第39卷，第10期
2019年10月

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

Vol. 39, No. 10, pp3292-3296
October, 2019

基于 XGBOOST 的恒星光谱分类特征数值化

张 枭[1,2]，罗阿理[1*]

1. 中国科学院国家天文台光学天文重点实验室，北京 100101
2. 中国科学院大学，北京 100049

摘 要 恒星光谱分类是研究恒星的基础性工作之一，常用的光谱分类是基于20世纪70年代Morgan和Keenan建立起来的并逐步完善的MK分类系统。然而基于MK规则的交互式决策分类系统对处理海量天文光谱数据存在着一定的困难。目前光谱巡天一般采用的自动化分类则是模版匹配方法而忽略对谱线特征的测量。怎样自动、客观地提取海量光谱中的分类特征并应用这些特征进行分类可以对天体的物理化学性质的统计分析至关重要。针对此问题，通过机器学习和计算光谱的谱线指数结合的方法，提取光谱特征，并通过大数据分析定量地确定对光谱特征谱线的分类判据(数值化)，确定每一类光谱具有物理意义的特征谱线的强度分布。首先对LAMOST DR4恒星光谱测量其谱线指数作为输入，光谱的分类标记采用官方发布的分类结果。使用XGBoost算法进行自动分类及特征排序，从而获得已知或未知的对于分类决策最为敏感的谱线。首先，选取高信噪比($S/N>30$)、被LAMOST标记为B，A，F和M的恒星光谱数据，总计约414万个。然后，对光谱数据计算谱线指数从而使其得到降维处理，过滤冗余信息。其次，将处理后的恒星光谱数据随机划分为训练集和测试集，通过适当调整算法参数，用训练集得到所需要的分类决策树模型，用测试集测试其稳定性和可用性，以防止出现过拟合，同时使用算法自带函数进行提取分类特征。最后，输出并整理实验中算法所得的决策树模型，并挑选其概率比较大的分支作为最终的决策树模型。通过实验，可以发现在固定参数下，XGBoost所得的模型有一定的自适应性，较少受数据集影响，总体准确率可达88.5%；同时其所输出的分类决策树与已知的特征较为吻合，而且可以获得基于大数据的、数值化的特征谱线对应分类的范围，为完善基于特征的分类提供定量的规则。

关键词 光谱分类；线指数；XGBoost；决策树；LAMOST

中图分类号：P152 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)10-3292-05

引 言

LAMOST，全称是“大天区面积多目标光线光谱望远镜”，是一架自适应光学的施密特反射式望远镜，位于北京兴隆观测站。因纪念元代天文学家郭守敬而冠名为郭守敬望远镜[1]。

LAMOST产生的约一千万条恒星的光谱是目前世界上最大的恒星光谱库。对这些光谱分类对研究银河系的各种规律是十分必要的。LAMOST光谱数据处理的官方软件流程是采用模板匹配的方法[1]，但存在缺陷：(1)数据质量存在限制，连续谱的质量直接影响了分类结果；(2)本质上是高维数据的距离，意味着数据里包含一些重复、可约简的信息。

在过去的数十年中，很多自动分类方法已经应用于恒星光谱分类。人工神经网络(ANN)算法在天文领域应用广泛。2011年，Schierscher和Paunzen使用ANN算法对SDSS DR7数据的部分恒星光谱进行分类[2]；在2017年Hampton等通过ANN对积分场光谱型的发射线进行分类[3]。

支持向量机(SVM)算法在天文也有广泛应用。近些年来，Liu等在2015年对LAMOST数据使用线指数和SVM算法进行MK分类[4]；2016年Du等使用Baysian SVM和PCA方法对SDSS DR10进行M型星子型M0，M1，M2，M3，M4的分类[5]。

PCA是一种常用的特征提取方法，但是PCA提取的是数学特征，不具有物理意义。

收稿日期：2018-09-07，**修订日期**：2019-01-18

基金项目：国家自然科学基金项目(11390371)资助

作者简介：张 枭，1990年生，中国科学院大学国家天文台硕士研究生 e-mail：cyscum@outlook.com

*通讯联系人 e-mail：lal@nao.cas.cn

采用谱线指数进行光谱自动分类除了前述 Liu 等的工作外[4]，王光沛等也对 lick 线指数进行了聚类[7]。这些工作都揭示了谱线指数与物理特征之间的某种联系。但这种多特征与恒星类型之间的联系并非简单关系，而是一种需要通过树来表达的复杂关系。

基于树的机器学习算法是一类有监督的机器学习算法，具有模型结构相对简单、运算量相对较小，同时准确率相对较高等优点。其中，Chen 等[8]提出的 XGBoost 算法，是一种迭代型树类算法，有着更容易实现的并行处理、更快的运算处理速度、比起传统决策树更高的准确性等备受瞩目，成为一种流行的机器学习算法，应用于诸多领域。此外，XGBoost 克服了 ANN 算法复杂的、难以解读的隐藏层问题；相对 SVM 等算法，有更高的准确率。而对于光谱分类最为重要的特点是其能够容易、直观地提取分类特征，对这些分类特征的物理解释可以帮助我们理解唯像光谱分类背后的物理本质。

通过所获得的分类的数值特征，我们就可以将基于 MK 原则的分类决策树中某些定性的决策规则定量化，使得分类决策更加简单并容易解释。

1　方　法

1.1　线指数

采用线指数是等值宽度定义，即

$$EW = \sum\left[1 - \frac{f_{\text{line}}(\lambda)}{f_{\text{count}}(\lambda)}\right]\mathrm{d}\lambda$$

其中 $f_{\text{line}}(\lambda)$和 $f_{\text{count}}(\lambda)$是在伪连续谱上的流量，$f_{\text{count}}$是通过在伪连续谱插值而得到[1]。每个线指数定义见表 1。

表 1　采用的线指数定义[4, 6]

Table 1　Line index definitions[4, 6]

谱线名称	通带/Å	伪连续谱范围/Å
H_delta	4 083.500～4 122.250	4 041.600～4 079.750, 4 128.000～4 161.000
CN1	4 143.375～4 178.375	4 081.375～4 118.875, 4 245.375～4 285.375
CN2	4 142.125～4 177.125	4 083.875～4 096.375, 4 244.125～4 284.125
Ca4227	4 223.500～4 236.000	4 212.250～4 221.000, 4 242.250～4 252.250
G_band	4 282.625～4 317.625	4 267.625～4 283.875, 4 320.125～4 336.375
H_gamma	4 319.750～4 363.500	4 283.500～4 319.750, 4 367.250～4 419.750
Fe4383	4 370.375～4 421.625	4 360.375～4 371.625, 4 444.125～4 456.625
Ca4455	4 453.375～4 475.875	4 447.125～4 455.875, 4 478.375～4 493.375
Fe4531	4 515.500～4 565.500	4 505.500～4 515.500, 4 561.750～4 580.500
Fe4668	4 635.250～4 721.500	4 612.750～4 631.500, 4 744.000～4 757.750
H_beta	4 847.875～4 876.625	4 827.875～4 847.875, 4 876.625～4 891.625
Fe5015	4 977.750～5 054.000	4 946.500～4 977.750, 5 054.000～5 065.250
Mg1	5 069.125～5 134.125	4 895.125～4 957.125, 5 301.125～5 366.125
Mg2	5 154.125～5 196.625	4 895.125～4 957.625, 5 301.125～5 366.125
Mgb	5 160.125～5 192.625	5 142.625～5 161.375, 5 191.375～5 206.375
Fe5270	5 245.650～5 285.650	5 233.150～5 248.150, 5 285.650～5 318.150
Fe5335	5 312.125～5 352.125	5 304.625～5 315.875, 5 353.375～5 363.375
Fe5406	5 387.500～5 415.000	5 376.250～5 387.500, 5 415.000～5 425.000
Fe5709	5 698.375～5 722.125	5 674.625～5 698.375, 5 724.625～5 738.375
Fe5782	5 778.375～5 798.375	5 767.125～5 777.125, 5 799.625～5 813.375
NaD	5 878.625～5 911.125	5 862.375～5 877.375, 5 923.875～5 949.875
TiO1	5 938.375～5 995.875	5 818.375～5 850.875, 6 040.375～6 105.375
TiO2	6 191.375～6 273.875	6 068.375～6 143.375, 6 374.375～6 416.875
H_alpha	6 548.000～6 578.000	6 420.000～6 455.000, 6 600.000～6 640.000
CaK6	3 930.700～3 936.700	3 903.000～3 923.000, 4 000.000～4 020.000
CaK12	3 927.700～3 939.700	3 903.000～3 923.000, 4 000.000～4 020.000
CaK18	3 924.700～3 942.700	3 903.000～3 923.000, 4 000.000～4 020.000

1.2　XGBoost 算法

提升树决策树算法(GBDT)采用决策树为基函数，GBDT 是利用梯度下降法的近似方法；XGBoost 是梯度提升算法的改进版本，相比梯度提升算法，对损失函数进行二阶泰勒展开，可以进行凸优化。同时可以进行分布式计算来提高速度。算法步骤如下：

目标函数

$$L(\Phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

其中

$$\Omega(f_k) = \gamma T + \frac{1}{2}\gamma \| w \|^2$$

对每步训练目标函数二阶泰勒展开

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

记

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}),\ h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$$

得到

$$L^{(t)} \approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

求出目标函数最优解

$$\hat{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T$$

而上式即可作为对应树子叶的分数；分数越大，该树结构越好[2]。而算法的其中一点就是搜索给定参数中最大的所得的值；一旦再分裂后所得结果小于原结果，算法随即停止继续增长子叶深度[9]。

2 数据处理实验

2.1 数据准备

我们数据取自 LAMOST DR4，考虑到以生成模型、统计分析为主要目的，所以尽可能多选取样本；为了得到较为准确的结果，同时选取高信噪比($S/N>30$)恒星光谱数据。

同时对光谱数据进行线指数计算，使得原本的伪连续谱简化为 27 维数据的矩阵，大大降低原数据量。

表 2 各类恒星光谱数量

Table 2 The quantity of stellar spectra

类别	数量
B	4 466
A	221 024
F	1 120 500
G	2 009 758
K	477 003
M	229 192

2.2 实验步骤

将 414 万条光谱线指数数据随机分为训练集和测试集，两者数据比约为 3∶1。

调整 XGBoost 的参数为：采用决策树，booster 为 gbtree，因为多分类并希望得到属于每一类的概率，objective 为 multi：softpro，学习率 eta=0.7，为了提高模型准确率，设置最大可能树深度为 12；为了模型更加稳定，设置正则项 lambda 为 4，也使得算法保守。

为了得到模型，迭代次数只设为 1 次，这样因为没有对残差进行拟合，牺牲一定的准确率，但是可以较容易的得到所需要的决策树模型。其他大多采取默认。

2.3 结果与分析

实验得到 XGBoot 分类正确率如表 3。从中可以看出，实验分类结果和 LAMOST 所标定的结果大体一致，虽然舍弃了对残差的拟合，总体正确率仍较高，在 88.5%左右。和 Liu 等 2015 年使用 SVM 所做的恒星线指数分类结果类似，相比其他类型，B 型星分类正确率偏低。考虑到 B 型特征线指数仅为 He 线，而本次分类实验未包括 He 线；同时，相比其他类型，B 型数量明显偏少，这对实验结果同样有明显影响。

表 3 基于 XGBoost 分类误差矩阵

Table 3 The confusion matrix in terms of percentage of the XGBoost-based MK classification

	B	A	F	G	K	M
B	74.90%	0.74%	0	0	0	0
A	24.72%	94.38%	2.81%	0.01%	0	0
F	0.09%	4.82%	89.88%	8.27%	0.02%	0.01%
G	0.09%	0.03%	7.25%	85.37%	6.00%	0.07%
K	0	0.01%	0.05%	6.32%	92.34%	5.76%
M	0.19%	0.02%	0.01%	0.02%	1.64%	94.16%

通过 XGBoost 自带函数，我们可以得到每类线指数所占权重的打分，这里只保留权重最大的 7 个特征，如图 1。

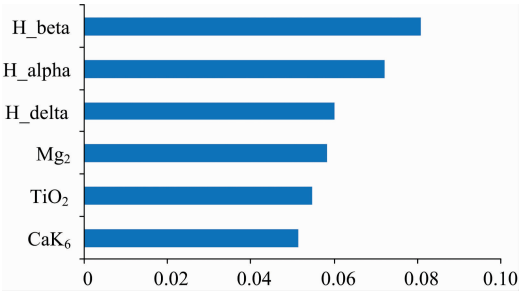

图 1 XGBoost 对特征的打分

Fig. 1 The scores given by XGBoost

这个评分会随着参数不同而发生变化，但总体差异不会很大。没有选取 H_alpha，是因为某些数据 H_alpha 存在缺失情况。下面分别以这些特征作为横纵坐标，每一类取 1 000 个最高信噪比的光谱数据绘制散点图和边缘分布图如图 2。

3 结 论

对 411 万光谱进行了基于 XGBoost 的 MK 分类，和 LAMOST 结果相比，达到了较高的正确率。除了约四分之一的 B 型分类错误，其他正确率均在 85%以上。

同时得到了 A，F，G，K，M 的分类决策树，这里我们以 A 型星为例，挑选了每一类相对应的可能性最大的 7 个分支。图 3 即为 A 型星的分类决策树图。其中左向箭头为 YES，右向箭头为 NO，菱形方框内的参数即为表 1 中所述的线指数及其取值范围，长方形为输出的可能性较大的结果，圆形为停止，即该分支可能性相对较小而被舍去。图 3 中的“P”为程序认为该分支为概率最大的一个分支。

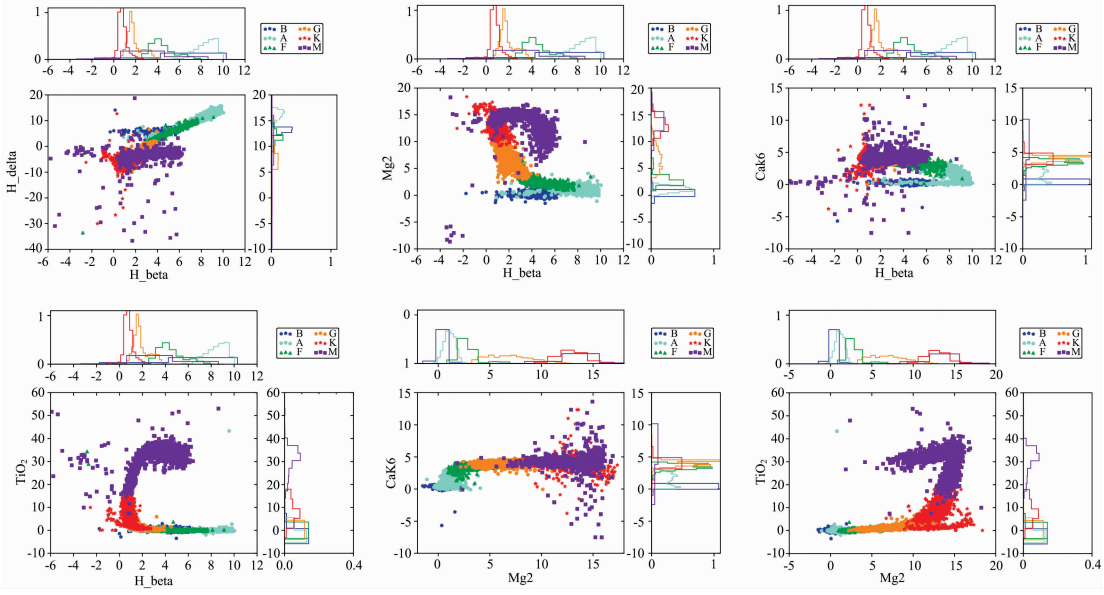

图 2　XGBoost 所得到的线指数分布图

Fig. 2　The distribution of the MK classes of the test data resulting from an XGBoost applied to the parameter space defined by line indices

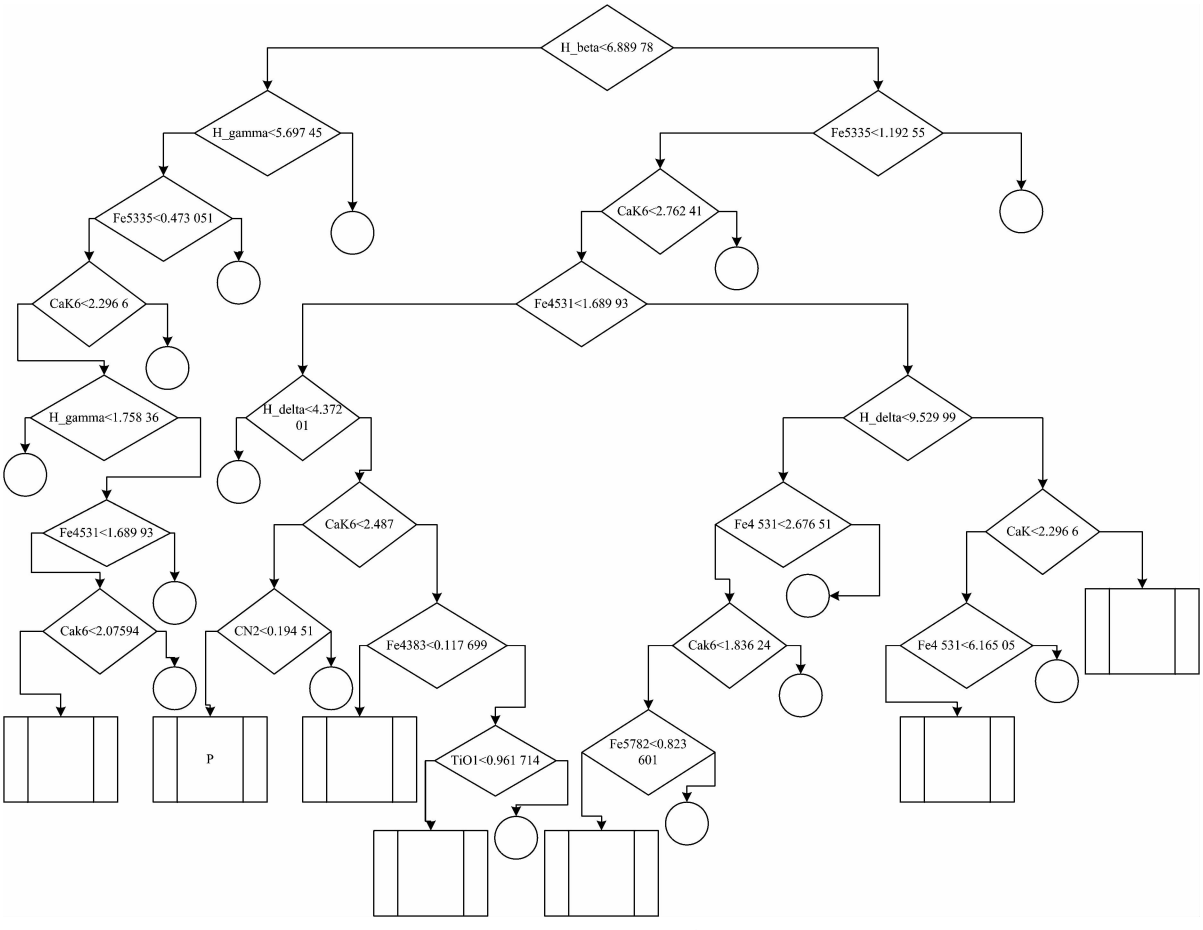

图 3　A 型星分类决策树

Fig. 3　The decision tree of A type star

可以看出，判别决策树基本符合已知结论，A 型星的特征谱线是很强的巴尔末线，特别是 H_beta 和 H_gamma。例如，绝大多数 A 型星的特征谱线之一的 H_beta＞6.889 78；而对于巴尔末线和 Fe4531 线均较弱情况下，即 H_beta＜6.889 78 并且 H _ gamma ＜ 5.697 45，算法以 CaK6 在 2.075 94的分界线，取小于此数值的部分，作为与 F0 等类型的分界面。同样，对于带有金属 Fe 线且 Fe4531＜1.689 93，位于 B 和 A 之间，以 H_delta＞4.372 01作为分界线；对于 Fe4383＞0.117 699，以 TiO1＜0.961 714区分含有 Fe 线丛 B9，M 和 A0。

References

[1] Luo Ali, Zhao Yongheng, et al. Research in Astronomy and Astrophysics, 2015, 15(8): 1095.

[2] Schierscher F, Paunzen E. Astronomische Nachrichten, 2011, 332(6): 597.

[3] Hampton E J, Medling A M, Groves B, et al. Monthly Notices of the Royal Astronomical Society, 2017, 470: 3395.

[4] Liu Chao, Cui Wenyuan, et al. Research in Astronomy and Astrophysics, 2015, 15(8): 1137.

[5] Du Changde, Luo A, Yang H. New Astronomy, 2017, 51: 51.

[6] Worthey G, Faber S M, Gonzalez J Jesus, et al. The Astrophysical Journal Supplement Series, 94(2): 687.

[7] WANG Guang-pei, PAN Jing-chang, YI Zhen-ping, et al(王光沛，潘景昌，衣振萍，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(8): 2646.

[8] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016. 785.

[9] Gray R O, Corbally C J. Stellar Spectral Classification. Princeton University Press, 2009. 160.

XGBOOST Based Stellar Spectral Classification and Quantized Feature

ZHANG Xiao[1,2], LUO A-li[1*]

1. Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China

2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Star spectral classification is a foundational work of stellar research. The Morgan-Keenan (MK) classification system which was developed in 1970s is the most widely used classical classification system. However, MK based interactive decision classification system has some difficulties when dealing with massive quantity of astronomical spectral data. Nowadays the most widely used method of automatically classification is template match which neglects measuring the spectral line. As a result, one of the most popular topics is how to extract features from massive data objectively and precisely and to apply the features for making classification decisions. In this paper, we processed the spectral data of LAMOST DR4 stars to obtain the line index as input data and used the official released labels of the spectrum as outcome. The XGBoost algorithm was applied to automatically classify the stellar spectra and rank the features. In this way, the identified and potential line indices which are sensitive to classification were revealed. Firstly, we labeled and selected the spectral data of stars with B, A, F and M by LAMOST high signal-to-noise ratio (S/N＞30) with the sample size amounting to around 41. 4 million. Then, the line indices of spectral data was calculated to reduce the dimension and to filter out the redundant information. Secondly, the processed star spectral data were randomly divided to a training set and a test set. By modifying the parameters, the required classification decision tree model was fitted by training set using XGBoost algorithm and the stability and availability of the model were validated by test set to avoid over-fitting. In the meantime, the classification features were extracted by the algorithm's own function. Finally, the branch with the highest probabilities was selected as the final decision tree model. Through experiments, it is shown that the XGBoost model has a better performance in self-adaptability under fixed parameters with less affection in data sets and the overall accuracy rate as high as 88.5%. Moreover, the output classification decision tree is more consistent with identified features and the numerical characteristics of spectrum and its corresponding range are obtainable through the model. This would shed light on providing quantitative rules for evaluating classification decision trees with numerical spectral features.

Keywords Spectral classification; Line index; XGBoost; Decision tree; LAMOST

* Corresponding author

(Received Sep. 7, 2018; accepted Jan. 18, 2019)