DOI:10.16136/j.joel.2022.02.0302

一种自动选择特征的激光诱导击穿光谱定量分析方法

王 凯[1],史晋芳[1*],邱 荣[1],万 情[2],张志威[1],潘高威[1]

(1.西南科技大学 制造科学与工程学院,四川 绵阳 621010;2.陆军勤务学院 教研保障中心,重庆 401331)

摘要:本文针对激光诱导击穿光谱技术(laser-induced breakdown spectroscopy,LIBS)定量分析中的特征选择问题,提出一种基于Pearson相关系数的排序、主成分分析和L1正则项相结合的自动选择特征的定量分析方法,建立了土壤中Co元素的定量分析模型。该模型训练集和测试集的R^2(决定系数)分别为0.995和0.991,均方根误差(root mean square error,*RMSE*)分别为4.634 mg/kg和6.078 mg/kg,平均绝对误差(mean absolute error,*MAE*)分别为6.100%和6.441%,特征个数由原始数据的42870个降至5个,耗时仅0.97 s。结果表明:采用该方法可降低特征子集维度并提高模型的泛化性和精确度,为LIBS技术定量分析的特征选择提供一种高效的方法。

关键词:激光诱导击穿光谱;特征选择;Pearson相关系数;主成分分析;L1正则项
中图分类号:TN249 **文献标识码**:A **文章编号**:1005-0086(2022)02-0187-06

An automatic feature selection method for laser induced breakdown spectroscopy quantitative analysis

WANG Kai[1], SHI Jinfang[1*], QIU Rong[1], WAN Qing[2], ZHANG Zhiwei[1], PAN Gaowei[1]

(1. School of Manufacturing Science and Engineering, Southwest University of Science and Technology, Mianyang, Sichuan 621010, China; 2. Amy Logistics University of PLA, Chongqing 401331, China)

Abstract: Aiming at the problems of prior knowledge and slow algorithm convergence for feature selection in the quantitative analysis of laser-induced breakdown spectroscopy (LIBS) technology, this paper proposes a quantitative analysis method for automatically selecting features with a combination of Pearson correlation coefficient-based ranking, principal component analysis and L1 regular term. This method first selects the feature that has the greatest correlation with the target element, then compresses the feature dimension to within the number of samples, and finally sparses the feature weight coefficient and establishes a quantitative analysis model. This method is used to screen the characteristic subsets of Co elements in the soil and establish a quantitative analysis model. The R^2 (coefficient of determination) of the training set and test set of the model reached 0.995 and 0.991, root mean square error (*RMSE*) were 4.634 mg/kg and 6.078 mg/kg, mean absolute error (*MAE*) were 6.100% and 6.441%. The number of features is reduced from 42870 of the original spectral data to 5, which takes only 0.97 s. The results show that the method proposed in this paper can reduce the dimension of feature subsets and improve the generalization and accuracy of quantitative analysis models, providing an efficient method for feature selection in quantitative analysis of LIBS technology.

Key words: laser-induced breakdown spectroscopy; feature selection; Pearson correlation coefficient; principal component analysis; L1 regularization

* **E-mail**:603071939@qq.com
收稿日期:2021-05-06 **修订日期**:2021-06-01
基金项目:国家自然科学基金(11972313,11902277)资助项目

1 引 言

激光诱导击穿光谱技术(laser-induced breakdown spectroscopy,LIBS)是一种新兴的原子发射光谱技术,因其具有无需样本预处理、可对各种物态分析、多元素同时检测、可实时在线检测等优点,被广泛地应用于化工、食品、生物等多个领域。随着核极限学习机(kernel-based extreme learning machine,KELM)[1]、人工神经网络(artificial neural network,ANN)[2]、支持向量机[3]、随机森林[4]等多元分析方法的引入,LIBS技术的应用范围得以不断拓宽。由于检测样本和实验环境的复杂性,LIBS原始数据包含了噪声、无关特征和冗余特征[5]。若以原始数据作为多元分析方法的输入,易产生"维数灾难"现象,因此,对于LIBS数据,选择合适的特征子集用于提升定量分析模型性能[6]成为近年来的研究热点。

综合国内外研究,针对LIBS数据特征子集的选择,主要有两类方法[7,8]:第一类为基于先验知识的手动筛选方法,例如REMUS等[9]利用偏最小二乘-判别分析(partial least squares discriminant analysis,PLS-DA)模型对黑曜石进行分类,发现手动剔除无关变量可明显提升分类正确率;第二类为自动选择特征子集方法,例如YAN等[10]将V-WSP(variable-reduction Wootton, Sergent, Phan-Tan-Luu's)方法和粒子群优化(particle swarm optimization,PSO)方法相结合对数据进行特征选择,不仅剔除了冗余特征,还降低了特征选择过程所需时间,但V-WSP方法需大量试验才能得到较好的阈值;马双双等[11]利用遗传算法对畜禽粪便中钙含量进行研究,发现遗传算法可减少数据维度,提高建模效率,但同时也发现遗传算法的收敛速度较慢且结果具有不确定性。

针对上述特征选择中存在算法收敛慢及算法阈值难确定的问题,本文提出一种基于Pearson相关系数的排序、主成分分析和L1正则项相结合的自动选择特征的LIBS定量分析方法,通过对土壤中Co元素检测,验证了该方法的有效性。

2 实验部分

2.1 样品制备

分别称取17 g标准土壤(GSS-32,泛滥平原沉积物成分分析标准物质)、2.924 g PE微粉和0.076 g一氧化钴样品,使用研钵研磨均匀并烘干,得到Co的质量浓度为3 000 mg/kg的粉末样品,编号C0。取1 g C0,再按质量比3∶1分别取土壤和PE微粉共11.5 g,将三者混合研磨均匀,得到Co的质量浓度为240 mg/kg,编号C1。按照此方法,依次制样,编号C2—C8,土壤样品中Co元素浓度如表1所示。

表1 土壤样品中Co元素浓度(mg/kg)

Tab. 1 Co concentration in soil samples (mg/kg)

	Co concentration							
Sample No.	C1	C2	C3	C4	C5	C6	C7	C8
Concentration	240	150	130	74	53	45	31	23

2.2 实验设备

图1所示为激光诱导击穿光谱技术实验系统示意图,激光脉冲通过透镜汇聚于放置在三维平移台上的样品表面,激光焦点在样品表面以下2 mm位置,激光烧蚀样品产生等离子体,光谱信号由光纤探头收集,经光纤传递至光谱仪。其中激光器与光谱仪之间的时序关系由数字脉冲发生器(DG645)进行控制。LIBS数据采集过程中,样品表面每个烧蚀区只辐照一次。

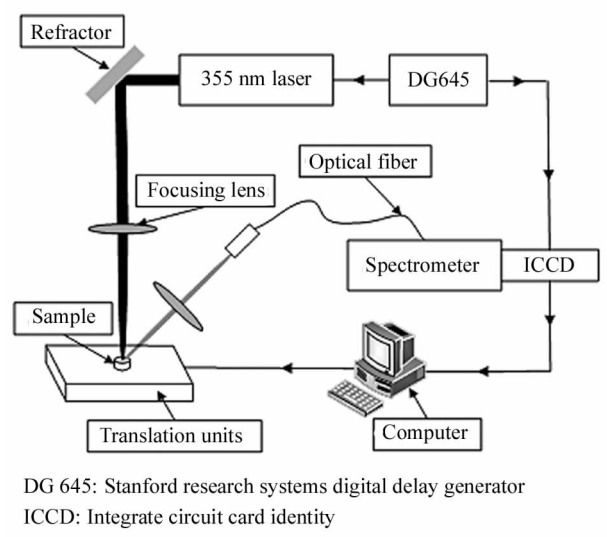

图1 LIBS实验系统示意图

Fig. 1 Detection system of soil heavy metals by LIBS

对s个不同浓度的样品,每个浓度采集p幅光谱,得到$s\times p=n$幅光谱,每幅光谱包含m组数据(波长、强度),m为探测器总像素(实验所用ICCD像素值为42 870),可组成光谱强度数据矩阵$\boldsymbol{X}[n,m]$,其对应浓度矩阵$\boldsymbol{y}[n,1]$。为避免偶然因素,减小数值波动和降低噪声,每个样品采集6幅有效光谱,每幅光谱为激光激发10次取平均。故本实验数据所组成的光谱强度矩阵为$\boldsymbol{X}[48,42\,870]$,浓度矩阵为$\boldsymbol{y}[48,1]$。取C2和C8为测试集,其余为训练集,故训练数据集矩阵为$\boldsymbol{X}_{TR}[36,42\,870]$,测试数据集矩阵为$\boldsymbol{X}_{TE}[12,42\,870]$。

3　特征子集选择方法及流程

3.1　特征子集选择方法

特征子集选择方法分为基于 Pearson 相关系数排序、主成分分析和 L1 正则项等 3 个阶段。

3.1.1　基于 Pearson 相关系数的排序

在对目标元素进行定量分析时,元素的谱线强度 I 与浓度 c 的关系式称为罗马金-赛伯公式,一定条件下,可表示为:

$$I = A \times c^b, \tag{1}$$

式中,A 和 b 是常数。

常数 b 称为自吸收系数,它的数值与谱线的自吸收有关,当不考虑谱线自吸收时,目标元素的谱线强度与其浓度呈线性关系。

Pearson 相关系数通常被用于研究两组数据之间的线性关系,其数学式如下:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{2}$$

式中,cov()表示协方差,σ 表示标准差。

基于Pearson相关系数的排序方法步骤为:1) 计算原光谱数据集 $\boldsymbol{X}[n,m]$中的每一个特征 i 和对应浓度 $\boldsymbol{y}[n,1]$之间的样本相关系数 $\rho^i_{X,Y}$; 2) 计算每个特征 i 与浓度 y 之间的统计量 f 得分,f 值越大,表示第 i 个特征与浓度 y 之间的相关性越大; 3) 选取与目标元素相关性最大的前 r 个特征,即得分排名前 r 的特征集合,得到特征子集 $\boldsymbol{A}_1[n,r]$。如式(3)、(4)表示为:

$$\rho^i_{X,Y} = \frac{(X[:,i] - mean(X[:,i]))(y - mean(y))}{\sigma_{X[:,i]}\sigma_Y}, \tag{3}$$

$$f = \frac{\rho^i_{X,Y}{}^2}{1 - \rho^i_{X,Y^2}} \times (n-2), \tag{4}$$

式中,$\rho^i_{X,Y}$表示第 i 个特征和浓度 y 之间的样本相关系数,$X[:,i]$表示所有光谱中的 i 号特征,$mean(X[:,i])$表示所有光谱中 i 号特征的平均值,$mean(y)$表示y的平均值,$(n-2)$表示f服从$F(1,n-2)$分布。

3.1.2　主成分分析(principal component analysis, PCA)

PCA[12] 的主要思想是将 r 维特征映射到 k 维上,即通过计算数据的协方差矩阵,得到其特征值和特征向量,选择特征值最大的前 k 个特征所对应的特征向量,组成特征矩阵。

对排序得到的特征子集 $\boldsymbol{A}_1$ 使用 PCA 方法,得到 $\boldsymbol{A}_2$,其步骤为:1) 对 $\boldsymbol{A}_1$ 中所有元素进行中心化; 2) 计算 $\boldsymbol{A}_1$ 的协方差矩阵; 3) 对 $\boldsymbol{A}_1$ 的协方差矩阵做特征值分解; 4) 取累积方差为 99.9%的前 k 个特征值所对应的特征向量,得到特征子集 $\boldsymbol{A}_2[n,k]$。

3.1.3　L1 正则项

L1 正则项,又被称作“稀疏规则算子”(Lasso regularization),可实现特征的自动选择[13]。一般来说,在最小化目标函数时,考虑更多的特征虽可获得更小的训练误差,但在预测新的样本时,这种做法会干扰对正确目标的预测,L1 正则项的引入就是为了去掉没有贡献或贡献很小的特征,将这些特征对应的权重置零。其数学表达式为:

$$\tilde{J}(w;x,y) = J(w;x,y) + \lambda \| w \|_1, \tag{5}$$

式中,x、y 分别为训练样本和相应标签,w 为权重系数向量,$J()$为目标函数,$\| w \|_1$即为正则项,参数 λ 控制正则化强弱,其中 $\lambda>0$。

利用基于 L1 正则项对特征子集 $\boldsymbol{A}_2$ 进行稀疏化操作,得到最终的特征子集 $\boldsymbol{A}_3$。

3.2　特征子集选择流程

特征子集选择流程为:首先,使用排序方法选择原数据集 $\boldsymbol{X}[n,m]$中与目标元素相关性最大的前 r 个特征,去除数据中的无关信息,得到特征子集 $\boldsymbol{A}_1[n,r]$;然后,使用主成分分析方法对 $\boldsymbol{A}_1[n,r]$进行主成分选择,选择 $\boldsymbol{A}_1[n,r]$中累积方差为 99.9%的前 k 个主成分,得到特征子集 $\boldsymbol{A}_2[n,k]$;最后,利用 L1 正则项对 $\boldsymbol{A}_2[n,k]$中各个特征的权重系数进行重分布,去除权重系数为零的特征,得到包含 t 个特征的特征子集 $\boldsymbol{A}_3[n,t]$ 。

特征子集选择流程如图 2 所示,因实验环境及土壤样品基体效应[14] 的影响,原始光谱数据包含大量噪声且存在基线漂移现象,需对原始光谱数据降噪、去基线。本文使用离散小波变换对原始光谱进行降噪、去基线操作,得到样本矩阵 $\boldsymbol{A}_{\mathrm{total}}[48,42870]$,再按上述特征子集选择流程分别得到 $\boldsymbol{A}_1$、$\boldsymbol{A}_2$ 和 $\boldsymbol{A}_3$,因训练集和测试集的处理过程相同,后续以训练集为例进行特征子集的选择。

第一步,对训练集样本矩阵 $\boldsymbol{A}_{TR}[36,42870]$进行排序,通过排序步骤去除数据中的无关变量,选择 169 个(此阶段保留的特征数目参考 NIST 数据库)与 Co 元素相关性最大的特征,得到样本矩阵 $\boldsymbol{A}_1[36,169]$,此处以其中一幅光谱数据为例,排序选择的 Co 的特征子集如图 3 所示。

第二步,通过主成分分析对 $\boldsymbol{A}_1$ 进行降维,得到累积方差为 99.9%的特征子集 $\boldsymbol{A}_2[36,27]$。Co 元素特征子集各主成分方差百分比如图4所示。从图4

可以看出，第一主成分的方差百分比为95%，在数值上明显大于其他主成分，但经实验验证，当只取图示第一主成分作为特征子集对Co进行定量分析时，训练集 R^2 仅为0.96，这是由于主成分个数较少，所包含数据信息不足，造成算法模型因训练样本信息量过少出现学习不足的现象，进而导致模型性能下降。所以为得到更好的定量分析结果，此处选择累计方差为99.9%的主成分。

第三步，利用最小绝对收缩选择算子（Lasso）对 $\boldsymbol{A}_2$ 进行稀疏化处理并建立定量分析模型，Co元素特征子集L1正则项权重系数分布如图5所示，共选出5个特征，其余特征系数全部被置零，得到特征子集 $\boldsymbol{A}_3$[36,5]。

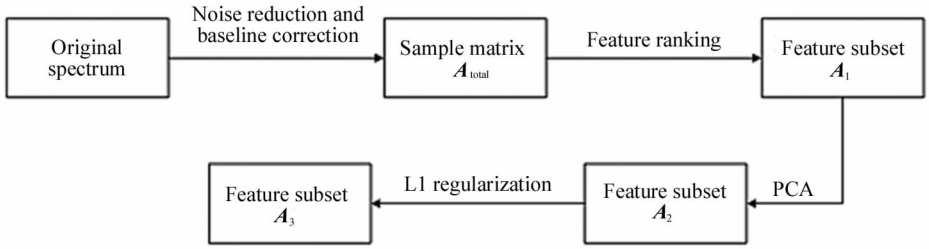

图2　特征子集选择流程

Fig. 2　Feature subset selection process

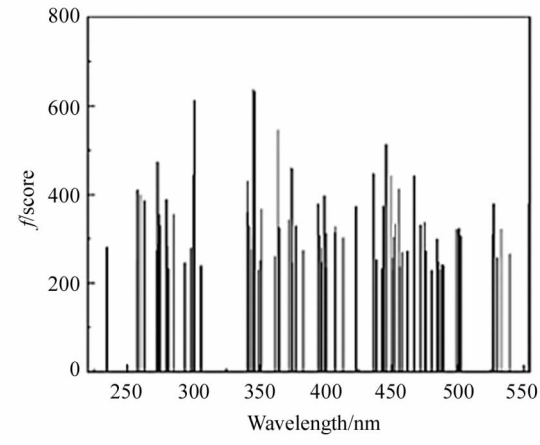

图3　排序选择的Co的特征子集

Fig. 3　The feature subset A1 of Co selected by the sorting method

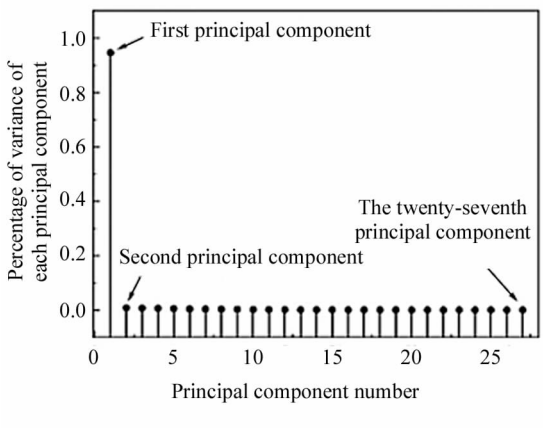

图4　Co元素特征子集各主成分方差百分比

Fig. 4　The percentage of principal component variance of the characteristic subset of Co element

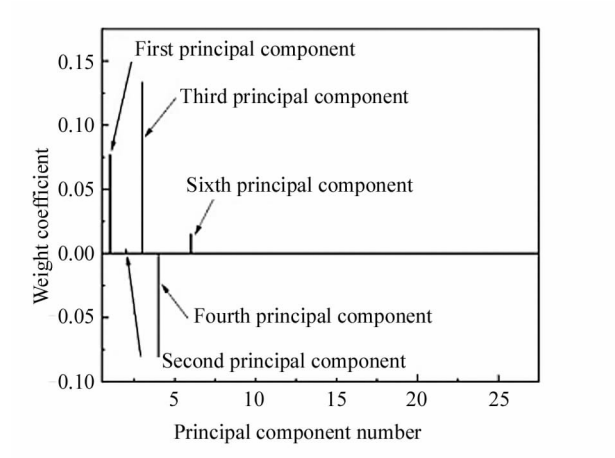

图5　Co元素特征子集L1正则项权重系数分布

Fig. 5　The regular term weight coefficient distribution of characteristic subset L1 of Co element

4　定量分析

选择① Lasso（最小绝对收缩选择算子）、② PCA+Lasso（主成分分析+最小绝对收缩选择算子）、③ Sort+Lasso（排序+最小绝对收缩选择算子）、④ 支持向量机回归（support vector regression，SVR）与本方法建立的Co定量模型进行对比以评估本方法所建模型的性能。其中，选择①、②、③是为验证本方法相比单独使用其中一阶段或使用两阶段方法的有效性，选择④的原因为：在处理特征尺寸大于样本数量的数据时，SVR使用广泛且较为有效[15]，Co的定量模型结果对比如表2所示。②、③、④3种模型的正确率均不及本方法，且在测试集的差距更显著，①的训练集正确率虽大于本方法，但其出现测试集正确率远不及训练集的现象，这是由于

Lasso 对高维小样本数据集的特征选择容易出现过拟合问题[16]。②在①的基础上先对数据进行排序处理,选出与目标元素相关性最大的特征,故模型表现有所提升。③在②之前先对数据进行降维(PCA)处理,将特征维数降至样本数以内,再使用 Lasso 进行后续操作,所以模型正确率有明显提升,但③方法并没有去除数据集中的无关特征,不能正确表示目标元素的浓度变化。⑤(本方法)结合②、③之所长,不仅剔除数据中的无关特征,而且将特征维数降至样本数以内,便于 Lasso 稀疏化特征权重系数并建立定量模型,Co 的定标曲线如图 6 所示。此外,对比 SVR,本方法依旧有明显的优势。

表 2　Co 的定量模型结果对比

Tab. 2　Comparison of Co quantitative model results

No.	Models	Calibration set			Validation set			Time/s
		R^2	*RMSE*/(mg/kg)	*MAE*/%	R^2	*RMSE*/(mg/kg)	*MAE*/%	
①	Lasso	0.999	0.268	0.353	0.635	38.367	59.251	1.49
②	Sort+Lasso	0.903	19.782	20.061	0.901	20.015	31.011	0.91
③	PCA+Lasso	0.979	9.065	11.053	0.968	11.300	13.646	0.18
④	SVR	0.991	5.989	9.314	0.961	12.541	17.040	1.35
⑤	Sort+PCA+Lasso	0.995	4.634	6.100	0.991	6.078	6.441	0.97

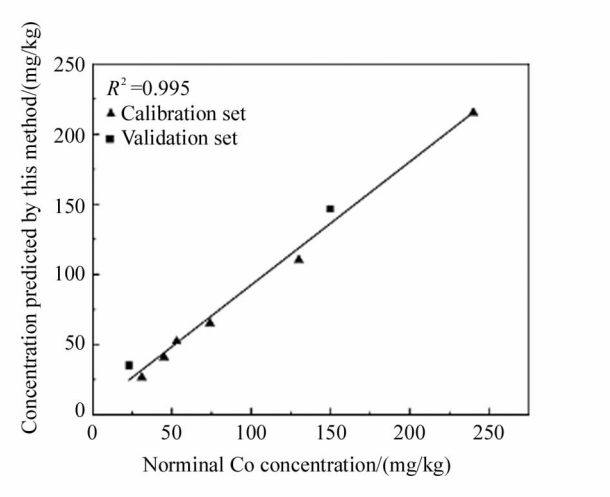

图 6　Co 的定标曲线

Fig. 6　Calibration curve of Co

5　结　论

本文将基于 Pearson 相关系数的排序、主成分分析和 L1 正则项三者相结合的自动选择特征方法应用于 LIBS 的定量分析中,对土壤中 Co 元素进行检测,将该方法与其他 4 种方法的检测结果进行比较,相比其他方法,本方法建立的定量模型准确度和泛化性均有明显提升,其训练集和测试集的 R^2(决定系数)分别为 0.995 和 0.991,均方根误差(root mean square error,*RMSE*)分别为 4.634 mg/kg 和 6.078 mg/kg,平均绝对误差(mean absolute error,*MAE*)分别为 6.100% 和 6.441%,所使用特征个数为 5 个,耗时为 0.97 s。这是由于排序选择出与 Co 元素相关性最大的特征变量,主成分分析将特征维数降至样本数以内,L1 正则项稀疏化了特征权重系数。由此得出:该方法在未增加时间开销的基础上,不仅降低了特征子集的维度,而且提升了定量分析模型的性能,这为 LIBS 数据定量分析领域提供了一种高效便捷的方法。

参考文献:

[1] YAN C,ZHANG T,SUN Y,et al. A hybrid variable selection method based on wavelet transform and mean impact value for calorific value determination of coal using laser-induced breakdown spectroscopy and kernel extreme learning machine[J]. Spectrochimica Acta Part B Atomic Spectroscopy,2019,154:75-81.

[2] CAMPANELLA B,GRIFONI E,LEGNAIOLI S,et al. Classification of wrought aluminum alloys by ANN evaluation of LIBS spectra from aluminum scrap samples[J]. Spectrochimica Acta Part B Atomic Spectroscopy,2017,134:52-57.

[3] BOUCHER T F,OZANNE M V,CARMOSINO M L,et al. A study of machine learning regression methods for major elemental analysis of rocks using laser-induced breakdown spectroscopy[J]. Spectrochimica Acta Part B Atomic Spectroscopy,2015,107:1-10.

[4] QI J,ZHANG T,TANG H,et al. Rapid classification of archaeological ceramics via laser-induced breakdown spectroscopy coupled with random forest[J]. Spectrochimica Acta Part B Atomic Spectroscopy,2018,149:288-293.

[5] LI Z Q,DU J Q,NIE B,et al. Review of feature selection

methods[J]. Computer Engineering and Applications, 2019,55(24):10-19.
李郅琴,杜建强,聂斌,等.特征选择方法综述[J].计算机工程与应用,2019,55(24):10-19.

[6] KUMAR M A, SPEGAZZINI N, ZHANG C, et al. Less is more:avoiding the LIBS dimensionality curse through judicious feature selection for explosive detection[J]. Scientific Reports,2015,5:1-10.

[7] MEHMOOD T,LILAND K H,SNIPEN L,et al. A review of variable selection methods in partial least squares regression[J]. Chemometrics & Intelligent Laboratory Systems,2012,118:62-69.

[8] GUEZENOC J,BASSEL L,GALLET-BUDYNEK A,et al. Variables selection: a critical issue for quantitative laser-induced breakdown spectroscopy[J]. Spectrochimica Acta Part B Atomic Spectroscopy,2017,134:6-10.

[9] REMUS J J,GOTTFRIED J L,HARMON R S,et al. Archaeological applications of laser-induced breakdown spectroscopy:an example from the Coso Volcanic Field, California,using advanced statistical signal processing analysis[J]. Applied Optics,2010,49(13):120-131.

[10] YAN C H,LIANG J,ZHAO M J,et al. A novel hybrid feature selection strategy in quantitative analysis of laser-induced breakdown spectroscopy[J]. Analytica Chimica Acta,2019,1080:35-42.

[11] MA S S,MA Q L,HAN L J,et al. Study on calcium content in livestock manure based on LIBS and genetic algorithm [J]. Spectroscopy and Spectral Analysis,2017,37(5):1530-1534.
马双双,马秋林,韩鲁佳,等.基于LIBS和遗传算法的畜禽粪便中钙含量研究[J].光谱学与光谱分析,2017,37(5):1530-1534.

[12] ALAA T. Principal component analysis-a tutorial[J]. Int. J. of Applied Pattern Recognition,2016,3(3):197-240.

[13] LENG W,LI J P,ZHANG C Q. LASSO variable selection of high-dimensional mixture model[J]. Mathematical Statistics and Management,2019,38(1):85-90.
冷薇,李俊鹏,张崇岐.高维混料模型的LASSO变量选择[J].数理统计与管理,2019,38(1):85-90.

[14] LI H L,XIE H J,LU H S,et al. Research on metal matrix-assisted measurement based on laser-inducedbreakdown spectroscopy technology[J]. Journal of Optoelectronics · Laser,2021,32(2):166-172.
李红莲,谢红杰,吕贺帅,等.基于激光诱导击穿光谱技术的金属基体辅助测量研究[J].光电子 · 激光,2021,32(2):166-172.

[15] SHI W F,HU X Q,YU K. An iterative lasso feature selection method for high-dimensional data[J]. ApplicationResearch of Computers,2011,28(12):4463-4466.
施万锋,胡学钢,俞奎.一种面向高维数据的迭代式Lasso特征选择方法[J].计算机应用研究,2011,28(12):4463-4466.

[16] HADDAD J E,CANIONI L,BOUSQUET B. Good practices in LIBS analysis:review and advices[J]. Spectrochimica Acta Part B:Atomic Spectroscopy,2014,101:171-182.

作者简介:

史晋芳 (1977—),女,副教授,硕士生导师,主要从事智能化测控与图像处理技术的研究.