第39卷，第5期
2019年5月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol 39，No. 5，pp1618-1622
May，2019

马氏距离度量LAMOST早型星光谱的分类研究

陈淑鑫[1,2]，孙伟民[1*]，宋轶晗[3]

1. 哈尔滨工程大学理学院，纤维集成光学教育部重点实验室，黑龙江 哈尔滨 150001
2. 齐齐哈尔大学机电工程学院，黑龙江 齐齐哈尔 161006
3. 中国科学院光学天文重点实验室(国家天文台)，北京 100012

摘 要 随着天文大数据不断积累，我国大天区多目标光纤光谱望远镜LAMOST已完成6年的大规模巡天观测，获得DR5数据集已达到900多万条光谱，其中含有观测比例较低的早型恒星光谱，具备重要的研究价值。利用准确的恒星分类模板库可提升恒星的分类精度与可靠性，由于LAMOST第一年的巡天光谱中并没有完整覆盖B型恒星包含的所有子类型，造成后续观测数据分类的子类型范围受限。依据LAMOST已发布DR5数据中B型恒星光谱为研究对象，选取ELODIE发布的B型恒星实测光谱模板库来检测LAMOST在用的分类光谱。首先完成ELODIE发布37条B型光谱模板的相关性分析，去掉相关性弱的三条光谱后，筛选出ELODIE 34条B型恒星实测模板作为中心，通过计算LAMOST DR5发布的绝大多数被标记为B6型(7 662条)和B9型(3 969条)实测光谱的马氏距离，经有监督聚类LAMOST早型恒星光谱数据，标记13个子类型在涵盖B2—B9子类的34条ELODIE光谱模板中的分布。经线性分析判别每条谱线子类型的类内距离，确保波长覆盖范围和分辨率与LAMOST数据完全一致，去掉距离数值偏差较大的数据，计算相应子类的平均谱线，得到LAMOST源于DR5观测数据早型B型恒星的13条子类型光谱分类模板，为后期完善模板提供较好的参考性。

关键词 马氏距离；早型恒星；光谱模板；LAMOST；ELODIE

中图分类号：TP391.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)05-1618-05

引 言

天文大数据时代，海量天体光谱数据是天体物理学研究的基础和论证依据[1]，其中恒星光谱分类是该领域研究的热点议题之一，采用模板匹配法是最直接有效且可靠性较高的分类方法，已经成功地运用在SDSS光谱巡天、LAMOST[2]光谱巡天等大型光谱巡天项目。LAMOST银河系巡天所获得的恒星光谱涵盖了各种温度类型的光谱，但目前发布数据的分类模板仍是来自LAMOST DR1巡天观测数据，第一年巡天的B型星本身数量稀少，且没有高质量B型恒星数据，最终模板库保留了原模板库四个B0，B5，B6和B9类型[3]，已发布数据主要集中在B6和B9两种类型的恒星谱线。

LAMOST已成功获取六年的巡天数据，包含各种B型恒星子类型的光谱，正确标记这些积累的数据，从中选择高质量的光谱进行模板构造更有重要的研究意义。本文选取国际上高度认可的ELODIE[4]高分辨实测光谱模板库，经降低ELODIE分辨率后标记LAMOST光谱数据。早型星光谱的谱线特征主要集中蓝端波段，且ELODIE光谱的波段覆盖范围只能达到6 800 Å，文中实验截取LAMOST光谱对应波段，再进行标记。利用R语言分析了该B型样本恒星间的相关性，去掉不可靠的ELODIE模板光谱，设置优化后的34条ELODIE数据光谱为中心，计算马氏距离度量已获取的LAMOST观测B型星光谱实施有监督聚类，从而构造出基于LAMOST DR5积累数据的实测模板，并构造出一套较完整的B型星光谱子型模板，为LAMOST早型恒星的光谱分类提供重要的依据。

1 恒星光谱模板

中国科学院国家天文台运行着特殊的反射Schmid(斯密

收稿日期：2018-04-12，**修订日期**：2018-08-31

基金项目：国家自然科学基金项目(U1631239)，国家自然科学基金青年科学基金项目(11803013)，黑龙江省省属高等学校基本科研业务专项(135209253)资助

作者简介：陈淑鑫，女，1978年生，齐齐哈尔大学机电工程学院教授 e-mail：shuxinfriend@126.com

*通讯联系人 e-mail：sunweimin@hrbeu.edu.cn

特)望远镜——大天区多目标光纤光谱望远镜(LAMOST)。截止到 2017 年 7 月，LAMOST 经过六年的巡天(http://dr5.lamost.org/)共获得了9 017 844个光谱，虚拟天文台已公布的大数据光谱包含8 171 443颗恒星。

1.1 LAMOST 恒星模板库

LAMOST 在用模板光谱来自约 100 万条大量的先导巡天恒星实测光谱数据，LAMOST 巡天光谱数据按 MK 分类标准系统分类，波长覆盖范围为 3 690～9 100 Å，步长为 1 Å(总采样点数 $N=5\ 491$)，分辨率为 1 800，利用局部孤立性概率算法(local outlier probability)及主成分分析(principle component analysis)重构方法[5]分组光谱数据，构建出模板光谱再通过人眼检查判断恒星模板库去掉其中低质量光谱，得到目前在用的 183 条模板光谱，包含 61 个不同子类型的恒星分类模板库。LAMOST 的 1D Pipeline 光谱分析软件利用这套模板对观测光谱数据进行分类和视向速度测量，通过 χ^2 距离最小化进行交叉相关匹配[6]，然而构建模板时 DR1 积累的 B 型光谱线类型不足，在用的这套模板中 B 型星的光谱子型并不完整。LAMOST 经过 6 年巡天后，已经积累足够的 B 型恒星数据重构，可建全完善子类型模板库。

1.2 ELODIE 实测模板库

ELODIE 模板数据是由高分辨的实测光谱构建，早在 1993 年底 Observatorie de Haute-Provence 1.93 m 望远镜率先使用 filber-fed echelle spectrograph ELODIE。后期公布了实测分辨率为 0.1 Å 的高分辨光谱库，由 1998 年 211 条只有 F，G 和 K 型恒星光谱增加到 917 条全模板光谱数据，提取 ELODIE 模板的 FITS 文件数据 header 文件波长信息，包含 CRVAL1 是对数下波长的起始位置，CDELT1 是对数下步长，NAXIS1 是数据点的个数，后续实验数据读取模板中 *.fits 头文件的 CRVAL 变量起始位置 4 104 Å，CDELT 步长变量为 0.2，截止波长为 6 800 Å 递增，每条光谱共对应 13 501 个波长的流量数据，此套数据覆盖 B 型恒星各种类型光谱，选取高质量的作为标准对 LAMOST 的观测光谱进行标记。

2 ELODIE 模板相似性分析

因光谱数据是高维数据，可将波长不同的光谱数据采样看成不同能量、不同维度的分布。本节研究 ELODIE 光谱库中 B 型恒星子类型之间的相关性关系，选取 B 型相关性较高的谱线，去掉相关性低的异常光谱。

2.1 数据相关分析

传统的高维数据相关分析是以协方差矩阵为基础构造划分数据模型，但经典的样本协方差矩阵的估计方法难以适用于高维的光谱数据。早期英国科学家高尔顿 Galton 给出相关性的概念：描述一个变量变化时，另一个变量也随之相应地变化，测量相关关系的统计量为相关系数。相关系数为“0”代表不相关、“1”代表全相关，则介于“0～1”之间的数值越大表示相关性越强，当因变量增加时，另一个变量也随之增加称为正相关，用正数表示同方向；相反随因变量而减少则称为负相关，用负数表示反方向。按照相应规则分析天文大数据集拆分成若干子数据集，借鉴经典的统计抽样思想设计有效的数据分类，然后分析每个小分类聚类后的数据集，后文 3.4 节开展实验描述理论分析数据集的拆分与拟合。

2.2 马氏距离计算方法

目前，欧氏距离(Euclidean distance)被广泛使用求 m 维空间中两个点之间的真实距离[7]，本文采用马氏距离(Mahalanobis distance)是一种充分考虑各变量间协方差的广义距离，利用采样协方差来计算两点之间距离的方法。与常用欧氏距离相比，马氏距离能消除量纲及各变量间相关性的影响，本质上认为它是某样本点属于某集合的概率度量。将 m 维马氏空间视为一组点集，它的每个点可以表示为(x_1，x_2，…，x_m)，其中 x_i($i=1$，2，…，n)是实数为 x 的第 i 维坐标，其与 y_i($i=1$，2，…，n)马氏空间的 m 维马氏距离 $dm(x,y)$ 如式(1)所示。

$$d_m(x,y)=\sqrt{(x_i-y_i)^{\mathrm{T}}\boldsymbol{C}^{-1}(x_i-y_i)} \tag{1}$$

其中 T 表示转置，$\boldsymbol{C}$ 为协方差矩阵如式(2)所示。

$$\boldsymbol{C}=\frac{(n_1-1)C_1+(n_2-1)C_2}{n_1+n_2-2} \tag{2}$$

2.3 基于马氏距离的监督聚类

马氏距离实际上采用 19 世纪末的法国数学家 Cholesky 提出的分解方法(Cholesky transformation)消除不同维度之间的相关性和尺度不同的性质，Cholesky 分解矩阵的效率较高，无需归一化处理，只需将矩阵分解为一个下三角矩阵以及它的共轭转置矩阵的乘积，即可消除不同维度之间的相关性尺度，而欧氏距离必须首先完成归一化处理后，再计算两两之间的距离，否则距离值无意义。后续第 3 节将详细介绍根据马氏距离对 LAMOST 中 B6 和 B9 型光谱数据，以 ELODIE 光谱模板为中心的聚类详见表 1 所示。

3 实验数据分析

实验构建相异度矩阵中，每条 ELODIE 光谱数据视为高维空间中的一个点，利用马氏距离值进行相关性度量分析，借助绘制相关性分析图，实现可视化相异度矩阵，相异度低表示相关性高，即可筛选出用于标记的样本。

3.1 线性插值模板波长

LAMOST 低分辨率光谱数据表示波长范围 3 865～9 000 Å，而 ELODIE 高分辨率光谱波长范围仅为 4 104～6 800 Å，为保证 ELODIE 所表示的蓝端范围不变，实验以 LAMOST 的光谱波长值作为参照，采用线性插值方法降低 ELODIE 分辨率，使之对应的波长保持一致，便于标准化处理光谱流量值进行数据比对分析。

3.2 计算 ELODIE 恒星相关性

实验基于 R 语言环境计算模板间的马氏距离，经线性插值处理 ELODIE 中 B 恒星分类 .csv 文件中每列为一条模板的一维数组，生成 37 列×2 193 行数据集。采用 R 语言 corrgram( )函数转换分析光谱数据，corrgram(x，order＝，panel＝，text.panel＝，diag.panel＝)函数中，行向量 x 表示每条模块光谱的数据框。当 order＝TRUE 时相关矩阵将利用

主成分分析法对变量重排序，二元变量的关系模式运行后得

表 1 比对 LAMOST 实测光谱与 ELODIE 模板光谱统计匹配数据

Table 1 Comparison of spectral statistics data of LAMOST and ELODIE spectra templates

ELODIE 模板		LAMOST 数据	
恒星类型	模板文件	B6 型条数	B9 型条数
B2：Ⅲ pshe	00235. fits	6	24
	00236. fits	9	9
B3Ve	00584. fits	1	2
B4Ⅲp	00156. fits	3 718	1 285
	00157. fits	1	3
B4Ⅴne	00863. fits	2	2
	00864. fits		2
B6Ⅲp	00433. fits	4	
	00432. fits		4
B6Ⅴ	00518. fits	1	2
	00521. fits	3	1
B6Ⅴnn	00475. fits	1 046	728
	00476. fits	7	14
B7p	00099. fits	4	5
	00100. fits	2	1
	00101. fits		1
	00102. fits	3	
	00103. fits	11	19
	00104. fits	1	
	00105. fits		2
B8	00161. fits	1	
	00529. fits	1	2
B8Ⅲp	00754. fits	1	1
B9	00576. fits	2	
	00164. fits	3	11
	00165. fits	82	154
B9Ⅲ	00533. fits		4
	00662. fits	1	
B9p	00002. fits	12	5
	00484. fits		2
	00485. fits		1

到与所有(包含自身)的马氏距离即该类型模板间距离为 37×37 阶矩阵，可视化特征提取图像距离值如图 1 所示，展现了 37 条光谱简单的相关性方向和强弱，主对角线显示所分析的模板光谱

色单元格表示两个变量呈正相关。反之红色表示变量呈负相关，饱和度越高则颜色越深，则光谱包含变量相关性越大。矩阵上三角单元格显示饼型图中，颜色相同为同信息，被填充的饼型图中块的大小展示相关性的数值，正相关性从 12 点钟方向开始顺时针填充饼图，而负相关性则逆时针方向填充饼图，相关性接近于 0 时单元格基本无色，B 类型光谱模板相互间呈正相关，如图 2(a—c)所示模板文件 B5 型 00872. fits、B9 型 00893. fits、和 B9. 5Ve 型 00888. fits 出现 3 条异常谱线，为方便比较给出图 2(d)正常 B9p 型光谱模板文件 00485. fits。

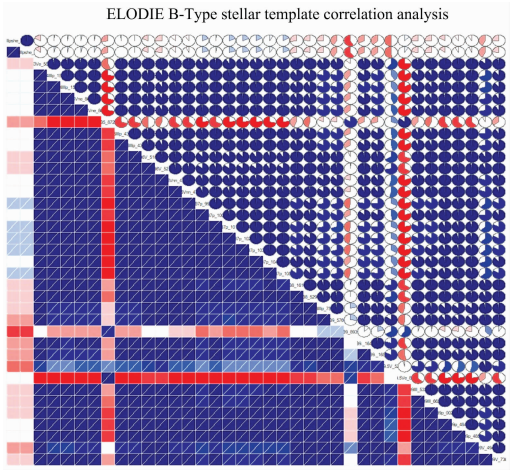


图 1 马氏距离提取 ELODIE 实测 B 型恒星 37 条模板间相关性分析

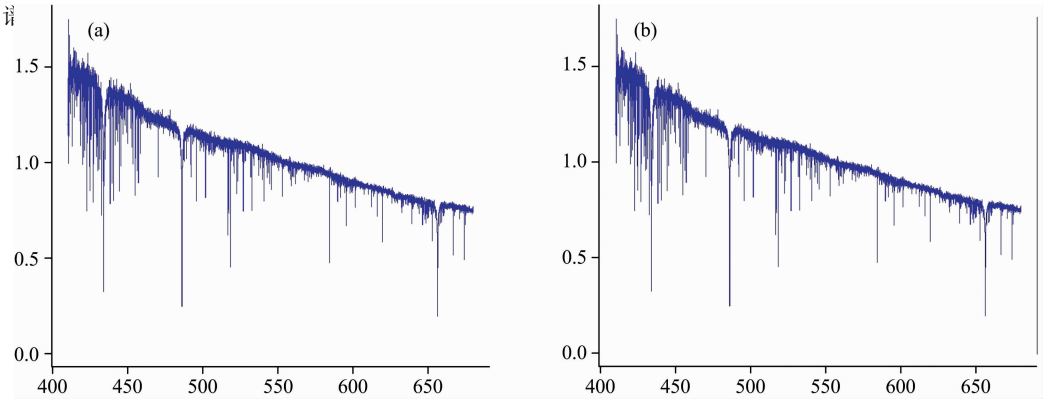
Fig. 1 Analysis of the correlation between 37 types of B-type stars of ELDIE measured by Mahalanobis distance

3.3 LAMOST 的数据聚类

LAMOSTDR5 已发布 B 型星主要集中在 B6 型(7 662 条)和 B9 型(3 969 条)恒星，扣除 DR1 数据后剩余 B6 型 4 920条 *. fits 观测数据和 B9 型共 2 286 条数据。依据 3.2 节分析采用 ELODIE 相关性较好的 34 条实测模板作为中心进行有监督聚类。

3.4 比对分析

实验采用有监督方法经清理 LAMOST 巡天 DR1 数据后，利用 DR5 发布数据与 ELODIE 相关性较好的 34 条实测 B型模板间的马氏距离度量进行聚类。如表1所示分别列出



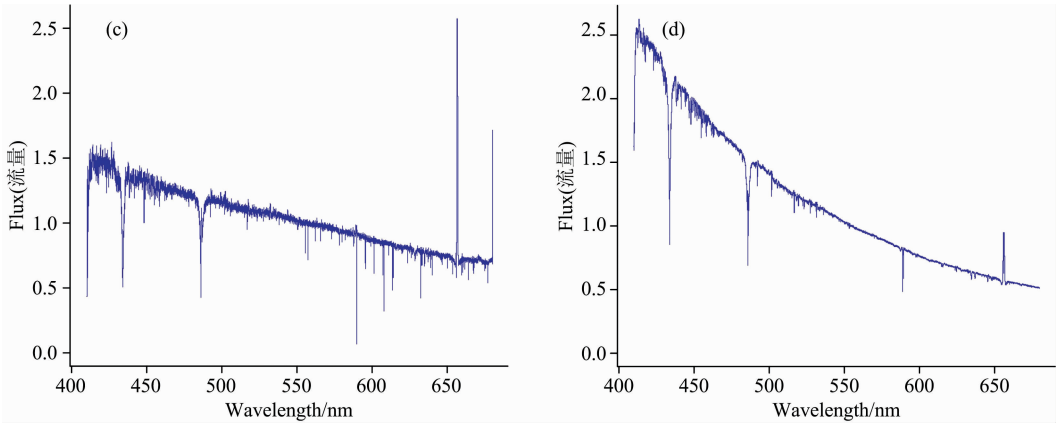


图 2　读取 ELODIE 实测模板中 B 型恒星相关性异常 . fits 谱线与正常图像比较

(a)：异常 B5 型 00872. fits；(b)：异常 B9 型 00893. fits；(c)：异常 B9. 5Ve 型 00888. fits；(d)：B9p 型 00485. fits

Fig. 2　Contrast the . fits spectral line images of abnormal B-type stars in the measured template of ELODIE

(a)：Abnormal type B5 is the file of 00872. fits；(b)：Abnormal type B9 is the file of 00893. fits；
(c)：Abnormal type B9 is the file of 00888. fits；(d)：Normal type B9p is the file of 00485. fits

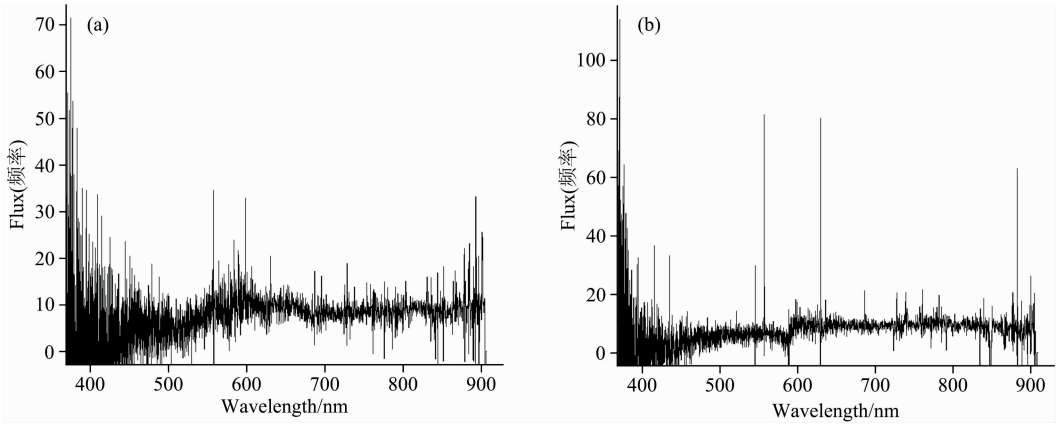


图 3　匹配 LAMOST 实测 B 型恒星光谱数据马氏距离分析异常的 fits 谱线图像

Fig. 3　Matching LAMOST measured B-type spectral data Mahalanobis distance analysis abnormal fits line image

(a)：spec-55859-F5907_sp02-184. fits；(b)：spec-55859-F5902_sp05-199. fits

ELODIE 的恒星类型和模板对应 * . fits 数据文件名，LAMOST 观测高质量的 B 型光谱中 B6 型 4 920 条和 B9 型 2 286 条实施硬聚类统计最小马氏距离，得到 34 条 ELODIE 模板数据，匹配出 13 条有效的样本模板，其中多数聚集在 B4Ⅲ与 B6Ⅴ类型，约占 B6 型分类的 97%，B4Ⅲ，B6Ⅴ和 B9 类型约占 B9 型分类的 95%。

聚类所得 13 个子类判别线性，并去掉那些检查马氏距离偏差较大的光谱如图 3(a)列出的 B6 型光谱 spec-55859-F5907_sp02-184. fits、如图 3(b)所示 spec-55859-F5902_sp05-199. fits 等。然后在整个 LAMOST 波段范围计算这 13 个类型的中值谱，得出每个类型的模板，利用聚类后的实测数据再重新构建出可用的模板库。

清理后光谱数据按表 1 显示 13 类的条数，计算每个类型光谱的中值，再执行归一化数据处理后取每个波长的中值，构造出新分类的光谱模板。结合 LAMOST 自身特点直接影响采集数据流量改变存在一定误差的连续谱，如图 4 所示给出去掉连续谱后构建的 B9Ⅲ型模板。

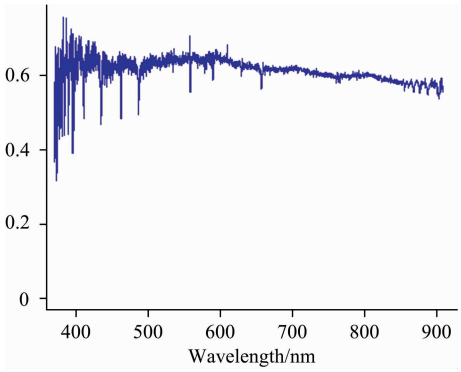


图 4　LAMOST 实测光谱构建 B9Ⅲ型光谱模板

Fig. 4　Examples of B type templates constructed with LAMOST empirical spectra

4 结 论

基于 R 语言实现不受量纲影响的高维光谱模板间的马氏距离计算，呈现 ELODIE 模板库 B 型恒星子类型模板距离阵分布，运行可视化函数提取图像距离值，得到与所有(包括自身)的谱线特征相关性分析，从而利用相关性度量值优化恒星模板库。经 LAMOST 实测数据计算结果拟合出蓝端及红端连续的光谱模板，优化的模板从海量的 LAMOST 巡天光谱数据中谱线匹配，获取更多有价值的信息，为后续判别恒星类型提供有效的论证依据。

References

[1] ZHAO Yong-heng(赵永恒). Scientia Sinica: Physica, Mechancia & Astronomica(中国科学：物理学力学天文学), 2014, 44(10): 1041.
[2] Luo A L, Zhao Y H, Zhao G, et al. RAA, 2015, 15(8): 1095.
[3] Wei P, Luo A L, Li Y B, et al. The Astronomical Journal, 2014, 147(5): 101.
[4] Bouchy F, Ségransan D, Díaz R F, et al. Astronomy Astrophysics, 2016, 585:A46.
[5] ZHONG Shou-bo, HAN Bo, ZHANG Yan-xia, et al(钟守波，韩 波，张彦霞，等). Astronomical Research & Technology(天文研究与技术), 2015, 12(4): 510.
[6] CUI Chen-zhou, YU Ce, XIAO Jian, et al(崔辰州，于 策，肖 健，等). Chinese Science Bulletin(科学通报), 2015, 60(5-6): 445.
[7] CHEN Shu-xin, SUN Wei-min, KONG Xiao(陈淑鑫，孙伟民，孔 啸). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(6): 1951.

Study on the Classification of LAMOST Early Stellar Spectrum Template by Mahalanobis Distance

CHEN Shu-xin[1,2], SUN Wei-min[1*], SONG Yi-han[3]

1. Key Lab of In-fiber Integrated Optics, Ministry of Education, Harbin Engineering University, Harbin 150001, China
2. College of Mechanical and Electrical Engineering, Qiqihar University, Qiqihar 161006, China
3. Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China

Abstract With the continuous accumulation of astronomical data, Large Sky Area Multi-Object Fiber Spectroscopy Telescope (LAMOST) has completed six years of large-scale sky surveys. The DR5 dataset has obtained more than 9 million spectrum data, including early-type stellar spectra with the lower observation proportion. The correct stellar classification template library can improve the classification accuracy. The currently used classification templates in LAMOST pipeline which don't completely cover all the subtypes such as B-type stars, because they were constructed using DR1 data without enough early type stars. In this paper, the B-type stellar spectra in LAMOST DR5 have been collected as our research object, and the reference B-type spectra are from the library of ELODIE. Firstly, we complete the correlation analysis of 37 spectra of ELODIE B-type stars. After removing three weakly correlated spectra, 34 spectra of ELODIE B-type stars spectra were selected as the cluster center. The majority of the published LAMOST DR5 labels were marked as B6 (7662) and B9 (3969) and spectra were measured by Mahalanobis distances, with the supervised clustering, 34 LAMOST early-type stellar spectral data were marked as 13 subtypes according to ELODIE labels covering from B2 to B9 subclasses. The intra-class distance of each spectral subtype is determined by linear analysis to ensure that the wavelength coverage and resolution are completely consistent with the LAMOST data. The average spectral line of the corresponding subclass is calculated removing the outliers, thus 13 subtype spectral classification templates of B-type stars provide a good reference for later template completion.

Keywords Mahalanobis distance; Early-type stellar; Spectrum template; LAMOST (large sky area multi-object fiber spectroscopy telescope); ELODIE

(Received Apr. 12, 2018; accepted Aug. 31, 2018)

* Corresponding author