第42卷，第4期 光 谱 学 与 光 谱 分 析 Vol. 42，No. 4，pp1186-1191
2022年4月 Spectroscopy and Spectral Analysis April，2022

基于影响空间和数据场的LAMOST低质量光谱分析

杨雨晴[1]，蔡江辉[1,2*]，杨海峰[1*]，赵旭俊[1]，殷晓娜[1]

1. 太原科技大学计算机科学与技术学院，山西 太原 030024
2. 中北大学计算机科学与技术学院，山西 太原 030051

摘 要 针对LAMOST DR5 pipeline分类为Unknown的光谱数据对其进行了特征提取和聚类分析。主要工作如下：(1) 基于影响空间及数据场的特征提取。首先基于影响空间从低信噪比光谱中提取出大量小集团；然后计算各小集团内部的场并根据场对光谱排序，依次访问光谱序列及其小集团内的成员来获得特征谱；(2) 对上述特征谱进行K-means聚类，并统计了每一类目标所在天区、观测视宁度、各波段信噪比、亮度、光谱仪/光纤的分布情况。(3) 低质量光谱聚类结果的理论分析。通过聚类所有低质量光谱被分为了5大簇：A光谱信噪比较低或与传统分类模板差异较大，但通过特征分析可确定其类别(占比2.7%)；B光谱蓝端或红端出现疑似特征线或分子带，但与线表无法匹配(占比23.6%)；C光谱蓝端信噪比极低，且该波长区域噪声值较强，其他波长区域的连续谱和线的特征较弱(占比48.0%)；D红蓝两端拼接问题导致5 700～5 900 Å局部光谱突起明显，其他波长区域的连续谱和线的特征较弱(占比24.2%)；E存在大量缺省值导致无法确定其类别(占比1.5%)。实验结果表明，该方法不仅能够有效提取低信噪比光谱的特征谱，同时能够通过特征谱的聚类分析揭示低质量光谱的成因，从而为制定光谱观测计划提供参考，为低信噪比光谱分析及处理提供方法借鉴。

关键词 低信噪比光谱；光谱分解；特征分析；数据场；聚类分析
中图分类号：P114.1 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)04-1186-06

引 言

LAMOST[1-2]作为目前世界上光谱获取率最高的天文望远镜，为包括天文学以内的众多领域的科学研究提供了大量的珍贵样本[3-4]。然而随着巡天工作的不断深入，待观测的目标越来越暗，低信噪比光谱的数量也越来越多。如何有效处理低信噪比光谱一直是业内公认的难题。

为了获得低信噪比光谱中的有价值的信息，研究者们提出了诸多方法[5-8]。比如说：基于Hilbert-Huang变换[5]的方法将含噪短波信号进行经验模态分解，通过最大相关选择包含短波信号信息的固有模态函数进行信号重构，然后对重构信号进行谱减法降噪。基于傅里叶变换的方法[6]利用傅里叶变换得到二维光谱的频率域，然后通过加权滤波、低通滤波过滤噪声。Wigner变换[7]与加权滤波、低通滤波的结合有效地将噪声和信号分离，但其对截止频率的参考信号质量要求更高。Robnik和Seljak[8]提出了一种基于高斯匹配滤波的行星检测技术。目前，低信噪比光谱的分析和处理方法存在较多问题，针对低信噪比光谱流量分布特征开展分析的研究较少。本文从低信噪比光谱流量分布特征分析出发，介绍了低信噪比光谱的分解及聚类分析方法。该方法利用影响空间和数据场分析低信噪比光谱的流量分布特征并实现低信噪比光谱的特征提取，然后对特征谱进行聚类分析，最后探讨了各类低信噪比光谱的成因。

1 实验部分

1.1 数据选择

实验从LAMOST DR5 pipeline分类为Unknown的光谱数据中选取了50 000条光谱进行了实验。以可能存在特征线的波长范围的局部谱(4 000～5 500，6 300～7 000，8 400～8 800 Å，如图1中黑色曲线所示)为数据对象，利用影响空

收稿日期：2021-07-23，**修订日期**：2022-01-19
基金项目：国家自然科学基金项目(U1931209)，山西省重点研发项目(201903D121116)，山西省基础研究计划项目(20210302123223)，中央政府指导地方科技发展基金项目(20201070)资助
作者简介：杨雨晴，1992年生，太原科技大学计算机科学与技术学院博士研究生 e-mail：B20180012@stu.tyust.edu.cn
*通讯作者 e-mail：Jianghui@tyust.edu.cn；hfyang@tyust.edu.cn

间和数据场对局部谱进行分解和特征提取，并对特征谱(图 1 中粉色曲线)进行聚类，进而分析了各类光谱的差异，揭示了各类低质量光谱的形成原因。

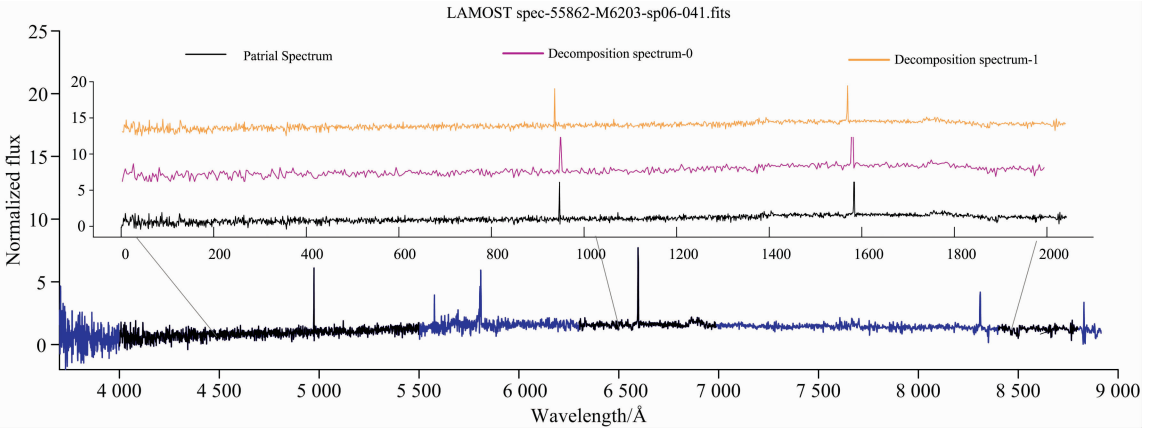

图 1　实验数据样例

Fig. 1　A sample of test data

1.2　特征分析方法

1.2.1　基于影响空间和数据场的低信噪比光谱分解

本文以位于同一影响空间[9]中的点为一个小集团且数据分布的稠密程度与小集团中的成员数成正比，每个数据通过场[10]发射的能量随距离的增加而降低。光谱中的特征线分布稀疏，特征线相对远离其小集团内的其他非特征线则其数据场相对较弱。各流量的数据场计算方式如式(1)

$$\varphi_{X_i}(X_j)=|IS(X_i)|\times\sum_{X_j\in IS(X_i)}\exp\left(-\left(\frac{\|X_i-X_j\|}{\sigma_i}\right)^2\right) \tag{1}$$

式(1)，X_i 为样本点，$X_i=(f_i,w_i)$，(f_i 和 w_i 为 X_i 的流量和波长)，$|IS(X_i)|$ 为 X_i 的影响空间的成员数，$\|X_i-X_j\|$ 为样本 X_i 到 X_j 的欧式距离。

根据数据场对各样本点降序排列，并依次访问排序后的各点及其所在小集团的所有元素。点的初始访问标志置为 0，访问后其访问标志修改为 1。原始光谱被分解为 0、1 对应的两条谱线，以访问标志位为 0 的点为特征谱(图 1 中的粉色曲线)开展聚类分析。

1.2.2　聚类分析

以各特征谱所在小集团的数据场均值为插值对特征谱进行波长统一和流量插值，并对插值后的特征谱开展 K-means 聚类。

2　结果与讨论

对 50 000 条 Unknown 光谱进行了 K-means 聚类分析，将特征谱划分为 5 大类，每类的均值谱及光谱范例如图 2—图 6 所示，其中左、中、右图分别为聚类中心光谱及其该类中随机的两条光谱。

Type1(图 2)，主要特征为连续谱信噪比较低，导致 LAMOST pipeline 模板匹配结果置信度较低，被分类为 Unknown。但是通过特定波长段较强的发射线特征，可以计算其视向速度(或红移)，从而对其光谱类型作出初步诊断。该类光谱占比较少，约 2.7%。

Type2(图 3)，主要特征为光谱信噪比不低，且光谱蓝端或红端出现疑似特征线或分子带，但与线表无法匹配，此类光谱类型基本无法识别，约占 Unknown 总数的 23.6%。

Type3(图 4)，主要特征为光谱蓝端信噪比极低，其他波长区域的连续谱和线的特征较弱，特征分析能发现部分疑似特征线，无法判定其类别，此类光谱占比达 48.0%。

Type4(图 5)主要特征为红蓝两端拼接问题导致 5 700～5 900 Å 局部光谱突起明显，其他波长区域的连续谱和线的

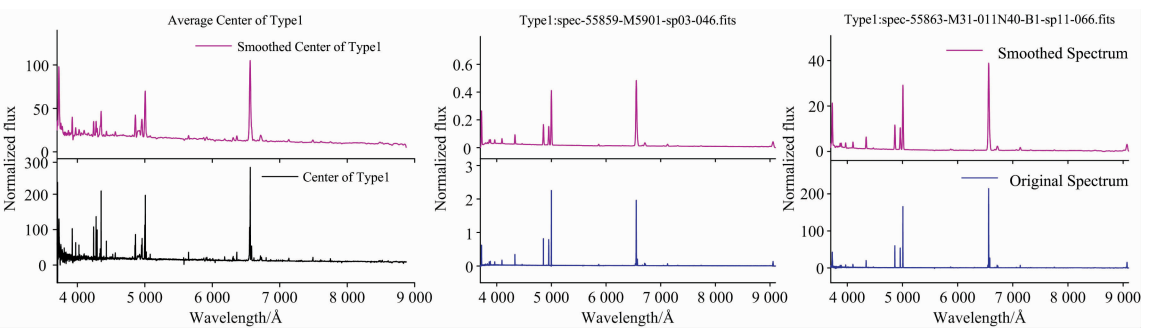

图 2　Type1 低信噪比光谱聚类中心及范例

Fig. 2　The center and examples of the first type of low SNR spectrum

特征较弱，模板匹配的效果较差，此类目标在屏蔽掉 5 700～5 900 Å 波长区域突起的光谱后，可能进一步识别其光谱类型，尤其对于某些珍稀天体的搜寻具有较好的价值，约占 24.2%。

Type5(图 6)主要特征为存在大量缺省值，曲线中部分位置为一条平直线，特征信息丢失而无法分辨其类别，约占 1.5%。

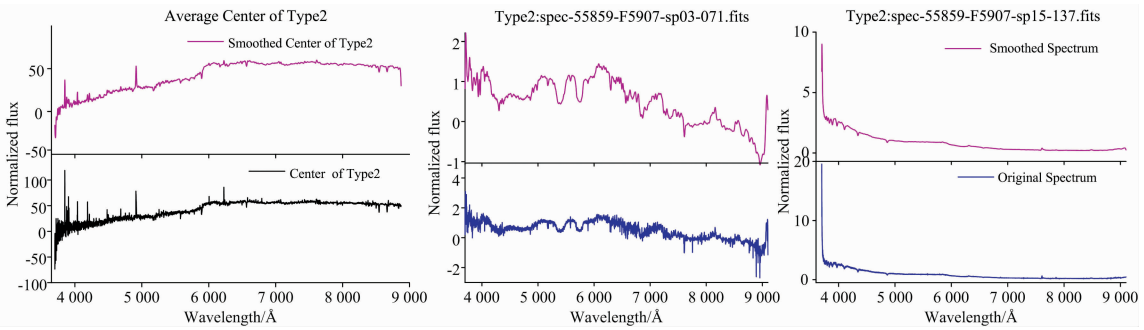

图 3 Type2 低信噪比光谱聚类中心及范例

Fig. 3 The center and examples of the second type of low SNR spectrum

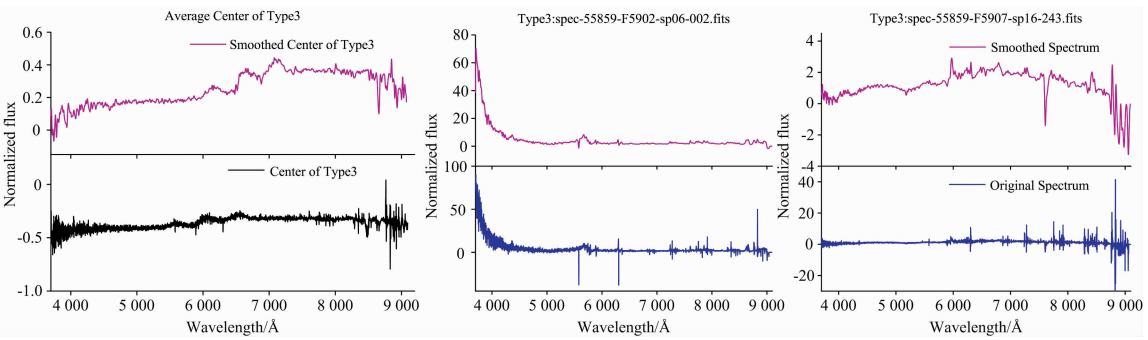

图 4 Type3 低信噪比光谱聚类中心及范例

Fig. 4 The center and examples of the third type of low SNR spectrum

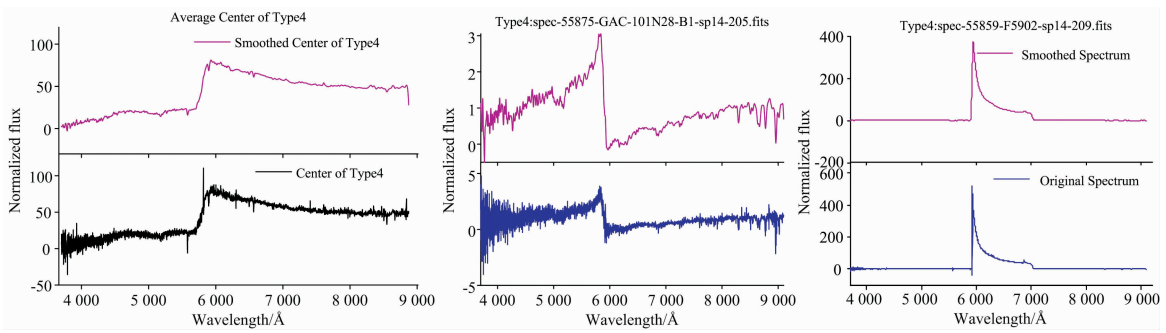

图 5 Type4 低信噪比光谱聚类中心及范例

Fig. 5 The center and examples of the fourth type of low SNR spectrum

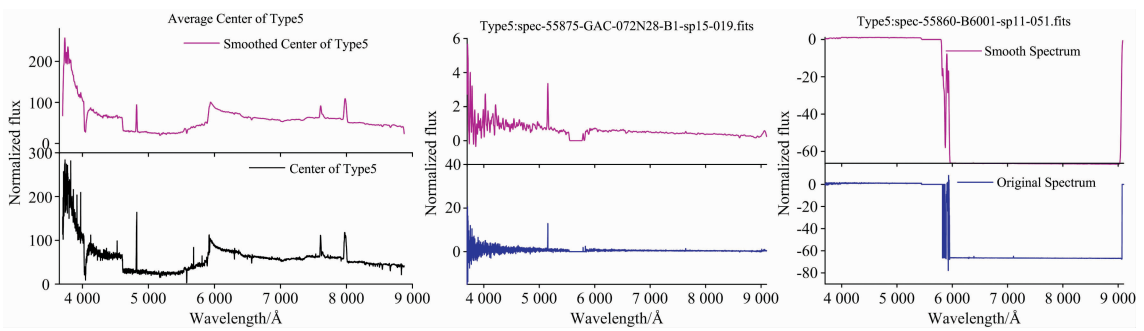

图 6 Type5 低信噪比光谱聚类中心及范例

Fig. 6 The center and examples of the fifth type of low SNR spectrum

2.1　光谱的天区分布分析

图 7 统计了各类光谱的比例以及其在天区的分布。可以看出，五类光谱总体分布在 B，F，GAC 和 M 天区，基本未呈现 HD 和 VB 天区的光谱，其中以 M 和 F 等较暗天区比例较高，而 B 和 GAC 天区相对较少；相比于其他类型，Type1 在 M 天区所观测的比例较高，这和该类光谱信噪比总体较低，而发射线可以识别的特征是一致的。

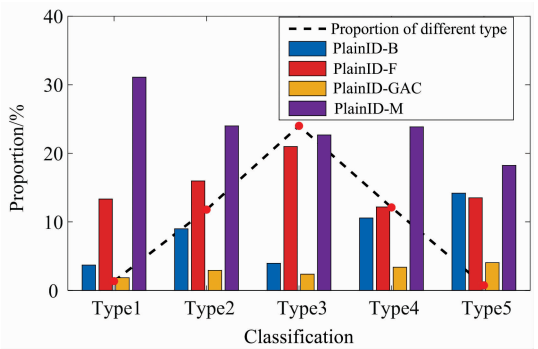

图 7　各类低信噪比光谱的比例及其在天区的分布

Fig. 7　The proportion of five types low SNR spectra and their distribution in the sky area

2.2　观测时的视宁度分析

图 8 给出了各类光谱在观测时的视宁度分布，图中未呈现明显的规律，视宁度大于 2.8 的比例较大，也即这些 Unknown 光谱观测时的环境较差，而 Type3(图中绿色部分)有小部分光谱观测环境较好，是由于其他原因导致 pipeline 无法有效识别。

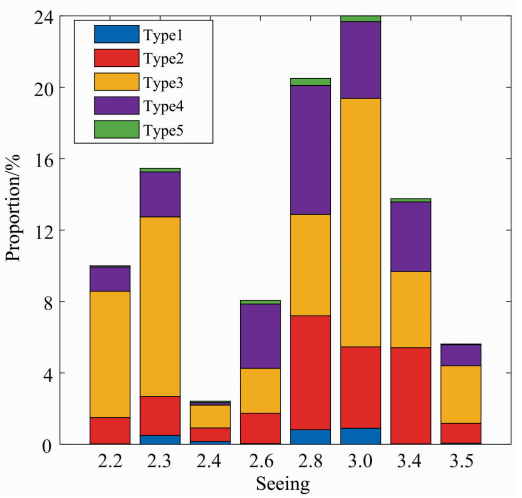

图 8　各类低信噪比光谱的视宁度分布

Fig. 8　Seeing distribution of low SNR spectra

2.3　观测目标的亮度分析

图 9 为各类光谱的亮度分布。可以看出各类天体的星等峰值集中在 17 mag，统计图的轮廓有向较暗的方向偏峰的特点，但特别暗的接近 LAMOST 极限星等的极少。因此，目标天体的亮度影响光谱质量，从而无法被 LAMOST pipeline 识别的可能性比较微小。

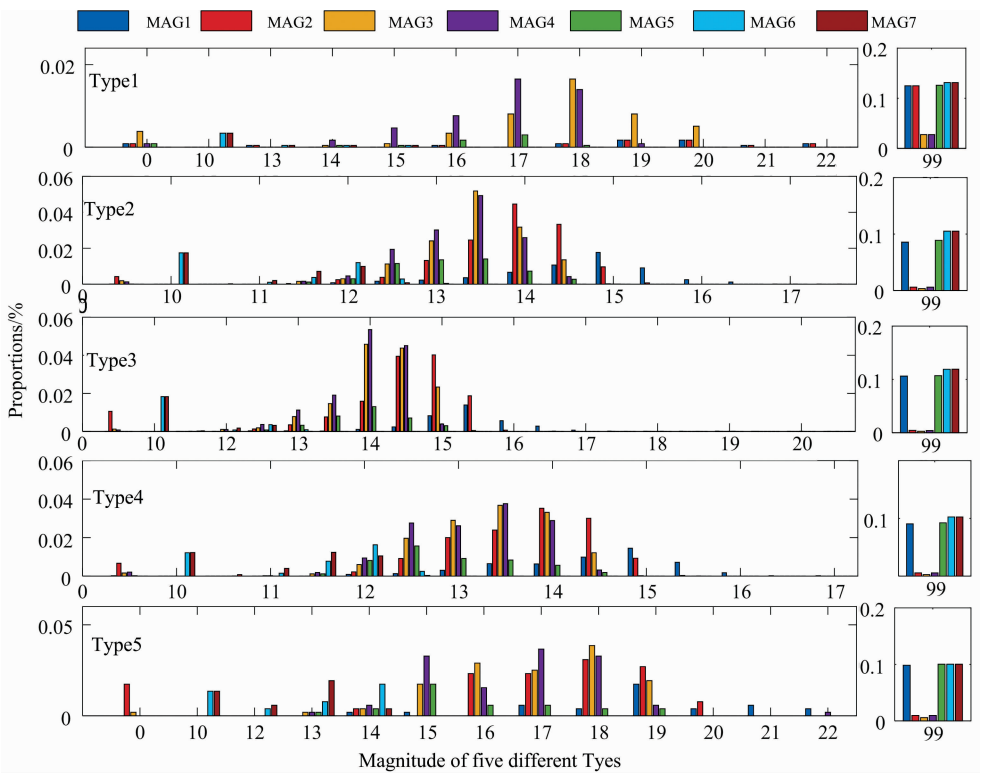

图 9　各类低信噪比光谱的亮度分布

Fig. 9　Magnitude distribution of low SNR spectra

2.4 各类光谱的光谱质量分布

我们在图 10 中统计了每类光谱在各波段的信噪比分布。光谱信噪比大致集中在(0，30)，Type5 信噪比相对较高；红蓝两端，尤其是蓝端的光谱质量对整条光谱的影响较大，而 i 和 z 波段的质量相对较高，这与 LAMOST 整体的信噪比分布基本一致，即本次分类与光谱质量的分布没有直接关系。

2.5 光谱仪与光纤分布

图 11 为低质量光谱所对应的光谱仪和光纤号的统计结果。各类光谱分布的主要光谱仪分别集中在：3，12，15；1，12，16；1，6，13，15，16；1，12，13，14；1，5，10，11，15 上，说明上述光谱仪观测的光谱质量总体较差。以 Type5 为例，除 4 和 16 号光谱仪上没有出现数据点外，其余光谱仪上均有数据点分布；1 号光谱仪的 19 号光纤，5 号光谱仪的 19 号光纤，10 号光谱仪的 105 和 19 号光纤，11 号光谱仪的 49 号光纤以及 15 号光谱仪的 19 号光纤出现的低质量光谱较多，分别占该类总比例的 5.41%，5.41%，6.76% 和 5.41%，9.46%以及 5.41%，对这些光谱仪和光纤所对应的光谱，在深入分析时需进行检验。上述统计数据对数据处理与分析乃至设备维护等工作具有一定的指导意义。

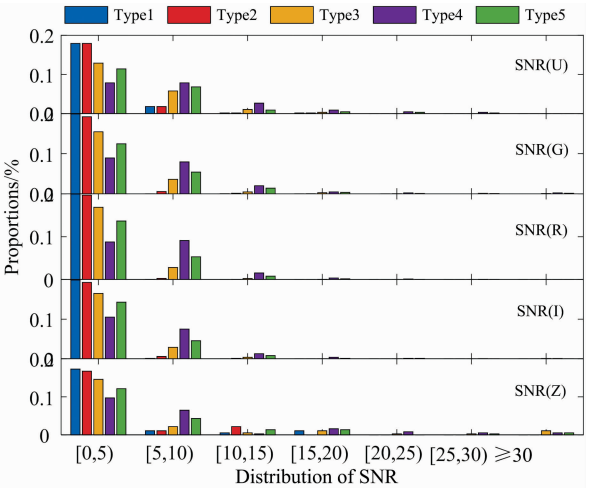

图 10 各类低信噪比光谱的信噪比分布

Fig. 10 SNR distribution of low SNR spectra

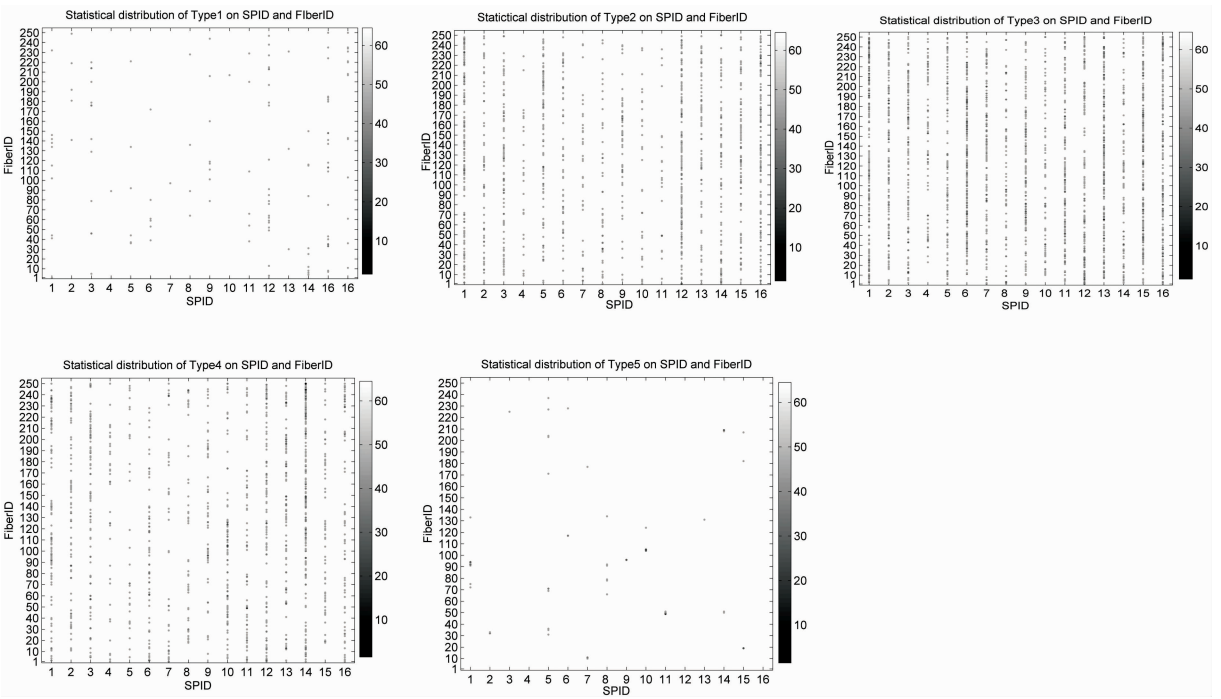

图 11 各类低信噪比光谱在光纤和光谱仪上的统计分布

Fig. 11 The statistics distribution of different types of low SNR spectra on SPID and FiberID

3 结 论

分析了低信噪比光谱的分布特征，给出了一种低信噪比光谱分解和聚类分析方法。该方法借助影响空间和数据域的相关技术分析了光谱流量的空间分布，然后依据上述分布对低信噪比光谱进行分解和特征提取，最后在特征光谱上完成了最终的聚类和结果分析。将所有低信噪比光谱分成了 5 类并揭示了各类光谱的形成原因，不仅为光谱观测计划的制定提供了依据，同时为低信噪比光谱分析和处理提供了新的手段。

References

[1] Luo A L, Zhang Y X, Zhao Y H. Advanced Software, Control, and Communication Systems for Astronomy, 2004, 5496: 756.
[2] Luo A L, Wu Y, Zhao J K, et al. Advanced Software and Control for Astronomy II, 2008, 7019: 701935.
[3] QU Cai-xia, YANG Hai-feng, CAI Jiang-hui, et al(屈彩霞, 杨海峰, 蔡江辉, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(4): 1304.
[4] Luo A L, Zhang J N, Chen J J, et al. Setting the Scene for Gaia and LAMOST, 2014, 9(298): 428.
[5] Xie L. Chinese Astronomy and Astrophysics, 2019, 43(4) : 579.
[6] Pan J. Research in Astronomy and Astrophysics, 2020, 20(9): 146.
[7] Chakrabarti D, Maji T, Mondal C, et al. Physical Review D, 2017, 95(7): 074028.
[8] Robnik J, Seljak U. Monthly Notices of the Royal Astronomical Society, 2021, 504(4): 5829.
[9] Yang Y Q, Cai J H, Yang H F, et al. Expert Systems with Application, 2020, 139: 112846.
[10] Thébault E, Lesur V, Kauristie K, et al. Space Science Reviews, 2017, 206(1-4): 191.

LAMOST Unknown Spectral Analysis Based on Influence Space and Data Field

YANG Yu-qing[1], CAI Jiang-hui[1,2*], YANG Hai-feng[1*], ZHAO Xu-jun[1], YIN Xiao-na[1]

1. School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China
2. School of Computer Science and Technology, North University of China, Taiyuan 030051, China

Abstract Based on the spectral data classified as Unknown by LAMOST DR5 Pipeline, the characteristics of low-quality spectra are extracted, and clustering analysis is conducted in this paper. The main work includes: (1) Feature extraction based on influence space and the data field. Firstly, a large number of small clusters are extracted from the low SNR spectrum based on influence space; secondly, each small cluster's data field is calculated, and the spectrum is sorted using the above field; and then, access the sorted spectrum and the members in its small cluster to obtain the characteristic spectrum. (2) Carry out K-means clustering with the above characteristic spectrum and statistics on the sky area, observed visual ninety, the signal-to-noise ratio in each band, brightness, and spectrometer/fiber distribution for each class of targets. (3) Analysis of clustering results of the low SNR spectra. All low-quality spectra are divided into five categories through cluster analysis: A. The spectral SNR is low, or the spectrum is different from the traditional classification template, but its category can be determined by feature analysis (accounting for 2.7%); B. Suspected characteristic lines or molecular bands that do not match the line table appear at the blue or red end of the spectrum (accounting for 23.6%); C. The SNR at the spectrum's blue end is very low, and the noise value in this wavelength region is strong. While in other wavelength regions, the features of continuous spectrum and line are weak (accounting for 48%); D. Due to the splicing problem, a protrusion can be seen in the local spectrum between 5 700 and 5 900 Å, and the continuum and line characteristics are poor at other wavelengths (accounting for 24.2%); E. Many default values make it impossible to determine the category of the spectrum (accounting for 1.5%). The experimental results show that this method can not only effectively extract the characteristic spectrum of low SNR spectrum, but also effectively carry out clustering analysis on the characteristic spectrum to reveal their causes, to provide a reference for the formulation of spectrum observation plan and the analysis and processing of low SNR spectrum.

Keywords Low-SNR spectra; Spectral decomposition; Feature analysis; Data field; Clustering analysis

(Received Jul. 23, 2021; accepted Jan. 19, 2022)

* Corresponding authors