The Automatic Recognition of the Abnormal Sky-subtraction Spectra Based on Hadoop

An An and Jingchang Pan*
School of Mechanical Electronic and Information Engineering, Shandong University, Weihai, China
Email: anran0708@163.com, *jingchangpan@163.com

Abstract. The skylines, superimposing on the target spectrum as a main noise, If the spectrum still contains a large number of high strength skylight residuals after sky-subtraction processing, it will not be conducive to the follow-up analysis of the target spectrum. At the same time, the LAMOST can observe a quantity of spectroscopic data in every night. We need an efficient platform to proceed the recognition of the larger numbers of abnormal sky-subtraction spectra quickly. Hadoop, as a distributed parallel data computing platform, can deal with large amounts of data effectively. In this paper, we conduct the continuum normalization firstly and then a simple and effective method will be presented to automatic recognize the abnormal sky-subtraction spectra based on Hadoop platform. Obtain through the experiment, the Hadoop platform can implement the recognition with more speed and efficiency, and the simple method can recognize the abnormal sky-subtraction spectra and find the abnormal skyline positions of different residual strength effectively, can be applied to the automatic detection of abnormal sky-subtraction of large number of spectra.

1. Hadoop Platform
Hadoop\cite{1} is an open source distributed computing platform under the Apache Software Foundation, The core of Hadoop consists of HDFS distributed file system and MapReduce programming model. The HDFS distributed file system is mainly used to solve the problem of large-scale data storage, and the MapReduce programming model is used to solve the problem of distributed parallel computing of large-scale data. The underlying details of the Hadoop are transparent to the user, and we can form a distributed system by making use of a lot of cheap, low - profile minicomputers. The applications are deployed on each node to achieve parallel processing, speed up the processing speed and expand the storage space, so as to be more effective to complete the processing of massive data.

The HDFS cluster is based on the Master/Slaves model, which takes the block as the storage unit, and the default size of each block is 64MB. The figure 1 is the HDFS Structure. From this figure, we can see that Hadoop is mainly composed of Client terminal, NameNode, SecondaryNameNode and DataNode. For Hadoop version 1, a Hadoop cluster has only one NameNode. NameNode is running on the master node of HDFS, and responsible for the management of the namespace about the file system and recording the position in the DataNode of all data blocks and its replication. This information is stored on disks as the HDFS metadata image files and HDFS file change logs. The SecondaryNameNode is used as the slave date node to monitor the HDFS auxiliary background process, and regular merge the HDFS metadata image files and the HDFS file change logs to the disks.

The DataNode is actually used to store the data and also responsible for serving read and write requests about the data block, and through the heartbeat mechanism to communicate with NameNode, to tell NameNode its storage capacity and the current storage of data block information, etc.
Figure 1 HDFS Structure[1]

Figure 2 MapReduce Structure[1]

MapReduce is a distributed programming framework, which is mainly composed of JobTracker, TaskTracker and Task. The figure 2 is the MapReduce Structure. The JobTracker, as the MapReduce background process, is running on the master node, and responsible for monitoring resources, scheduling jobs. Such as it will monitor the state of TaskTracker, and transfer the failed tasks to other nodes. The TaskTracker is running on the slave node, for receiving the commands sent by JobTracker, such as starting a new task, killing the failed tasks, etc. In addition, it will regularly report to the JobTracker according to the Heartbeat mechanism about the implementation of tasks and the current storage of data block information, etc. Task is activated by TaskTracker, it is the executor of the task.
One task is generally divided into a number of Map Tasks for the implementation of map() function, a Reduce Task to perform reduce() function.

2. Abnormal Sky-subtraction Recognition algorithm

LAMOST\(^{2,3}\) is one of the national key scientific projects, which is approved by the state, began in 2001, taking 7 years, costing $235 million to establish, and is China's largest astronomical telescope. Its site located in the National Astronomical Observatory of the Chinese Academy of Sciences. LAMOST contains 16 spectrometers, you can simultaneously observe 4000 target objects, each of the spectrograph will be connected to 250 fibers, and we use 240 fibers to observe the target objects, and the other 10 optical fibers are used to observe the skylight.

The skylight refers to the sky conditions without month and city light interference. When the sun falls 18 degrees below the horizon, the night sky will show a faint diffuse light, we generally call this as sky background or skylight in astronomical observation. The more common skylight emission line are oxygen atomic emission line \([\text{OI}5577\text{Å}]\) and the OH emission band. The LAMOST process uses this limited skylight fibers to fit a "super skylight" spectrum, and then the spectral data of each observation spectrum is subtracted from the corresponding "super skylight" spectral data to get the sky-subtraction\(^4\) spectral data. On account of the limited technology, the fitted "super skylight" is not a complete representation of the sky background component contained in the target spectrum, this led to the accuracy of the sky is not too high and generated a large number of abnormal sky-subtraction spectra.

Through the analysis of the abnormal sky-subtraction spectra, we can get the anomaly as shown in Figure 3, and the upper line shows the original spectrum, and the middle line represents the fitting of the super skylight spectrum, the lower line represents the spectrum after sky-subtraction processing. It can be seen from the diagram that the spectrum after the sky-subtraction processing is not a relatively smooth curve, but there is uneven band, we can take this phenomenon as some sky-subtraction residual components.

By analyzing the characteristics of the abnormal sky-subtraction spectra, we can know that when identifying the abnormal sky-subtraction spectra, we only need to identify the residuals which are similar to the emission line or the absorption line. So we use continuum normalization method to remove the components of the continuum spectrum, leaving only the characteristic spectral line information and noise information. The continuum normalization is achieved by fitting the spectral continuum, and then the original spectral flow is divided by the continuous spectral flow at the corresponding wavelength, finally, only the spectral lines and noise information remained. As a nonlinear filtering algorithm, median filtering algorithm\(^5\) can effectively remove the transient pulse noise of positive or negative value, and preserve the shape of the continuous spectrum. When the skylight is abnormal, there will be more or less positive and negative pulse noise. In addition, the median filtering algorithm is simple and feasible, we just need to select the appropriate window size to process the algorithm. Therefore, the use of median filtering algorithm will be better suited to carry out the continuous spectrum fitting.

![Figure 3 Abnormal sky-subtraction features\(^6\)](image)

In this paper, we first carry out the continuum normalization, and then use the following algorithm to process the recognition of the abnormal sky-subtraction spectra:
Assume that the number of sky emission lines selected is \( m \), the central wavelength of each sky emission line is \( \text{skyline}[i], i = 1, 2, 3, \cdots, m \), the length of the spectrum is \( n \), and the starting pixel position which gained in the FITS files is \( \text{start} \), the spacing between adjacent pixels is \( \text{bata} \), the spectral flow array is \( \text{flux}[j], j = 1, 2, 3, \cdots, n \) after the continuous normalization, and the wavelength array is \( \text{wave}[j], j = 1, 2, 3, \cdots, n \), now we assume \( \theta_1 \) as the calculating threshold and \( \theta_2 \) as the sky line window threshold.

1. According to the formula \( \log_{10}(\text{skyline}[i] - \text{start})/\text{bata} \), we calculated the subscript position of the skyline wavelength in the wavelength array. Here we assume the position is \( \text{skyIndex} \).

2. The average flow \( \mu \) and standard deviation \( \sigma \) in the wavelength range \( \text{wave[skyIndex} - \theta_1], \text{wave[skyIndex} - \theta_2] \) and the wavelength range \( \text{wave[skyIndex} + \theta_1], \text{wave[skyIndex} + \theta_2] \) of each wavelength are calculated near each skylight emission line after the continuum normalization. As for the selection of the threshold \( \theta_2 \) and \( \theta_2 \), it is determined by the intensity and the width of the skylight, so as to make the two intervals in the relatively smooth range between the two sides of the skylight.

3. Between the wavelength range \( \{\text{wave[skyIndex} - \theta_2], \text{wave[skyIndex} + \theta_2]\} \):
   a. If the continuum normalization flux on the position of this skyline \( \text{flux[skyline[i]} > u] \), in the interval to find whether there is a wavelength \( \lambda \), so that the establishment of \( \text{flux[\lambda]} > \mu + \theta_2 \), If it exists, it is judged that the skylight position is abnormal.
   b. If the continuum normalization flux on the position of this skyline \( \text{flux[skyline[i]} < u] \), in the interval to find whether there is a wavelength \( \lambda \), so that the establishment of \( \text{flux[\lambda]} < \mu + \theta_2 \), If it exists, it is judged that the skylight position is abnormal.

The threshold \( \theta_2 \) is chosen to reduce the influence of the LAMOST wavelength calibration error, which makes it possible to find the abnormal skylight positions in a certain range near the wavelength of the skylight. In this experiment, we set \( \theta_1 \)'s value to 2. The threshold \( \theta_2 \) indicates the degree of deviation of the skylight flow relative to the whole flow, In this experiment, we set \( \theta_2 \)'s value to 3.

3. Hadoop Cluster Processing

3.1 Pretreatment

In this paper, the experiment uses 226000 data which are released by the DR1 from LAMOST, and the data is stored in fits format files. The original spectrum file is composed of several smaller Fits files, Hadoop as a large data distributed processing platform, when the size of the file is much smaller than the size of a single block, its performance will be greatly reduced. We need to do the data preprocessing, the small files will be integrated into large files which are more suitable for the Hadoop platform to deal with, and the information needed in this paper will be extracted, irrelevant spectral information will be filtered out at the same time. In this paper, each of the Fits spectral files are treated as follows: First of all, through the Java program to extract MJD, PLANID, SPID, FIBERID data values from the Fits files, and write them to the text files, The information uniquely identifies the different spectral data. Secondly, we need to extract the initial pixel value as CRVAL1, the distance between adjacent pixels as CD1_1 and the Spectral length, and write them to the text files. Finally, the Java program in turn to find the wavelength and flow value of each pixel according to the information, and then write them to the text files, too. These values are separated by spaces, each line represents the data information of a spectral file. In this experiment, this total size of the fits file is 35.2G before the pretreatment, after the pretreatment, this size is compressed to 15.6G. The size of the uploaded file can be greatly reduced by the pretreatment step.
3.2 Upload File to HDFS
Write java programs to upload the text files to HDFS file system of Hadoop as the input file of Hadoop Mapper function.

3.3 Submit Tasks
Submit tasks to Hadoop cluster. Hadoop divides the input files into several pieces based on the size of the block, and each piece starts a Map Task for processing. Each Map Task performs the Map function to read the file information, generates a number of input <Key, Value> pairs according to the contents of the file. The Key is the byte offset of the row in the TXT file which is currently being read, the Value is the content of the current read line, in this paper, the format of the read line is MJD PLANID SPID FIBERID CRVAL1 CD1_1 NAXIS1 Wave[1] Flux[1] Wave[2] Flux[2] … … Wave[NAXIS1] Flux[NAXIS1]. We can find the meaning of each symbol in document 7. Then the Map function is used to perform continuum normalization and the recognition of the abnormal sky-subtraction spectra. Finally, we output the result in the form of <Key, Value> pairs, the Key’s format is spec-MJD-PLANID-spFIBERID-SPID, it will uniquely identify a spectral data, and the value is the calculation of the abnormal sky-subtraction positions.

4. Experiment Analysis

4.1 Performance Analysis of Hadoop
Table1 shows the running time about the automatic identification of the sky-subtraction residual components of the 226000 spectra between the Hadoop cluster model with only three computing nodes and the single model. Table2 shows the performance analysis. In this experiment, 63 Map Tasks and 1 Reduce Task were used to synchronize processing the data in the Hadoop cluster at the same time. So it can be seen from the table that the time spent by the Hadoop cluster is less. In addition, the storage capacity of the single mode is limited by the memory and the disk, while the Hadoop cluster can be easily extended to make it easier to store large data.

| operation mode       | runtime   |
|----------------------|-----------|
| Single model         | About 960s|
| Cluster model        | About 120s|

| Table 2 Performance analysis |
|-----------------------------|
| Speedup[^10]                | 8          |
| Efficiency[^11]             | 2.6        |
| Cost[^12]                   | 360s       |

4.2 Recognition of Sky-subtraction Residual Components Results
In practical application, we can increase or decrease the threshold value to detect sky-subtraction residual components with different strength according to the actual demand. If we only need to detect some high strength skylight residuals, we can increase the value of the standard deviation multiple correlation threshold in the experiment. In the experiment, we choose 68 high strength and undoped skylines. The following figure 5 is the recognition result of the abnormal sky-subtraction spectra, the black line represents the spectral flux after sky-subtraction process, and the blue line indicates the abnormal sky-subtraction wavelength positions. From the analysis to the experiment we discover that this method can not only effectively reduce the time of identification, but also find out the positions of the skylight residuals. Figure 4 shows the sky-subtraction residuals at the wavelength of 6863 Å and 8344 Å, the position of 8344 Å has the positive and negative feature, probably because the wavelength calibration is not accurate. In the following step, the sky-subtraction precision can be improved by
improving the wavelength calibration accuracy. The position of 6863 Å has the anomaly characteristics like a “W” shape, probably because there are contour inconsistency or resolution inconsistency anomalies occurred between the sky spectrum and the original spectrum around the sky emission line position.

Based on the analysis of the experimental results, it is found that the method can find out the residual distribution of abnormal sky-subtraction stellar spectra in varying degrees by adjusting the threshold value. In addition, through the experimental analysis for the abnormal sky-subtraction stellar spectra, the abnormal situation is mainly by fitting the wavelength calibration of imprecise, inconsistent resolution or outline, incorrect “super skylight” caused by background subtraction too much or too little. Because, in reality, the factors that affect the results of the sky-subtraction result is complex, most of which are combined with a variety of factors, so it is difficult to do the repair work about the abnormal sky-subtraction spectra. The most fundamental method to reduce the error of sky-subtraction by increasing the wavelength calibration accuracy, increase the number of skylight fibers, get method to fit the "super light" better agreement with the target spectrum.

![Abnormal sky-subtraction bands](image1)

**Figure 4** Abnormal sky-subtraction bands

![Abnormal sky-subtraction Spectra](image2)

**Figure 5** Abnormal sky-subtraction Spectra
5. Conclusions
After the completion of the spectral continuum fitting and the continuum normalization steps, the paper proposed a simple and effective method for the automatic detection of different degrees of the residual components based on the Hadoop platform. Through the analysis of the time performance, the experiment proved the efficiency of Hadoop cluster mode compared to single mode in large data processing, such as, the speedup is 8 in this experiment when the Hadoop cluster only have three nodes. In addition, the method proposed in the paper can recognize the abnormal skyline positions of different residual strength effectively and flexibly by adjusting the threshold.

6. Acknowledgment
This work was supported by the National Natural Science Foundation of China (UI431102).

7. References
[1] Dong Xicheng. Hadoop technology insider. BeiJing: China Machine Press.2013.
[2] The official website of LAMOST: http://www.lamost.org/.
[3] Cui, X., et al. The Large Sky Area Multi-Object Fiber Spectroscopic Telescope. Research in Astronomy and Astrophysics, 12(9): 1197-1242, 2012.
[4] Bai Zhongrui. Research on the sky-subtraction method of LAMOST[D]. BeiJing: university of science and technology of china.2007.
[5] S.J. Ko, Y.H. Lee. Center weighted median filter and their applications to image enhancement [J].IEEE Transactions on Circuits and System, 1991, 38 (1):984-993.
[6] An Ran, Pang Jingchang etc. The Automatic Recognition and Detection of Sky-subtraction Residual Component in The Stellar Spectra. Spectroscopy and Spectral Analysis. 2017, 37(1):273-277.
[7] Fits file format official website of LAMOST: http://dr1.lamost.org/doc/data-production-description.
[8] Basel A. Mahafzah. Parallel multithreaded IDA* heuristic search: Algorithm design and performance evaluation.International Journal of Parallel, Emergent and Distributed Systems, 26(1), 61-82, 2011.
[9] SW Al-Haj Baddar, BA Mahafzah. Bitonic sort on a chained-cubic tree interconnection network. Journal of Parallel and Distributed Computing, 74(1), 1744-1761, 2014.
[10] BA Mahafzah. Performance assessment of multithreaded quicksort algorithm on simultaneous multithreaded architecture. Journal of Parallel and Distributed Computing, 74(1), 1744-1761, 2014.