基于 CART 回归树的 LIBS 特征变量选择方法研究

尤 文[1]，夏阳鹏[1]，黄玉涛[1]，林京君[2*]，林晓梅[3*]

1. 长春工业大学电气与电子工程学院，吉林 长春 130012
2. 长春工业大学机电工程学院，吉林 长春 130012
3. 吉林建筑科技学院，吉林 长春 130012

摘 要 激光诱导击穿光谱技术(LIBS)用于检测时，由于谱线多且复杂，存在许多冗余的信息，这些都会对定量分析造成影响。因此，提取有效的特征变量在 LIBS 的定量分析中具有非常重要的意义。对 $CaCl_2$ 溶液中的 Ca 元素进行光谱特征选择方法分析，对比单变量模型、偏最小二乘回归和 CART 回归树定标模型的准确度和稳定性。针对水体表面的波动性较大，光谱稳定性差，同时光谱受基体效应和自吸收效应影响等问题，首先采用单变量模型得到的拟合系数(R^2)仅有 0.933 2，训练均方根误差(RMSEC)、预测均方根误差(RMSEP)和平均相对误差(ARE)分别为 0.019 2 Wt%，0.017 7 Wt%和 11.604%。经偏最小二乘回归优化后，模型 R^2 提高到 0.975 3，RMSEC，RMSEP 和 ARE 分别降低到 0.010 8 Wt%，0.013 Wt%和 7.49%。为了进一步提高定量分析的准确度，建立 CART 回归树定标模型。该方法在构建树模型时，通过平方误差最小化准则，从复杂的光谱信息中选取最优的特征变量组合做分类决策，从而建立 Ca 元素的定标曲线。通过 CART 回归树的变量选择，特征变量个数从 100 个减少到 6 个，变量的压缩率达到了 94%，显著降低了无关谱线的干扰，回归树模型的相关系数 R^2，RMSEC，RMSEP 和 ARE 分别为 0.997 5，0.003 5 Wt%，0.006 1 Wt%和 2.500%。相较于传统的单变量模型与偏最小二乘回归，CART 回归树模型具有更高的精度、更小的误差。通过对特征变量的有效筛选，剔除无关信号的干扰，显著降低了基体效应和自吸收效应对 LIBS 定量分析的影响，提高了定量分析的准确度和稳定性。

关键词 激光诱导击穿光谱；特征变量选择；CART 回归树；定量分析

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)10-3240-05

引 言

激光诱导击穿光谱(laser-induced breakdown spectroscopy，LIBS)技术是一种元素含量分析技术[1-2]，它具有原位、实时、快速、可远程、非接触、无需样品准备等优点[3-5]，可以分析元素周期表中的所有元素，并且可以对固体、液体、气体、气溶胶等任何状态下的物质进行检测[6-9]。但是，LIBS 光谱信息丰富，包含了大量的原子和离子谱线，实验重复性低，实验结果误差较大。在进行定量分析时，用单一的特征谱线定标，原始数据比较分散，拟合相关性不高，而且光谱利用率低，模型稳定性差。多变量分析可以扩展影响 LIBS 谱线的特征信息，在一定程度上减小样品的波动性，提高分析准确率[10]。但是，整个 LIBS 光谱数据稀疏且高维，大多数谱线与分析元素无关。此外，无关的冗余变量不仅会增加模型复杂程度，导致过拟合，而且会使模型学习到杂散的噪声信息，严重影响分析结果的准确性。因此，寻找一种高效的变量选择方法具有重要意义。

传统方法是根据光谱信息结合 NIST 数据库人为选择特征谱线作为分析变量，效率低，受主观因素影响较大。而且，手动选择无法识别元素间的相互作用，特征谱线容易受到基体效应影响。为了有效筛选 LIBS 特征光谱，减少定量分析误差，国内外学者对变量选择展开了大量的工作。吴宜青等用竞争性自适应重加权算法(CARS)选择 Cr 元素的特征变量，预测结果优于单变量、五变量和全波段模型[11]。胡丽等将 LIBS 与 PLS 相结合，分析了水中的 Pb 元素含量，结果表

收稿日期：2020-10-10，**修订日期**：2021-02-04

基金项目：国家重大科学仪器开发专项(2014YQ12035104)，吉林省科技厅项目(20180414017GH，20200403008SF)，吉林省发展改革委项目(2018C034-3)资助

作者简介：尤 文，1961 年生，长春工业大学电气与电子工程学院教授 e-mail：youwen@ccut.edu.cn

*通讯作者 e-mail：1124270941@qq.com；187049860@qq.com

明，PLS 适用于不同的水样，可以在一定程度上降低基体效应的影响[12]。郭恺琛等采用主成分分析载荷空间距离法筛选特征谱线，对矿物进行种类识别，识别精度达到了 92.8%，降低了识别难度[13]。Sun 等将 SelectKBest 算法用于 LIBS 特征变量的选择，发现其可以限制过拟合，有效提取重要的特征[14]。大量的研究证明，变量筛选技术可以有效减少基体效应的影响，提高 LIBS 定量分析精度。

LIBS 在用于液体检测时，样品波动性强，同一实验数据重复性较差[15]。为了验证 CART 回归树对特征谱线的选择能力，本文利用激光诱导击穿光谱技术结合 CART 回归树对溶液中 Ca 元素的含量进行检测。通过计算每个变量的重要性程度，选择对待测元素浓度贡献率较大的几个特征谱线作为分析变量，提高定量分析准确性。

1　实验部分

1.1　装置

本实验采用液体射流的实验方式，实验平台包含激光光路系统和液体射流系统两部分，如图 1 所示。光路系统主要有 Nd∶YAG 固体激光器(Surelite Ⅲ 10，USA Continuum)、光纤光谱仪(Avaspec-USB2，荷兰 Avantes)、数字延时脉冲发生器(DG645)等。激光经过格兰棱镜、反射镜、半波片和聚焦透镜(f=150 mm)聚焦到液柱表面。产生的等离子体经光纤探头收集，传输进入光谱仪，最终通过 PC 机得到光谱信息。液体射流系统主要由蠕动泵(Kamoer Lab UIP)、分液漏斗(喷嘴直径为 1 mm)和烧杯支架等组成。液体样品经过漏斗、乳胶管、蠕动泵和烧杯形成循环系统，整个系统放置在三维移动平台上，通过 X 轴和 Z 轴实时改变焦距以及聚焦点到喷嘴的距离。通过参数优化得到最佳的聚焦位置为液柱表面，焦点距离喷嘴的最佳距离为 2 mm。

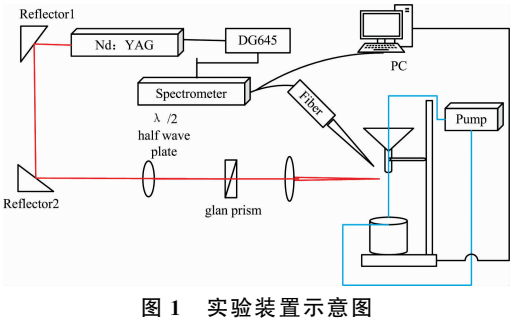

图 1　实验装置示意图

Fig. 1　Schematic diagram of experimental device

1.2　样品制备与数据准备

实验样品是利用母液稀释而成，标准溶质为北京化工厂生产的 $CaCl_2$，使用蒸馏水稀释母液，配置 7 种不同浓度梯度的 $CaCl_2$ 溶液，如表 1 所示。

采集光谱时，液面波动性较大，收集到的光谱重复性差。为了减少实验误差与不稳定性，本文对 7 个梯度进行多组实验，每组实验进行 6 次，取平均值作为定标的输入，将 37 组数据用于最终的定量分析。

表 1　标准样品中 Ca 元素浓度

Table 1　The concentration of Ca in standard samples

Sample number	Concentration of Ca/Wt%
1#	0.02
2#	0.05
3#	0.08
4#	0.1
5#	0.15
6#	0.2
7#	0.25

1.3　算法介绍

本文将 CART 算法中的回归树用于 LIBS 的定量分析，以平方误差最小化作为准则，进行特征变量的选择，逐步选择内部节点，从而生成回归树模型。其构建算法具体如下：

(1)选择最优变量 j 与变量切分点 s，求解

$$\min_{j,s}\left[\min_{c1}\sum_{x_i\in M_1(j,s)}(y_i-c1)^2+\min_{c2}\sum_{x_i\in M_2(j,s)}(y_i-c2)^2\right] \tag{1}$$

式(1)中：$c1=\text{average}(y_i\mid x_i\in M_{1(j,s)})$，$c2=\text{average}(y_i\mid x_i\in M_{2(j,s)})$ 遍历所有的特征 j 和切分点 s，即可找到最优的特征变量组合；

(2)用求得的分割点(j，s)将输入区域划分，并得到相应节点的输出值；

$$M_1(j,s)=\{x\mid x^{(j)}<s\}$$
$$M_2(j,s)=\{x\mid x^{(j)}>s\}$$
$$\hat{c}_m=\frac{1}{N_m}\sum_{x_i\in M_m(j,s)}y_i,\ x\in M_m,\ m=1,2 \tag{2}$$

(3)重复(1)和(2)，构建其他的子节点，直到满足停止条件；

(4)将输入空间划分为 M_1，M_2，…生成决策树。

为了限制回归树的规模，简化模型结构，需要对回归树进行剪枝。回归树的剪枝分为两部分。第一步是从生成的决策树底部开始剪枝，一直到根节点，如此生成一系列子树 $\{T_0,T_1,\cdots,T_m\}$；第二步是对所有的子树做交叉验证，选出效果最优的子树。

2　结果与讨论

2.1　单变量分析

由于 Ca Ⅱ 393.366 nm 谱线的轮廓清晰、波峰较为明显，而且受附近谱线干扰较小，因此可以作为单变量分析的分析谱线。将谱线的光谱强度与 Ca 元素浓度做线性回归，绘制真实浓度与预测浓度的拟合曲线如图 2 所示。单变量分析不会对数据做任何处理，真实的反映了原始数据的分布情况。从图中可以看到，数据分布比较分散，而且数据稳定性较低。拟合系数 R^2 只有 0.933 2，RMSEC，RMSEP 和 ARE 分别为 0.019 2 Wt%，0.017 7 Wt%和 11.604%。这可能是

由于激光与液体作用后，液体飞溅、波动，导致连续背景增加，无用的光谱信息增加，大大降低了实验的稳定性。其次可能存在着自吸收和基体效应，导致数据呈非线性分布。因此，单变量分析难以满足 LIBS 的定量分析要求。

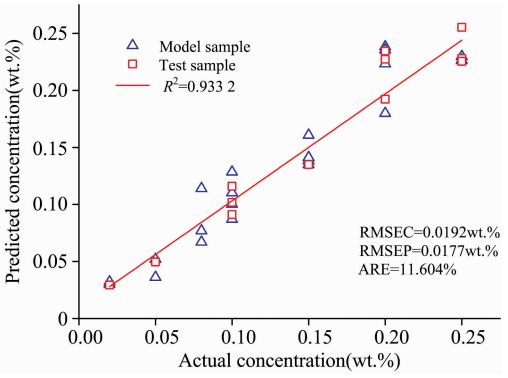

图 2 Ca Ⅱ 393.366 nm 的单变量定标曲线

Fig. 2 Single variable calibration curve of Ca Ⅱ 393.366 nm

2.2 偏最小二乘回归定标模型

为了改善单变量分析的不稳定性，提高分析精度，我们需要扩展表征待测元素浓度的光谱信息，充分利用光谱中的有用信息，实现多个变量之间信息互补，减小基体效应、波动等不确定因素的影响。同时，为了避免维度过高，模型过于复杂，我们需要对特征变量进行有效的筛选，因此引入了传统的变量选择方法—偏最小二乘回归(partial least squares regression，PLSR)。

在 PLSR 分析中，我们选取 392.818～397.61 nm 范围内的 100 条谱线作为多变量分析的输入，最终获得的主成分个数为 7。将得到的新的主成分作为变量与待测元素质量分数建立多变量关系。得到的定标模型如图 3 所示。横坐标为 Ca 元素的实际浓度，纵坐标为预测浓度，可以发现，曲线的拟合系数 R^2 达到了 0.975 3，RMSEC 和 RMSEP 分别为 0.010 8 Wt%和 0.013 Wt%，ARE 为 7.49%。

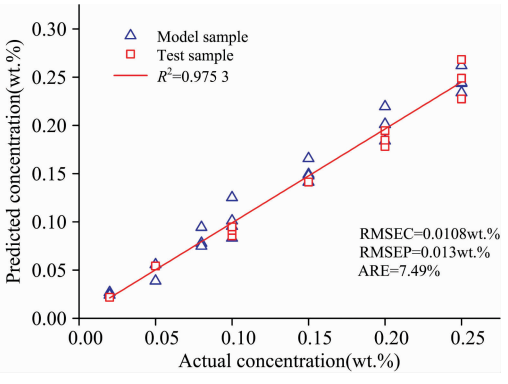

图 3 偏最小二乘回归定标曲线

Fig. 3 Partial least squares regression calibration curve

2.3 CART 回归树定标模型

CART 回归树利用最重要的特征信息构建树模型，可以分析所得变量的重要性程度，因此可以利用回归树对 LIBS 光谱进行特征变量的选取。本章节的数据处理在 Python 编程语言的框架内完成，利用机器学习模块 scikit-learn 进行特征变量选择，选择 392.818～397.61 nm 范围内的 100 条谱线作为回归树的输入，总样本的 70%作为训练集，30%作为预测集，检验模型的性能。

在 CART 回归树的构建过程中，波长变量数由 100 个减小到 6 个，变量压缩率达到了 94%。图 4 为回归树方法优选的 6 个波长变量的重要性分布情况。从图中可以看出优选的波长变量分别为 393.013 6，393.160 3，393.366，393.794 6，395.348 3 和 396.847 nm，得到的有效特征变量主要包括 Ca 的特征谱线(其中 Ca Ⅱ 393.366 nm 的重要性达到了 0.731 1)和相邻谱线。由此表明，Ca 元素含量不但与自身特征谱线强度有关，还受到了其他相邻特征谱线的干扰。

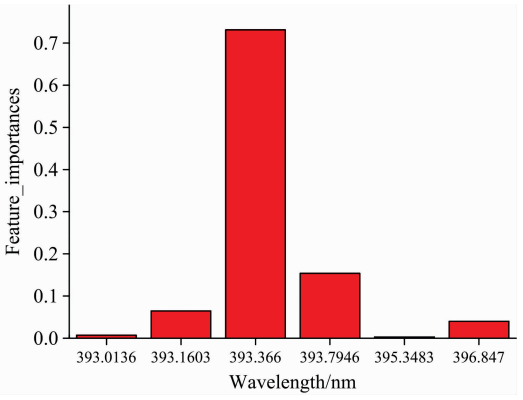

图 4 特征变量的重要性分布

Fig. 4 The importance distribution of feature variables

根据选择的最优变量组合，建立 Ca 元素的 CART 回归树定标模型。如图 5 所示，拟合系数 R^2 达到 0.997 5，RMSEC 和 RMSEP 分别达到 0.003 5 Wt%和 0.006 1 Wt%，ARE 降低到 2.500%。与单变量和 PLSR 相比，稳定性明显提高，模型的预测误差得到显著的降低，可见 CART 回归树可以用于 LIBS 特征变量的选择，具体对比结果如表 2 所示。

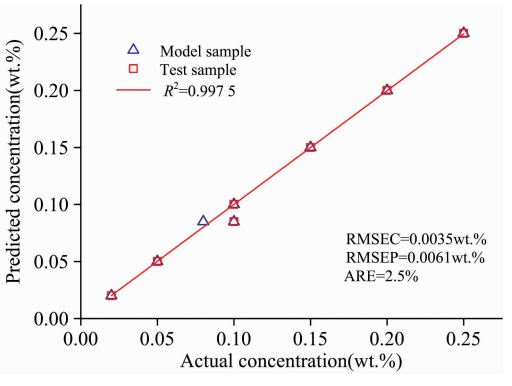

图 5 CART 回归树定标曲线

Fig. 5 CART regression tree calibration curve

表 2　定标模型参数比较

Table 2　Comparison of calibration model parameters

模型	R^2	RMSEC /Wt%	RMSEP /Wt%	ARE /%
单变量	0.933 2	0.019 2	0.017 7	11.604
PLSR	0.975 3	0.010 8	0.013 0	7.490
CART 回归树	0.997 5	0.003 5	0.006 1	2.500

3　结　论

研究了 CART 回归树对 LIBS 光谱中变量的筛选能力，通过构建 CART 回归树，以平方误差最小化为准则，从 100 个波长中获取到最重要的 6 个特征变量，变量压缩率达到了 94%，从而建立 Ca 含量的回归树定标模型。Ca 元素实际浓度与预测浓度的拟合系数达到 0.997 5，RMSEC，RMSEP 和 ARE 分别为 0.003 5 Wt%，0.006 1 Wt%和 2.500%，优于单变量和 PLSR 定标模型。由此表明，CART 回归树可以对变量进行有效的筛选，剔除无用信息，提高定量模型准确度和稳定性，因此，CART 回归树与 LIBS 结合可以作为一种快速、准确、鲁棒性强的检测方法。

References

[1] Guo L B, Zhu Z H, Li J M, et al. Optics Express, 2018, 26(3): 2634.

[2] YANG Ming-qing, WANG Chao, YAN Zhi-quan(杨明清，王　超，阎治全). China Petroleum Exploration(中国石油勘探), 2018, 23(1): 117.

[3] GUO Jin-jia, LU Yuan, et al(郭金家，卢　渊，等). Journal of Atmospheric and Environmental Optics(大气与环境光学学报), 2020, 15(1): 13.

[4] Shin Sungho, Moon Youngmin, Lee Jaepil. Plasma Science and Technology, 2019, 21(3): 34011.

[5] YANG You-sheng, ZHANG Yan, et al(杨友盛，张　岩，等). Laser & Optoelectronics Progress(激光与光电子学进展), 2015, 52(5): 053001.

[6] ZHANG Lei, WANG Zhe, DING Hong-bin, et al(张　磊，王　哲，丁洪斌，等). Journal of Atmospheric and Environmental Optics(大气与环境光学学报), 2016, 11(5): 338.

[7] XIN Yong, LI Yang, CAI Zhen-rong, et al(辛　勇，李　洋，蔡振荣，等). Metallurgical Analysis(冶金分析), 2019, 39(1): 15.

[8] Bilge G, Sezer B, Boyaci I H, et al. Spectrochimica Acta Part B: Atomic Spectroscopy, 2018, 145: 115.

[9] Jun H M, Kim J H, Lee S H, et al. Energy, 2018, 160: 225.

[10] ZHANG Le-hao, ZHANG Li, WU Zhong-chen, et al(张乐豪，张　立，武中臣，等). Acta Photonica Sinica(光子学报), 2020, (6): 0630002.

[11] WU Yi-qing, SUN Tong, LIU Jin, et al(吴宜青，孙　通，刘　津，等). Laser & Optoelectronics Progress(激光与光电子学进展), 2018, 55(1): 013005.

[12] HU Li, ZHAO Nan-jing, et al(胡　丽，赵南京，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(8): 2585.

[13] GUO Kai-chen, WU Zhong-chen, ZHU Xiang-ping, et al(郭恺琛，武中臣，朱香平，等). Acta Photonica Sinica(光子学报), 2019, 48(10): 1030002.

[14] Sun Chen, Tian Ye, Gao Liang, et al. Scientific Reports, 2019, 9(1): 11363.

[15] Ma F, et al. Journal of Analytical Atomic Spectrometry, 2020, 35: 478.

Research on Selection Method of LIBS Feature Variables Based on CART Regression Tree

YOU Wen[1], XIA Yang-peng[1], HUANG Yu-tao[1], LIN Jing-jun[2*], LIN Xiao-mei[3*]

1. Department of Electronics and Electrical Engineering, Changchun University of Technology, Changchun　130012, China

2. Department of Mechanical and Electrical Engineering, Changchun University of Technology, Changchun　130012, China

3. Jilin University of Architecture and Technology, Changchun　130012, China

Abstract　When laser induced breakdown spectroscopy (LIBS) is used for detection, due to the many and complex spectral lines, there are much redundant information, which will affect the quantitative analysis. Therefore, extracting effective feature variables is of great significance in the quantitative analysis of LIBS. In this paper, the method of selecting the spectral characteristics of the Ca element in the $CaCl_2$ solution was analyzed, and the accuracy and stability of the univariate model, partial least square regression and CART regression tree calibration model were compared. In view of the large volatility of the

surface of the water body, the poor spectral stability, and the fact that the spectrum is affected by the matrix effect and the self-absorption effect, the fitting coefficient (R^2) obtained by the univariate model is only 0.933 2, and the training root mean square error (RMSEC), prediction root mean square error (RMSEP) and average relative error (ARE) are 0.019 2 Wt%, 0.017 7 Wt% and 11.604% respectively. After partial least squares regression optimization, the model R^2 is increased to 0.975 3, and RMSEC, RMSEP and ARE are reduced to 0.010 8 Wt%, 0.013 Wt% and 7.49%, respectively. Although the model's accuracy has been improved, it is still difficult to meet the analysis requirements. In order to further improve the accuracy of quantitative analysis, a CART regression tree calibration model was established. When constructing the tree model, this method uses the square error minimization criterion to select the optimal combination of characteristic variables from the complex spectral information to make classification decisions, thereby establishing the calibration curve of the Ca element. Through the variable selection of the CART regression tree, the number of characteristic variables is reduced from 100 to 6, and the compression rate of variables reaches 94%, which significantly reduces the interference of irrelevant spectral lines. The correlation coefficients of the regression tree model are R^2, RMSEC, RMSEP and ARE is 0.997 5, 0.003 5 Wt%, 0.006 1 Wt% and 2.500%, respectively. Compared with the traditional univariate and partial least square regression, the CART regression tree model has higher accuracy and lower error. Through effective screening of characteristic variables, this paper eliminates the interference of irrelevant signals, significantly reduces the influence of matrix effect and self-absorption effect on LIBS quantitative analysis, and improves the accuracy and stability of quantitative analysis.

Keywords Laser-induced breakdown spectroscopy; Feature variable selection; CART regression tree; Quantitative analysis

(Received Oct. 10, 2020; accepted Feb. 4, 2021)

* Corresponding authors