Analysis Pipeline,147(5):101,2014
[9] Jacoby G.H., D.A. Hunter, C.A. Christian. The Astrophysical Journal Supplement Series, 1984, 56: 257-281.