基于集成树的M型星光谱分类

王 晶[1]，衣振萍[1*]，岳丽丽[1]，董慧芬[1]，潘景昌[1]，卜育德[2]

1. 山东大学(威海)机电与信息工程学院，山东 威海 264209
2. 山东大学(威海)数学与统计学院，山东 威海 264209

摘 要 在赫罗图中，M巨星位于红巨星的顶端，是由类太阳的主序星逐渐演化而成的最明亮的一类恒星。M巨星的研究对于理解银河系，特别是银河系晕的性质至关重要。中低分辨率的M巨星光谱，常因为特征不显著、噪声影响等因素而与M矮星的光谱混在一起，不易区分。现有研究一般利用 CaH_2+CaH_3 *vs.* TiO_5 分子谱指数初步筛选M巨星光谱候选体，再通过人眼检查确认。但这种方法仅利用了三个巨星相关的分子带指数，没有利用识别M巨星的其他光谱特征，可能会由于噪声对指数的污染而导致分类错误。而且，人眼检查数量众多的光谱不仅耗时而且检查质量依赖于人的经验，可靠性无法得到保证。LAMOST望远镜自2011年开始先导巡天到2017年6月，已经发布了900多万天体的光谱，最新释放的光谱数据DR5包含了52万的M型星光谱数据，需要采用自动、准确、有效的方法来区分其中不同光度级的M子样本。本研究利用集成树模型分类M巨星和M矮星光谱，分别采用随机森林、GBDT、XGBoost和LightGBM算法，构建区分M巨星和M矮星的光度分类器。四种分类器的测试准确率分别达到97.23%，98%，98.05%和98.32%。实验表明LightGBM模型比其他三种集成树模型准确率更高，训练时间更少，分类效率更高。对分类器模型获取到的重要特征分析的结果表明，集成树算法有效提取并表达了用于区分M巨星和M矮星的结构性特征，模型提取到的重要特征不仅包括原子线或分子带吸收的波长位置，还包含了它们相邻的伪连续谱，这与传统上计算指数所需要特征波长和伪连续谱是一致的。相比于传统M巨星和M矮星分类方法，集成树模型能够采用光谱中的多个重要特征组合进行分类，避免仅依赖某一种特征易受噪声影响而得出错误的分类结果。研究结果表明集成树算法在巨星识别过程中具有显著优势，完全可以替代传统上只利用CaH和TiO指数的巨星光谱判别方法。基于集成树模型对M型星光谱的分类研究，为LAMOST高效、准确地处理海量天体光谱提供了有效的方法。随着LAMOST巡天项目不断开展，积累的M巨星和M矮星样本将为研究银河系的结构和演化提供重要的数据基础。

关键词 M巨星；集成树；光谱分类；特征提取

中图分类号：P144.1 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)07-2288-05

引 言

郭守敬望远镜[1](large sky area multi-object fiber spectroscopy telescope，LAMOST，大天区面积多目标光纤光谱天文望远镜)是一架视场为5度横卧于南北方向的中星仪式反射施密特望远镜。在大规模光学光谱观测和大视场天文学研究方面，处于国际领先地位。LAMOST先后发布了包括先导巡天[2]和正式巡天的光谱数据DR1[3]，DR2，DR3和DR4数据集，最新释放的光谱数据DR5包含了52万的M型星光谱数据[4]。其中部分M巨星光谱由于跟M矮星的光谱特征区分不明显或受噪声影响，混杂在M矮星光谱中，不利于后期M矮星和M巨星样本的选择和科学研究。因此，需要先把巨星光谱识别出来。

巨星光谱识别的一般方法是计算光谱的几个关键特征指数，比如NaI，TiO和CaH指数，然后根据经验公式筛选[5]，最后通过人眼检查确认。然而用指数分类，没有综合利用整条光谱的特征，可能会由于噪声对指数的污染而导致分类错误。而且人眼检查数量众多的光谱不仅耗时而且检查质量依赖于人的经验，可靠性无法得到保证。图1展示了一条混在

收稿日期：2018-06-06，**修订日期**：2018-10-28

基金项目：国家自然科学基金项目(11603014，11603012)和山东大学青年学者未来计划(2016WHWLJH09)资助

作者简介：王 晶，女，1997年生，山东大学(威海)机电与信息工程学院本科生 e-mail：wangjing7dhr@163.com

*通讯联系人 e-mail：yizhenping@sdu.edu.cn

LAMOST M 矮星星表中的 M 巨星光谱。黑色光谱是 LAMOST 望远镜观测到的 M 巨星光谱，红色是一条 M4 光谱型的 M 矮星光谱，二者的光谱大致形态相似，仅在几处波长位置人眼能分辨出差别。

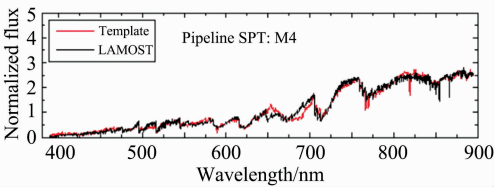

图 1　一条混杂在 M 矮星中的 M 巨星光谱

Fig. 1　A spectrum of M giant mixed in M dwarf

人工智能的发展使得处理数据可以高度自动化，许多机器学习算法处理数量大、维度高的光谱数据有很好的效果[6-7]。决策树模型在增长中产生高度自适应非线性的模型可能导致过拟合，而集成模型将非稳定的决策树模型组成一个集合以提高预测性能，即由多个弱分类器构成一个强分类器[8]。集成树模型在脉冲星群分类的研究中表现优秀[9]，在多种特征组合的自动提取中，较之于逻辑回归和支持向量机模型也具有优势[10]。

本研究调查随机森林(radom forest，RM)、梯度提升决策树(gradient boosling desision tree，GBDT)、极端梯度提升树(extreme gradient boosting，XGBoost)、LightGBM(light gradient boosting machine)四种集成树模型对 M 巨星和 M 矮星光谱分类的效果，并分析分类过程采用的重要特征与传统方法的差异。

1　M 型星光谱数据

从 LAMOST 第四年被分为 M 型的光谱中，去除坏谱、可能的双星光谱、K 型星光谱还有其他的奇异光谱后，剩下了97 849条，借助于 LAMOST 的 M 型星分类 pipeline，通过人眼检测确认，得到了巨星7 236条。为避免训练样本类型的不均衡，从矮星中随机抽取了包含 M0—M9 所有光谱次型的 M 矮星样本光谱数据 7 601 条，使巨星和矮星在最终样本中具有相似的比例。然后从巨星和矮星两个样本分别均匀采样(cvpartition)约 50%的光谱，组成了 7 349 条光谱的混合样本，将混合样本 70%即 5 145 条光谱用作训练集，30%即 2 204 条光谱留作测试集。测试样本中巨星在 r 和 i 波段的平均信噪比为 148.42 和 299.51，矮星的平均信噪比为 24.17 和 49.97。

LAMOST 的光谱数据存储于 Fits 格式文件中。首先读取数据并对数据进行预处理。把读出的光谱数据在 500～895 nm 范围内统一插值，间隔 0.1 nm。其次进行数据的归一化。最后将经过处理的数据存储到矩阵中，用来训练和测试集成树模型。

2　算法描述与调参

对四种集成树算法采用网格调参法，对决策树弱分类器个数、决策树最大深度、学习步长等多个参数进行网格搜索，确定最优参数，利用最优参数进行训练和测试，从多个方面对比分析测试结果。

2.1　随机森林算法

随机森林[11](random forest，RF)采用自助采样(bootstrap)技术，每棵树提供一个分类结果作为投票的依据，最终选择得票多的分类结果[8]。随机森林具有学习过程快、无需数据的归一化、易并行化等优点，已经成功应用于医学[8]、天文等多个领域。随机森林分类器选取的参数如表 1 所示。

表 1　随机森林模型参数列表

Table 1　Random forest model parameter list

Property	Value
n_estimators	40
max_depth	10
min_samples_split	9
min_samples_leaf	1

2.2　梯度提升决策树算法

梯度提升决策树[12](gradient boosting decision tree，GBDT)以决策树集合的形式产生预测模型，它通过梯度提升算法，每次在减少残差的方向建立新的决策树，提高预测准确性。GBDT 模型在天体物理学和粒子物理学方面均有广泛应用[13-14]。本实验中最终所使用的参数如表 2 所示。

表 2　GBDT 模型参数列表

Table 2　GBDT model parameter list

Property	Value
n_estimators	400
learining_rate	0.2
max_depth	4
random_state	10

2.3　极端梯度提升树算法

XGBoost[15](Extreme Gradient Boosting)是一种基于 GBDT 的集成树模型，实现了分裂节点寻找的近似算法，能够分布式处理高维稀疏特征[16]，并具有速度快、高准确度、不易过拟合的优点。最终所使用的参数如表 3 所示。

表 3　XGBoost 模型参数列表

Table 3　XGBoost model parameter list

Property	Value
n_estimators	1 000
learining_rate	0.1
max_depth	5
min_child_weight	1
seed	27

2.4　LightGBM 算法

LightGBM[17]是微软 2017 年推出的新的分布式梯度提

升框架，采用叶子分裂(leaf-wise)生长策略，在保证高效率分裂的同时防止过拟合。LightGBM 算法具有可并行化、效率高、占用内存小、精度大和提供可扩展解决方案等优势。在调参后使用的参数如表 4 所示。

表 4 LightGBM 模型参数列表

Table 4 LightGBM model parameter list

Property	Value
num_leaves	31
learning_rate	0.05
feature_fraction	0.9
bagging_fraction	0.8
num_boost_round	486

3 实验效果分析

3.1 4 种集成树模型预测能力

实验对比了 4 种集成树模型的 Accuracy，Precision，Recall 和 F-measure 指标，如表 5 所示。结果表明四种集成树模型对光谱数据分类均有较好的效果，LightGBM 模型达到了 98.32%的准确率，在四种集成树模型中表现最佳。最后一项指标模型运行时间(CPU Time)综合多次运行时间计算所得，可以看出 LightGBM 模型运行时间远小于其他集成树模型，分类效率高，适合在更大规模数据中得到推广应用。图 2 展示了 LightGBM 模型的混淆矩阵，可以看出测试样本中有 1 120 个矮星样本和 1 084 个巨星样本。其中 1 105 个矮星和 1 060 个巨星光谱被正确分类；有 15 个矮星光谱被误分到巨星类别，24 个巨星光谱被误分到矮星类别。光谱的误分多因为光谱质量差、信噪比低导致仅从光谱本身难以判定类别。

表 5 分类结果对比

Table 5 Comparison of classification results

Models	Accuracy /%	Precision /%	Recall /%	F-measure /%	CPU time/s
RF	97.23	97.27	97.23	97.23	37.36
GBDT	98.00	98.01	98.00	98.00	42.84
XGBoost	98.05	98.06	98.05	98.05	52.01
LightGBM	98.32	98.33	98.32	98.32	6.24

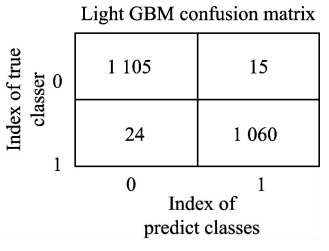

图 2 LightGBM 模型的混淆矩阵

Fig. 2 Confusion matrix of LightGBM model

3.2 重要特征

集成树模型在多种特征组合的自动提取中具有优势[17]，无需人工测量指数、提取特征，集成树模型自动发现巨星和矮星的光谱差异，给出分类结果。根据实验中得到的 XGBoost 模型和 LightGBM 模型的特征重要性排序，将二者排名前 100 的重要特征位置分别标注在光谱图上，如图 3 和图 4 所示。图(a)为一条实测的 M 矮星光谱，图(b)一条 M 巨星光谱。光谱上点的大小表示特征重要性评分的高低。越大代表分值越高，在分类决策中起到更加重要的作用。

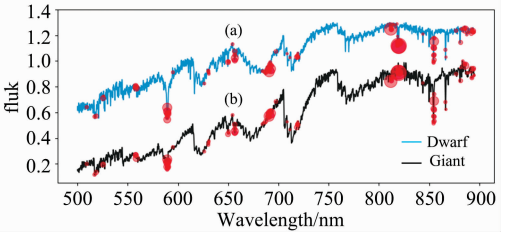

图 3 XGBoost 模型的前 100 特征

(a)：M 矮星；(b)：M 巨星

Fig. 3 Top 100 features of XGBoost model

(a)：M giants；(b)：M dwarfs

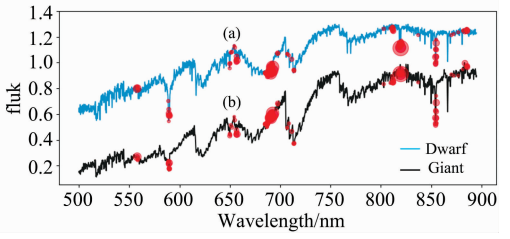

图 4 LightGBM 模型的前 100 特征

(a)：M 矮星；(b)：M 巨星

Fig. 4 Top 100 features of LightGBM model

(a)：M giants；(b)：M dwarfs

可以看出两个模型都提取到了区别巨星和矮星的重要特征：NaI，CaH 分子吸收带，在构筑树的过程中，被多次用做分裂节点，因此具有很高的重要性评分(importance score)；模型提取到的重要特征不仅包含原子线或分子带，还包含了与之相邻的伪连续谱，这与传统上计算指数所需要特征波长和伪连续谱是一致的。算法自动采用了光谱中多个特征的组合，避免仅依赖某一种特征容易受噪声影响而得出错误的结果。

4 结 论

使用集成树模型对 LAMOST 的 M 型星光谱数据进行分类，经过实验测试可知，随机森林、GBDT、XGBoost 和 LightGBM 四种模型分类准确率均可达 97%以上，这些集成算法都能够自动采用光谱中多个重要的特征组合进行分类，避免仅依赖某几种特征，易受噪声影响而得出错误结果。而 LightGBM 模型较之于其他 3 种模型，具有更高的分类准确率，模型训练时间远小于其他集成树模型，分类效率高，适合在更大规模数据中得到推广应用。

References

[1] Cui X Q, Zhao Y H, Chu Y Q, et al. Research in Astronomy and Astrophysics, 2012, 12(9): 1197.
[2] Luo A L, Zhang H T, Zhao Y H, et al. Research in Astronomy and Astrophysics, 2012, 12(9): 1243.
[3] Luo A L, Zhao Y H, Zhao G, et al. Research in Astronomy and Astrophysics, 2015, 15(8): 1095.
[4] ZHAO Yong-heng(赵永恒). Physics(物理), 2015, 44(4): 205.
[5] Zhong J, Lepine S, Li J, et al. Research in Astronomy and Astrophysics, 2015, 15(8): 1154.
[6] Yi Z, Luo A, Song Y, et al. The Astronomical Journal, 2014, 147(2): 33.
[7] Bates S D, Bailes M, Barsdell B R, et al. Monthly Notices of the Royal Astronomical Society, 2012, 427(2): 1052.
[8] Ichikawa D, Saito T, Ujita W, et al. Journal of Biomedical Informatics, 2016, 64: 20.
[9] Devine T R, Goseva-Popstojanova K, McLaughlin M. Monthly Notices of the Royal Astronomical Society, 2016, 459(2): 1519.
[10] Li N, Yu Y, Zhou Z H. Diversity Regularized Ensemble Pruning. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2012: 330.
[11] Breiman L. Machine Learning, 2001, 45(1): 5.
[12] Friedman J H. Annals of Statistics, 2001, 29(5): 1189.
[13] Roe B P, Yang H J, Zhu J, et al. Nuclear Instruments and Methods in Physics Research, 2005, 543(2-3): 577.
[14] Möller A, Ruhlmann-Kleider V, Leloup C, et al. Journal of Cosmology and Astroparticle Physics, 2016, 2016(12): 8.
[15] Chen T, Guestrin C. Xgboost: A Scalable Tree Boosting System. Proceedings of the 22nd Acmsigkdd International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785.
[16] Kadiyala A, Kumar A. Environmental Progress & Sustainable Energy, 2018, 37(2): 618.
[17] Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree (c). Advances in Neural Information Processing Systems, 2017. 3149.

Spectral Classification of M-Type Stars Based on Ensemble Tree Models

WANG Jing[1], YI Zhen-ping[1*], YUE Li-li[1], DONG Hui-fen[1], PAN Jing-chang[1], BU Yu-de[2]

1. School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai, Weihai 264209, China
2. School of Mathematics and Statistics, Shandong University, Weihai, Weihai 264209, China

Abstract Located at the top of the red giants in Hertzsprung-Russell diagram, M giants are the brightest stars that evolved from the sun-like main sequence stars. The study of M giants is crucial to understand the Milky Way, especially the Galactic haloes. The spectrum of an M giants in medium and low resolution is often mixed with spectra of M dwarfs because of insignificant features, noise effects, and other factors. Previous studies often used the molecular index of CaH_2+CaH_3 *vs.* TiO_5 to search for M giant candidates, then checked them with human eyes. However, this method only used three important molecular band indices associated with giants, without using other spectral features to identify the M giants, which may cause misclassification due to noise pollution of the index. Moreover, relying on human eyes to check a large number of spectra is time-consuming, and the quality of the inspection dependings on people's experience and its reliability is not guaranteed. Since 2011, LAMOST has released more than 9 million celestial spectra. The latest spectral data product data release 5(DR5) contains 520 000 M-type spectral data, which needs an automatic, accurate and effective method to distinguish the M sub-samples of different luminosity levels. This study uses four ensemble tree models: Random Forest, GBDT, XGBoost, and LightGBM to construct classifiers that distinguish between M giants and M dwarfs. The accuracy of four classifiers is 97.23%, 98%, 98.05%, and 98.32%, respectively. Experiments showed that LightGBM has higher accuracy and less training time when compared to the other threemodels. The analysis of important features obtained by the classifier models showed that ensemble tree model can efficiently extract and express the structural features that distinguish M giants and M dwarfs. These features include not only the atomic lines, molecular bands, but also their adjacent pseudo-continuum spectrum, which is consistent with the features and pseudo-continuum spectra that we traditionally need to calculate the indices. Compared to the traditional classification methods, ensemble tree can use the combination of tens or hundreds important features in the spectrum rather than only several features to avoid misclassification affected by noises. The results of this study showed that the ensemble tree algorithm has significant advantages in the process of M giant recognition, and it can completely replace the traditional M giant spectral discrimination method using only CaH and TiO indices. In this study an effective method has been provided for LAMOST to efficiently and effectively process the massive celes-

tial spectra. As the LAMOST survey continues, more and more M spectra will be accumulated, which provides massive data for the studies of structure and evolution of the Milky Way.

Keywords M giants; Ensemble tree; Spectral classification; Feature extraction

(Received Jun. 6, 2018; accepted Oct. 28, 2018)

* Corresponding author