基于 XGBoost 的铝合金 LIBS 光谱分类识别方法

李晨阳[1,2,3]，陈雄飞[1,2,3]，张 勇[4]，王亚文[1,2,3]，田中朝[4]，
王世功[4]，赵珍阳[4]，刘 英[1,2,3]，刘鹏宇[1,2,3]*

1. 有研科技集团有限公司国家有色金属及电子材料分析测试中心，北京 100088
2. 国合通用测试评价认证股份公司，北京 101400
3. 北京有色金属研究总院，北京 100088
4. 山东东仪光电仪器有限公司，山东 烟台 264670

摘 要 铝合金作为重要的金属材料，广泛应用于各领域，但大量的铝合金废料却难以进行分类回收。二次资源的回收利用是我国工业绿色、可持续发展的助推器，如何快速、简便地对铝合金废料进行识别分类则成为了铝合金废料回收利用的先决条件。激光诱导击穿光谱(LIBS)是近年来发展快速的一种分析技术，具有快速、全元素分析、实时、原位、远距离检测等优点，已广泛应用于塑料、土壤、肉类、钢铁等识别研究，大多采用最小二乘判别分析法、簇类独立软模式、人工神经网络、支持向量机、随机森林等算法来建立模型。基于迭代型树的 XGBoost 算法具有正则化、并行处理运算、内置交叉验证和高度的算法灵活性等优势，其模型结构相对简单、运算量较小，且准确率较高，成为近年来机器学习中极受欢迎的算法，因而被广泛应用。基于六种铝合金样品的 600 组光谱数据，根据 NIST 原子发射光谱数据库进行光谱特征提取，确定光谱特征谱线的分类依据。利用 XGBoost 算法进行自动分类及排序，将处理后的光谱数据随机划分为训练集和测试集，通过训练集构建算法模型，提取其分类特征；利用测试集检验模型的稳定性和可用性，防止出现过拟合。XGBoost 在固定参数下得到的模型具有一定的自适应性，较少受数据集的影响，总体准确率可达 96.67%。其分类特征与已知的元素含量信息相吻合，证明了基于光谱的特征谱线数据，可为分类识别提供参考；同时还可根据 XGBoost 生成的特征评分来对光谱谱线特征的重要性进行排序。实验结果表明，LIBS 可用于不同种类铝合金的快速识别，为废弃金属的分类回收提供了一种新的技术。

关键词 铝合金；激光诱导击穿光谱；识别；XGBoost；决策树

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)02-0624-05

引 言

随着我国工业的发展，铝合金材料因其优异的性能而被广泛应用于各个领域，但大量的铝合金废料的分类回收成了工业难题。当前多采用密度分离技术来对铝合金进行分离，但该技术有着很多缺陷，且铝合金之间的物理性能差异较小，其他传统分离技术如磁、涡流、颜色感应等分离效果不好，耗能严重[1]，因此需要新的识别方法来对进行铝合金分类回收。

激光诱导击穿光谱(laser induced breakdown spectroscopy，LIBS)是近年来快速发展的一门崭新的分析技术，因其能够快速、全元素分析、实时、原位、远距离检测等优点而受到广泛关注[2]。该技术利用高功率脉冲激光辐射聚焦在样品表面上，立即蒸发出少量样品，引发样品元素的雪崩电离，产生击穿效应。样品中各特征元素所产生的原子线和离子线，是进行定性鉴定的基础。将 LIBS 技术应用于合金金属废料回收，可以快速地对合金废料按不同成分进行识别和分类。

LIBS 技术在未经任何样品预处理的情况下原位使用时，灵敏度、重现性和准确性无法保证；但在进行快速分类及对未知对象的成分与参考标准进行比较等方面，灵敏度和准确性问题则起着较小的作用。由于 LIBS 的快速性和鲁棒性，已有报道在 LIBS 定性识别和定量分析中使用偏最小二乘判

收稿日期：2019-12-24，**修订日期**：2020-04-08

基金项目：国家新材料测试评价平台建设项目重点项目(TC170A5SU/分包号：1)资助

作者简介：李晨阳，1992 年生，北京有色金属研究总院硕士研究生 e-mail：lichenyang3000@163.com

*通讯作者 e-mail：liupengyu@cutc.net

别分析法(PLS-DA)、主成分分析法(PCA)、人工神经网络(ANN)、支持向量机(SVM)、随机森林算法(RF)等[3-4]。金属合金类样品因其所含杂质元素种类多、含量低等特点，成为 LIBS 进行识别研究的难点。现有金属合金类样品识别分析多集中在钢铁样品的分类研究。Prasanthi Inakollu[5] 等通过 ANN 来预测铝合金的元素浓度，预测效果优于传统方法。Liang[6] 等将 LIBS 同 SVM 相结合，实现了不同牌号钢材的快速识别，通过改进 SVM 模型算法，满足多分类目的的同时，避免了冗余信息对识别的干扰，提高了识别正确率。Xu[7] 等通过蒙特卡洛有放回的重采样技术，最大限度的提取有效信息来进行变量选择，提升了 SVM 对钢铁样品的分类效率，获得了更好的稳定性和普遍性。Campanella[8] 等将 LIBS 同“模糊化”的 ANN 相结合，在模拟工业环境的条件下实现对铝合金的有效分类。

基于树的 XGBoost 算法是一类有监督的机器学习算法，具有正则化、并行处理运算、内置交叉验证和高度的算法灵活性等优势[9]，其模型结构简单、运算量较小，且准确率较高，成为近年来机器学习挑战中极受欢迎的算法。此外，XGBoost 算法克服了 ANN 算法复杂的、难以解读的黑箱问题，能够对算法提供详细解释；同时较 SVM 等算法，具有更高的准确率[10]。对于光谱分析，XGBoost 算法能够直接提取谱线特征信息进行预测，对 LIBS 识别材料提供很大的帮助。

1　XGBoost 算法原理

XGBoost 是一种迭代型树类算法，将多个弱分类器一起组合成一个强的分类器，是梯度提升决策树(GBDT)的一种实现[11]。XGBoost 是一种强大的顺序集成技术，具有并行学习的模块结构来实现快速计算，其通过正则化来防止过度拟合，可以生成处理加权数据的加权分位数草图[12]。具体算法步骤如下：

目标函数

$$l(y_i, \hat{y}_i) = ((y_i - \hat{y}_i)^2$$

在 XGBoost 中，每棵树需逐个加入，以期效果能够得到提升。

$$\hat{y}_i^{(0)} = 0;\ \hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i);\ \cdots;$$

$$\hat{y}^{(t)} = \sum_{k-1}^{t} f_k(x_i) = \hat{y}^{(t-1)} + f_t(x_i)$$

如果叶子的节点太多，模型的过拟合风险就会增大。所以在目标函数中加入惩罚项 $\Omega(f_t)$来限制叶子节点个数。

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2$$

式中，γ 为惩罚力度；T 为叶子的个数；ω 为叶子节点的权重。

完整的目标函数即为

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

记

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

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

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

上式可作为树的子叶分数，树的结构随着分数的增加而优异。且一旦分裂后的结果小于给定参数的最大所得值，算法将停止增长子叶深度[13]。

2　实验部分

2.1　仪器及参数

采用自主设计搭建的 LIBS 系统，采用调 Q 开关 Nd：YAG 脉冲激光器(Quantel Brilliant B，波长 1 064 nm，脉冲宽度 8 ns，最大重复频率 10 Hz)。位移平台设计为手动调节的“弓”字型运动方式，以避免激光脉冲重复作用于样品同一位置。激光经反射镜、平凸透镜和平凹透镜聚焦到样品表面，激发出等离子体。受激发的等离子体产生的光谱信号通过收集器耦合，进入直径为 100 μm 的光纤中，传输至光谱仪(AVANTES，3 通道拼接，波长范围 170～950 nm，分辨率全波长范围内小于 0.1 nm)。该系统的光路设计如图 1。

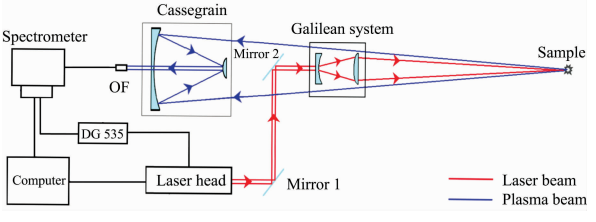

图 1　LIBS 实验装置示意图

Fig. 1　The schematic diagram of LIBS experiment setup

2.2　样品

采用 6 种铝合金样品作为实验样品，其形态为直径 Φ30×5 mm 圆柱。样品编号及杂质元素含量如表 1 所示。样品固定在自制的样品盒内，以保证 LIBS 的激光激发位置一致。

实验采用的激光脉冲能量为 200 mJ，脉冲频率为 2 Hz，采集延迟时间为 5 μs，均在一个大气压的实验环境下进行。每个样品取 100 个测量点，在各测量点上分别测量 20 次取平均，以减小误差。每个样品测试前，前 5 个激光脉冲用于对样品表面进行剥蚀，以清除样品表面的氧化物等杂质。最终得到 6 个样品的 LIBS 光谱数据共 600 组。随机打乱各样本光谱数据，其中取 420 组光谱数据为训练集，180 组光谱数据为测试集。采用 Python 环境下的 pandas 包对数据集进行数据处理，用 Python 环境下 XGBoost 包对模型进行训练。

2.3　特征谱线分析

为降低测定过程中波动影响，将各杂质元素谱线强度同基体元素 Al 394.40 nm光谱强度相除，得到六种样品的相对光谱强度如图2所示，部分波段因光谱仪多通道而被重复

表 1 6 种待测铝合金样品的杂质元素含量(Wt. %)

Table 1 The values concentration (Wt. %) of impurity elements in six alloys

编号 No.	Si	Fe	Cu	Mn	Mg	Cr	Ni	Zn	Ti
1#	0.541	0.206	0.040	0.864	1.170	0.035	0.059	0.297	0.114
2#	1.010	0.363	0.228	0.522	0.693	0.161	0.043	0.226	0.055
3#	1.240	0.428	0.487	0.233	0.933	0.354	0.027	0.094	0.026
4#	0.235	0.697	0.705	1.180	0.411	0.218	0.089	0.166	0.066
5#	1.580	0.848	0.078	1.480	0.068	0.288	0.115	0.041	0.008
6#	0.052	0.074	0.927	0.042	1.460	0.039	0.013	0.361	0.160

采集。考虑到以生成模型、统计分析为主要目标，应尽可能多地选取特征元素，根据 NIST 原子发射光谱数据库选择元素的特征谱线，根据元素含量进行曲线拟合，同时避开光谱重复波段，共筛选出 10 种 LIBS 光谱数据的特征谱峰，如表 2 所示。选取这些谱线数据使得原本的连续谱简化为 10 组光谱的特征谱峰，降低了原数据量，提高了模型的运行速率。

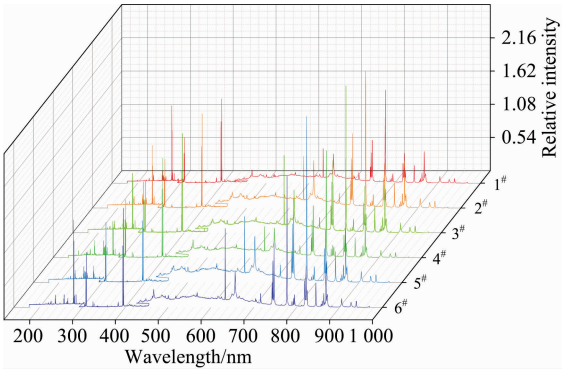

图 2 六种铝合金的典型光谱矩阵图

Fig. 2 The spectral matrices of six kinds of standard aluminium alloy

表 2 分析所用谱线

Table 2 Spectral lines for analysis

分析元素	分析谱线(NIST)/nm
Si	250.69, 251.61
Fe	259.94
Cu	324.75, 327.40
Mn	259.37
Cr	357.87
Ni	232.00
Zn	213.86
Ti	334.94

表 3 XGBoost 模型的超参数

Table 3 Super parameters of XGBoost model

超参数	取值
迭代模型	gbtree
损失函数	Multi：Softmax
学习率	0.1
树的深度	6
L_2 正则化参数	2
类别数	6

2.4 XGBoost 建模

设置训练集和测试集的数据比例为 7∶3，通过对 XGBoost 算法超参数进行调整，使得模型模型更加稳定，具体参数设置如表 3 所示。

3 结果与讨论

通过 XGBoost 自带函数算得每种谱线的权重评分如图 3 所示。结合表 4 中各样品杂质元素含量变异系数可知，XGBoost 算法对变异系数高的元素(如 Cu，Ti，Mn 和 Cr 等)赋予权重较高，有利于合理区分样品类别。该权重值会随着 XGBoost 超参数的不同而发生较小的变化，但总体不会有很大的改变。

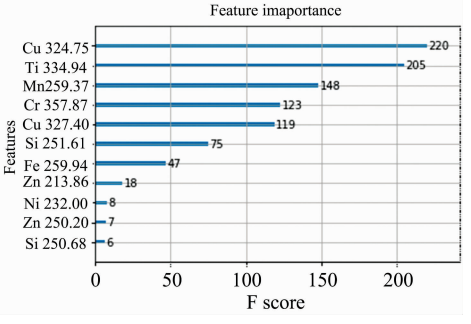

图 3 XGBoost 对特征的评分

Fig. 3 The feature importance scores given by XGBoost

表 4 6 种样品杂质元素含量的变异系数

Table 4 The Coefficient of Variation of all analytical elements

分析元素	变异系数
Cu	0.871 2
Ti	0.792 2
Mn	0.772 1
Si	0.770 4
Cr	0.712 8
Fe	0.670 4
Ni	0.668 1
Mg	0.643 2

下面分别以权重最高的四种元素(Cu，Ti，Mn 和 Cr)的相对强度值作为横、纵坐标，绘制散点图、直方图及等高线图如图 4。由散点图可以直观地看出，这四种元素相对强度值可以有效区分六种铝合金样品。从等高线图及直方图中可

以判断出，随着元素谱线特征权重的提高，各样品的集中趋势越明显，且样品之间重合区域越小；而较低的谱线特征权重下，各样品点分布较为分散，且不同样品之间重叠区域范围较大，不易区分样品。因此，XGBoost 算法可以有效地进行特征评价，判断出光谱中的特征谱线信息来进行样品判断和识别，克服了 ANN 等算法复杂的、难以解读的黑箱问题。

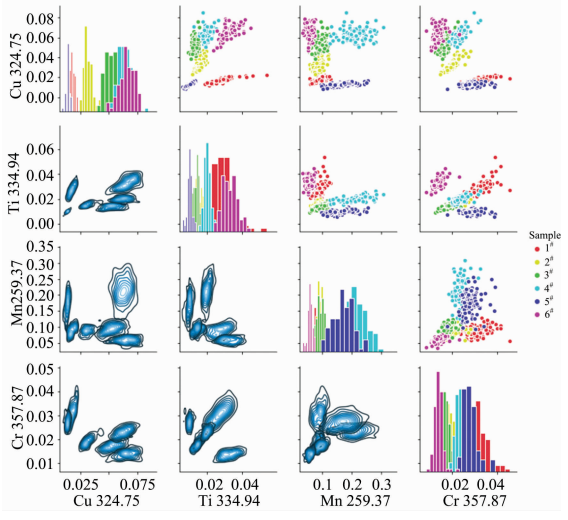

图 4　光谱数据在 XGBoost 的 4 种特征下的分布图

Fig. 4　Distribution map of spectral data under 4 featres of XGBoost model

实验得到 XGBoost 分类正确率如表 5 所示。从表 5 可以看出，样品识别正确率达到 96.67%，误报情况主要集中在 1# 及 3# 两类，其中 1# 与 5# 中的 Cu 含量接近，而 Cu 作为影响分类最大的因素，如果含量接近会使模型因分类界限接近而产生分类错误；而 3# 同 2# 相比，虽然 3# 中 Cu 的含量较 2# 多，但其他 Ti、Mn 等元素含量较少，在模型中所得分数较为接近而容易产生误判，因此这两类样品易产生误报现象。

如果训练集样本量足够多，模型将会更加的稳健，识别的准确率也会更高。模型将通过选用铝的非自吸收弱线(如 396.15 nm)来进行归一化；选用更多的特征谱线来进一步优化分类算法，以提高模型的识别准确率。

表 5　测试集识别结果

Table 5　Identification results of Test Sets

	1#	2#	3#	4#	5#	6#	准确率/%
1#	36				2		94.73
2#		22	1				95.65
3#		2	26	1			89.66
4#				35			100
5#					27		100
6#						28	100
总体							96.67

4　结　论

通过对 6 种铝合金进行 LIBS 光谱数据采集，选择 10 个特征波长的光谱强度作为特征变量，构建出铝合金的光谱矩阵，降低了模型负担。结合 XGBoost 算法对数据进行分析和建模，可对不同铝合金样品进行识别，识别准确率达到 96.67%，同时对元素谱线特征权重进行有效地区分。实验结果表明，LIBS 通过 XGBoost 算法可有效识别铝合金样品，为铝合金分类回收提供了一种新的技术。

References

[1] Koyanaka, Kobayashi K, Yamamoto, et al. Resourses Conservation and Recycling, 2013, 75: 63.

[2] Aderval S Luna, Fabiano B Gonzaga, Werickson F C da Rocha. Spectrochimica Acta Part B: Atomic Spectroscopy, 2018, 139: 20.

[3] Rosalba Gaudiusoa, Ebo Ewusi-Annana, Noureddine Melikechi, et al. Spectrochimica Acta Part B: Atomic Spectroscopy, 2018, 146: 106.

[4] Ke Liu, Di Tian, Xinxin Deng, et al. Journal of Analytical Atomic Spectrometry, 2019, 34: 1665.

[5] Prasanthi Inakollu, Thomas Philip, Awadhesh K. Rai, et al. Spectrochimica Acta Part B: Atomic Spectroscopy, 2009, 64: 99.

[6] Liang L, Zhang T, Wang K, et al. Applied Optics, 2014, 53(4): 544.

[7] Xu L, Liang L, Zhang T, et al. Analytical Methods, 2014, 6(20): 8374.

[8] Campanella B, Grifoni E, Legnaioli S, et al. Spectrochimica Acta Part B: Atomic Spectroscopy, 2017, 134: 52.

[9] Robert P Sherodan, Wei Ming Wang, Andy Liaw, et al. Journal of Chemical Information and Modeling, 2016, 56(12): 2353.

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

[11] Jin Zhang, Daniel Mucs, Ulf Norinder, et al. Journal of Chemical Information and Modeling, 2019, 59(10): 4150.

[12] Aman Agarwal, Liu Y A, Christopher McDowell. Industrial & Engineering Chemistry Research, 2019, 58(36): 16719.

[13] ZHANG Xiao, LUO A-li(张　枭，罗阿理). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(10): 3292.

Study on Identification Method Based on XGBoost Model for Aluminum Alloy Using Laser-Induced Breakdown Spectroscopy

LI Chen-yang[1,2,3], CHEN Xiong-fei[1,2,3], ZHANG Yong[4], WANG Ya-wen[1,2,3], TIAN Zhong-chao[4], WANG Shi-gong[4], ZHAO Zhen-yang[4], LIU Ying[1,2,3], LIU Peng-yu[1,2,3]*

1. National Analysis and Testing Center of Nonferrous Metals and Electronic Materials, GRINM Group Corporation Limited, Beijing 100088, China
2. China United Test & Certification Co., Ltd., Beijing 101400, China
3. General Research Institute for Nonferrous Metals, Beijing 100088, China
4. Shandong Dongyi Photoelectric Instruments Co., Ltd., Yantai 264670, China

Abstract As an important metal material, aluminum alloy is widely used in various fields, but a large amount of aluminum alloy waste is difficult to sort and recycle. The recycling of aluminum alloy resources is a booster for China's industrial green and sustainable development. How to quickly and easily identify and classify aluminum alloy waste has become a prerequisite for re-utilization. Laser-induced breakdown spectroscopy (LIBS) is an analytical technique that has developed rapidly in recent years. It has the advantages of fast, full-element analysis, real-time, in-situ, and long-distance detection. It has been widely used in plastics, soil, meat, steel, etc. For recognition research, most of them use the PLS-DA, SIMCA, ANN, SVM, Random Forest and other algorithms to build models. XGBoost algorithm has the advantages of regularization, parallel processing, built-in cross-validation, and high algorithm flexibility. Its model structure is relatively simple; it has a small amount of calculation and superior accuracy. It has become extremely popular in machine learning in recent years. Based on 600 sets of spectral data of six aluminum alloy samples, model extracts spectral features through machine learning to determine the classification. The processed spectral data is randomly divided into a training set, and a test set, and the XGBoost algorithm based on Decision Tree is used for automatic classification and sorting An algorithm model is constructed through the training set and its classification features are extracted; the test set is used to check the stability and usability of the model to qrevent over-fitting. The model obtained by XGBoost under fixed parameters has certain self-adaptability, is less affected by the data set, and the overall accuracy rate can reach 96.67%. Its classification characteristics are consistent with the known element content information, which proves that the characteristic spectral line data based on big data can provide a reference for classification identification; the importance of spectral line features can be ranked according to the feature score generated by XGBoost. Experimental results show that LIBS can be used for rapid identification of different types of aluminum alloys, and provides a new technology for the classification and recovery of waste metals.

Keywords Aluminum alloy; LIBS; Recognition; XGBoost; Decision Tree

(Received Dec. 24, 2019; accepted Apr. 8, 2020)

* Corresponding author