DOI: 10.11895/j.issn.0253-3820.140668

基于遗传算法和偏最小二乘法的土壤激光诱导击穿光谱定量分析研究

邹孝恒 郝中骐 易荣兴 郭连波 沈 萌 李祥友*
王泽敏 曾晓雁 陆永枫
(华中科技大学武汉光电国家实验室(筹) , 激光与太赫兹功能实验室,武汉 430074)

摘 要 在空气环境下,采用激光诱导击穿光谱(LIBS) 技术对土壤成分进行检测,建立了基于遗传算法(GA) 和偏最小二乘法(PLS) 的定量分析模型。将配制的 58 个土壤样品分为定标集、监控集和预测集,对 11 种组分 Mn, Cr, Cu, Pb, Ba, Al_2O_3, CaO, Fe_2O_3, MgO, Na_2O 和 K_2O 的含量分别进行预测。结果表明, GA 作为一种谱线选择的预处理方法,可以有效减少用于 PLS 建模的光谱谱线的数目,从而简化模型。对于土壤中的大部分组成成分,GA-PLS 模型能够显著改善传统 PLS 模型的预测能力。以 Mn 元素为例,浓度预测均方根误差(RMSEP) 从 0.0215% 降低至 0.0167%,平均百分比误差(MPE) 从 8.10% 降低至 5.20%。本研究为进一步提高土壤的 LIBS 定量分析准确度提供了方法参考。

关键词 激光诱导击穿光谱; 遗传算法; 偏最小二乘法; 土壤

1 引 言

激光诱导击穿光谱(LIBS) 技术是一种基于激光与物质相互作用的等离子体光谱物质组分分析技术,它具有检测速度快、无需样品预处理、可进行实时在线及远程测量等优点[1],在地质分析、生物医学、环境监测和文物鉴定等领域已得到广泛研究和应用[2,3]。LIBS 技术通过获取等离子体发射谱线的强度信息确定元素的含量。传统的定量分析方法包括基本定标法和内定标法[4],这些单变量模型仅利用待测元素的单条特征谱线进行定量分析,无法消除谱线干扰、自吸收效应和基体效应等因素的影响,因而准确度不能满足实际应用的需要。

偏最小二乘法(PLS) 是一种多元统计数据分析方法,已用于铝合金[5]、钢铁[6]、煤炭[7] 和植物[8] 等材料的 LIBS 成分检测。PLS 利用一段光谱或全谱的所有数据点作为自变量,分析元素浓度作为因变量,拟合光谱强度和元素浓度之间的关系,在一定程度上可以克服传统单变量模型的缺陷。但由于这些变量通常会包括一些与浓度相关性较弱,甚至完全不相关的信号,因此会降低 PLS 模型定量分析的真实性和准确度。通过对 LIBS 的原始光谱数据进行预处理,去除与分析元素浓度无关的噪声信号,保留相关信息,可以提高 PLS 的预测精度。遗传算法(GA) 是一种有效的全局优化算法[9],它借鉴生物界的自然选择和遗传机制,利用选择、交叉和变异等操作,以给定的优化标准对光谱谱线进行筛选。GA 与 PLS 结合(GA-PLS) 用于物质成分分析的优越性已经在近红外光谱中得到证实[10~13]。然而,在 LIBS 领域应用较少。Fink 等[14] 将 GA-PLS 应用于 LIBS,对热塑料中的 Ti 元素进行定量分析,与传统的 PLS 方法相比,光谱输入变量数目从 87 减小到 22,定标均方根误差(RMSEC) 和交互检验均方根误差均显著下降。Gottfried 等[15] 将 GA-PLS 与 LIBS 结合,并应用于碳酸盐、氟石和硅酸盐地质材料的分类中,用 GA 提取部分数据点可以补偿双脉冲光谱的分辨率。Anderson 等[16] 将 GA-PLS 应用于优化“好奇”号火星探测器上 ChemCam 仪器对岩石的远程定量分析,该方法能够减少数据量和节省计算时间,对于太空探测和快速分析具有重要意义。

目前,土壤元素的 LIBS 定量分析方法主要有基本定标法[17]、内定标法[18]、神经网络法[19] 和 PLS 法[20] 等。本研究将 GA-PLS 应用于土壤的 LIBS 分析,对重金属元素 Mn, Cr, Cu, Pb, Ba 和氧化物

2014-08-06 收稿; 2014-10-29 接受
本文系国家重大科学仪器设备开发专项(No. 2011YQ160017),中国博士后科学基金(No. 2013M542014) 和中央高校基本科研业务费(Nos. CXY13Q021, CXY13Q022) 资助
* E-mail: xyli@mail.hust.edu.cn

Al_2O_3，CaO，Fe_2O_3，MgO，Na_2O 和 K_2O 进行定量分析。通过与传统 PLS 方法的结果比较，证实了 GA-PLS 方法对土壤成分定量分析具有更高的准确度。

2 实验部分

2.1 实验装置

实验装置如图 1 所示。实验在空气环境下进行。采用调 Q 开关 Nd: YAG 脉冲激光器(Beamtech，Nimma 400，波长 532 nm，重复频率 3 Hz，脉冲宽度 8 ns)，激光能量为 40 mJ/脉冲。激光经过反射镜和平凸透镜(焦距 10 cm) 聚焦到样品表面，样品放置在二维电动平台上作“弓”字形运动。为了防止击穿空气，焦点位于样品表面下方 2 mm 处。激发出的等离子体辐射光由光收集器收集并耦合到光纤中(光纤收集信号的角度为 15°)，传输至光谱仪(Andor Technology，Mechelle 5000，波长范围 200 ~ 950 nm，分辨率 $\lambda/\Delta\lambda$ = 5000) 进行分光，光谱仪配备增强型电荷耦合器件(ICCD) (Andor Technology，iStar DH-334T 1024 × 1024 像素) 用于光电转换。最后将数据传输至计算机进行分析和处理。

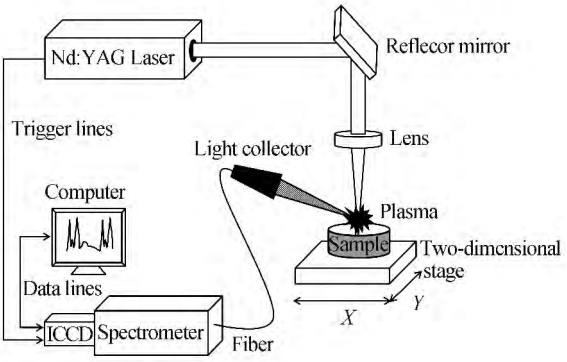

图 1 LIBS 实验装置原理图

Fig. 1 Schematic diagram of the experimental setup

为了获得较高的光谱强度和信背比，ICCD 门控采用时间积累模式，采集延迟时间设为 4 μs，门宽设为 10 μs。为了降低激光脉冲能量波动对光谱强度的影响，每张光谱在样品不同位置积累 60 个脉冲，每个样品重复采集 6 次。

2.2 样品制备

由于 PLS 模型应用于大样本数据中能获得更高的准确性，因此需要大量不同浓度的样品用于定量分析。选用 4 种国家标准土壤样品编号分别为 GBW(E) 070008，GBW(E) 070009，GBW07404 和 GBW07405，土壤中重金属元素 Mn，Cr，Cu，Pb，Ba 和氧化物 Al_2O_3，CaO，Fe_2O_3，MgO，Na_2O，K_2O，SiO_2 的含量如表 1 所示。每 2 种标准样品按配比分别为 1∶ 9，2∶ 8，……，9∶ 1 配制 9 种不同浓度的样品，经称量、搅拌、研磨使样品混合均匀，共配制得到 54 个样品。采用硼酸镶边垫底法对样品压片，压强 20 MPa，保压 3 min，制成直径4 cm，厚度 3 mm 的薄片。对于每种分析元素，各自将 58 个样品(4 个标准样品和 54 个自制样品) 按浓度从低到高编号，从中等间隔取出第 1，3，……，57 号样品(共 29 个) 作为定标集，第 2，6，……，58 号样品(共 15 个) 作为监控集，第 4，8，……，56 号样品(共 14 个) 作为预测集。由于每个样品重复采集 6 次，定标集、监控集和预测集分别包含 174，90 和 84 幅光谱。GA-PLS 程序在 Matlab R2010b 环境下编写。

表 1 土壤标准样品的主要成分和含量

Table 1 Major compositions and concentrations of standard soil samples

成分 Composition	标准值 Standard value (%)			
	GBW(E) 070008	GBW(E) 070009	GBW 07404	GBW 07405
Mn	0.210	0.709	0.142	0.136
Cr	0.0084	0.0099	0.0370	0.0118
Cu	0.0290	0.0854	0.0040	0.0144
Pb	0.0675	0.1141	0.0058	0.0552
Ba	0.0505	0.0522	0.0213	0.0296
Al_2O_3	14.32	18.20	23.45	21.58
CaO	3.58	1.14	0.26	0.10
Fe_2O_3	6.95	10.05	10.30	12.62
MgO	2.32	1.23	0.49	0.61
Na_2O	1.48	0.50	0.11	0.12
K_2O	2.67	2.60	1.03	1.50
SiO_2	60.57	54.31	50.95	52.57

3 结果与讨论

3.1 光谱数据的选取

土壤样品的 LIBS 光谱很复杂。将美国国家标准与技术研究院(National Institute of Standards and

Technology, NIST) 的标准原子光谱数据库[21] 与实际样品中分析元素的特征谱线逐一比对,得到了土壤中所含分析元素对应的部分特征谱线波长,以及用于分析和优化的光谱区间,如表 2 所示。这里的特征谱线是指谱线强度较高且能够分辨的谱峰。对于 PLS 模型,选取包含分析元素多条特征谱线的光谱区间(5 ~ 45 nm) 内的所有波长离散点强度值作为光谱数据; 而对于 GA-PLS 模型,通过 GA 从给定的光谱区间内挑选出部分强度值作为光谱数据。为了降低实验参数波动的影响,提高定量分析的精度,在建模之前需要对光谱做归一化处理。选择土壤中的主要元素 Si 的 288.16 nm 谱线作为内标线,所有谱线的强度均除以内标线强度。

表 2　土壤中所含元素、对应的特征谱线波长和光谱区间

Table 2　Compositions of soils, corresponding characteristic wavelengths and spectral regions

成分 Composition	特征谱线波长 Characteristic wavelengths (nm)							光谱区间 Spectral region (nm)
Mn	380.67	403.08	403.31	403.45	403.57	404.14	404.88	380 ~ 405
Cr	520.45	520.60	520.84					515 ~ 525
Cu	510.55	515.32	521.82					510 ~ 525
Pb	405.78							400 ~ 410
Ba	455.40							450 ~ 460
Al	308.22	309.27	309.27					305 ~ 315
Ca	317.93	318.13						310 ~ 320
Fe	404.58	406.36	427.18	432.58	438.35	441.51		400 ~ 445
Mg	517.27	518.36						515 ~ 520
Na	589.00	589.59						580 ~ 590
K	766.49	769.90						760 ~ 780

3.2　PLS 模型

将分析元素对应的光谱区间内的所有强度值作为 PLS 的自变量,分析元素浓度作为因变量,采用定标集建立定标模型。在 PLS 建模过程中,主成分个数的确定至关重要,关系到模型的稳定性和准确性。为了考察模型的预测能力,将 PLS 模型应用于监控集,根据监控集样品预测浓度的残差平方和(PRESS) 确定主成分个数。以 Mn 元素为例,PRESS 随主成分个数(取前 30 个主成分) 的变化情况如图 2 所示。PRESS 最小时对应的最佳主成分个数是 6,因此取 6 个主成分对 Mn 元素的定标集建立 PLS 模型。

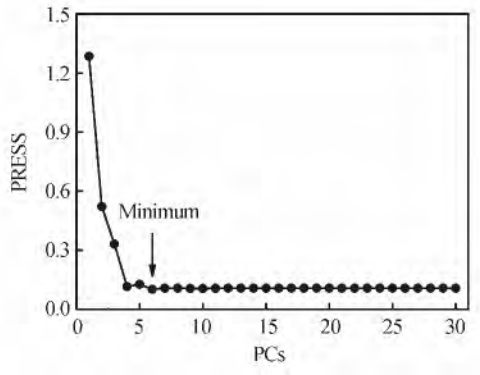

图 2　Mn 元素的残差平方和(PRESS) 随主成分个数的变化趋势

Fig. 2　Variation tendency of predicted residual error sum of squares (PRESS) with different numbers of principle components (PCs) for Mn

3.3　GA 优化谱线选择

GA 的主要影响因素有适应度函数(目标函数)、种群数目、终止代数、选择、交叉和变异算子等。遗传算法进行谱线选择的主要步骤为:

(1) 编码　将光谱区间中的每一个波长作为染色体的一个基因,采用 0-1 编码。若编码为 "1",则该波长处的强度值参与 PLS 建模; 若编码为 "0",则该波长处的强度值不参与建模。

(2) 种群初始化　设定种群数目为 200,采用随机初始化的方式,随机产生 200 条二进制编码的染色体。

(3) 评价　首先对染色体进行解码,将染色体中所有编码为 "1" 的基因对应的波长处的强度值代入 PLS 进行建模,PLS 中主成分个数由 3.2 节所述方法确定。然后计算该条染色体的适应度函数,设定适应度函数为监控均方根误差(RMSEM)。RMSEM 表示监控集样品浓度预测的均方根误差(RMSE),此函数也是 GA 待优化的目标函数。RMSEM 越小,暗示模型的预测准确性越高。

(4) 选择、交叉和变异操作　采用轮盘赌选择和单点交叉方法。交叉概率和变异概率分别设为 0.5

和0.1。

(5) 最优谱线组合　重复步骤(3)和(4)至最大遗传代数,设定终止代数为300。以目标函数RMSEM最小时的基因编码解码后得到的谱线组合作为最优谱线组合。

以Mn元素为例,每一代中所有种群的RMSEM的最小值随遗传代数的变化情况如图3所示。RMSEM的最小值随遗传代数的增大逐渐收敛,趋近于0。因为GA随机初始化种群,所以每次运行GA-PLS程序产生的结果会略有差异。在进行谱线选择时运行5次GA-PLS程序,产生5组待选谱线组合,同样以RMSEM为评价标准从中选出最优谱线组合。选取GA优化后的谱线用PLS对定标集建模,然后对监控集和测试集样品浓度进行预测。

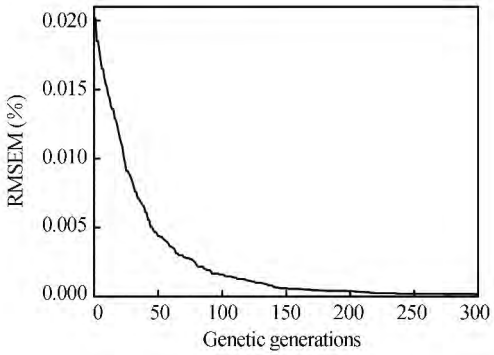

图3　Mn元素的监控均方根误差随遗传代数的变化趋势

Fig. 3　Variation tendency of root-mean-square error of monitoring (RMSEM) with different genetic generations for Mn

3.4　定量分析结果对比

以Mn元素为例,用于建立PLS和GA-PLS模型的谱线如图4所示,图中"↓"指向Mn的7条特征谱线。通过GA优化谱线选择后,谱线数目从原来的1042条减少至466条。从图4可见,对于Mn的单峰,优化后的谱线选择依然保留谱峰;而对于Mn的重叠峰,由于多条谱线叠加后的谱峰强度并不能反映原子特征谱线的真实强度,所以GA并没有选择重叠谱峰参与PLS建模,说明GA对于重叠峰干扰有一定的识别能力,可以在一定程度上降低重叠峰对定量分析结果的影响。Mn元素的定量分析结果如图5所示,使用GA-PLS模型后,定标集的RMSEC从0.0158%降低为0.0008%,预测集的预测均方根误差(RMSEP)从0.0215%降低为0.0167%,预测集的平均百分比误差(MPE)从8.10%降低为5.20%。可见,GA-PLS模型的定标和预测准确度明显高于PLS模型。由于只有少于一半的光谱数据应用于PLS模型中,并且GA-PLS模型的定量分析结果更好,说明GA有效地去除了与分析元素Mn浓度无关的噪声信号干扰,从而提高了定量分析的准确度。

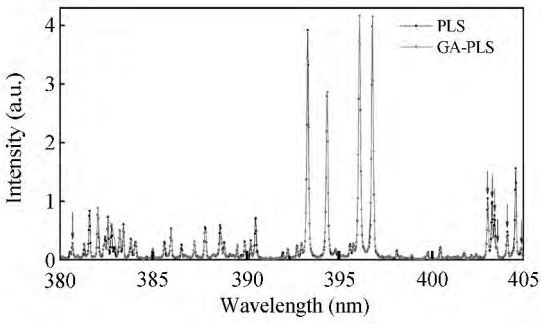

图4　建立Mn元素的PLS和GA-PLS模型的谱线

Fig. 4　Spectral lines of Mn used for building partial least squares (PLS) and genetic algorithm(GA)-PLS models

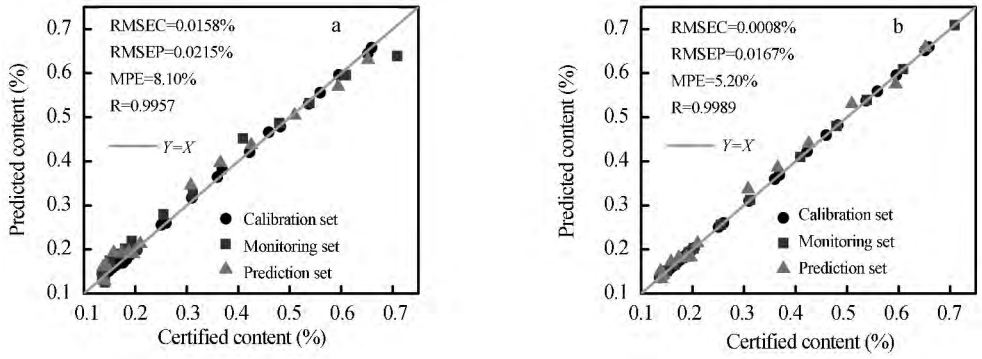

图5　Mn元素的定量分析结果对比

Fig. 5　Comparison of quantitative analysis results of Mn

(a) PLS model; (b) GA-PLS model.

分别应用 PLS 模型和 GA-PLS 模型对土壤中的其余 10 种组成成分 Cr，Cu，Pb，Ba，Al_2O_3，CaO，Fe_2O_3，MgO，Na_2O 和 K_2O 的含量进行预测，表 3 列出了全部定量分析结果，表中阴影表示两种模型中更优的数据。从建模过程中使用的谱线数目可以看出，与传统的 PLS 模型相比，GA-PLS 模型中自变量的数目明显减少，简化了模型。主成分个数与 PLS 保持相等或略有增加。从表征模型定标准确度的 RMSEC 来看，除 Fe_2O_3 外的所有分析成分的 GA-PLS 模型的 RMSEC 值都降低，说明 GA-PLS 模型的定标效果更好。从表征模型预测能力的 RMSEP 和 MPE 来看，在土壤的 11 种组成成分中有 7 种成分的 GA-PLS 模型的 RMSEP 值更小，有 8 种成分的 GA-PLS 模型的 MPE 值更小。因此，对于多数组分，GA-PLS 模型能够显著改善传统 PLS 模型的预测精度。而对于少数组分，GA-PLS模型比 PLS 模型的预测精度差，是由于 GA 在优化谱线选择的过程中去除了光谱区间内的一部分有用信息，因此 GA-PLS 模型有待进一步改进。

表 3　PLS 和 GA-PLS 模型定量分析结果对比

Table 3　Comparison of quantitative analysis results between PLS and GA-PLS models

成分 Composition	谱线数目 Number of spectral points		主成分个数 Factors		RMSEC (%)		RMSEP (%)		MPE (%)	
	PLS	GA-PLS	PLS	GA-PLS	PLS	GA-PLS	PLS	GA-PLS	PLS	GA-PLS
Mn	1042	466	6	12	0.0158	0.0008	0.0215	0.0167	8.10	5.20
Cr	313	148	5	6	0.0007	0.0003	0.00093	0.00086	6.64	5.00
Cu	471	213	5	5	0.0022	0.0014	0.0032	0.0031	15.40	12.54
Pb	404	191	4	4	0.0025	0.0019	0.0025	0.0029	4.82	5.07
Ba	358	179	4	9	0.0005	0.0002	0.00063	0.00061	1.26	1.29
Al_2O_3	528	234	4	4	0.1476	0.1349	0.2207	0.2112	0.91	0.81
CaO	520	257	4	6	0.0756	0.0170	0.0798	0.0831	11.43	9.58
Fe_2O_3	1741	777	7	3	0.0798	0.0860	0.1083	0.0902	0.87	0.72
MgO	157	60	4	5	0.0244	0.0180	0.0376	0.0277	3.61	2.44
Na_2O	277	130	3	3	0.0296	0.0222	0.0332	0.0360	11.63	10.93
K_2O	427	214	4	4	0.0604	0.0467	0.0944	0.1001	4.59	5.02

RMSEC: Root-mean-square error of calibration; RMSEP: Root-mean-square error of prediction; MPE: Mean percent prediction error.

4　结　论

本实验将 GA-PLS 模型应用于土壤的 LIBS 定量分析，对 Mn，Cr，Pb，Ba，Al_2O_3，CaO，Fe_2O_3 等 11 种组分的含量进行预测。结果表明，与传统的 PLS 方法相比，GA-PLS 可以有效去除光谱中重复、多余或不相关的变量，减少用于 PLS 建模的光谱谱线的数目，从而减少建模的时间，简化模型。对于土壤中的多数组分，GA-PLS 模型能够显著改善传统 PLS 模型的预测能力; 而对于少数组分，GA-PLS 模型比 PLS 模型的预测精度差，此部分 GA-PLS 模型的改进将是下一步研究的重点。

References

1　Wang Z，Yuan T B，Hou Z Y，Zhou W D，Lu J D，Ding H B，Zeng X Y. *Front. Phys.*，**2014**，DOI: 10.1007/s11467-013-0410-0

2　Lee W B，Wu J Y，Lee Y I，Sneddon J. *Appl. Spectrosc. Rev.*，**2004**，39(1): 27 - 97

3　SHAO Yan，ZHANG Yan-Bo，GAO Xun，DU Chuang，LIN Jing-Quan. *Spectroscopy and Spectral Analysis*，**2013**，33(10): 2593 - 2598

邵 妍，张艳波，高 勋，杜 闯，林景全. 光谱学与光谱分析，**2013**，33(10): 2593 - 2598

4　LU Cui-Ping，LIU Wen-Qing，ZHAO Nan-Jing，LIU Li-Tuo，CHEN Dong，ZHANG Yu-Jun，LIU Jian-Guo. *Acta Phys. Sin.*，**2011**，60(4): 045206

鲁翠萍，刘文清，赵南京，刘立拓，陈 东，张玉钧，刘建国. 物理学报，**2011**，60(4): 045206

5　Freedman A，Jr. F J I，Wormhoudt J C. *Spectrochim. Acta*，*B*，**2005**，60: 1076 - 1082

6　Gonzaga F B，Pasquini C. *Spectrochim. Acta*，*B*，**2008**，63: 56 - 63

7　Yuan T B，Wang Z，Liu S L，Fu Y T，Li Z，Liu J M，Ni W D. *J. Anal. At. Spectrom.*，**2013**，28: 1045 - 1053

8 Braga J W B, Trevizan L C, Nunes L C, Rufini I A, Jr. D S, Krug F J. *Spectrochim. Acta*, *B*, **2010**, 65: 66 – 74
9 XU Lu, SHAO Xue-Guang. *Methods of Chemometrics*. Beijing: Science Press, **2004**: 218 – 235
许 禄, 邵学广. 化学计量学方法. 北京: 科学出版社, 2004: 218 – 235
10 Bangalore A S, Shaffer R E, Small G W. *Anal. Chem.*, **1996**, 68: 4200 – 4212
11 Leardi R. *J. Chemometrics*, **2000**, 14: 643 – 655
12 Breitkreitz M C, Jr I M R, Rohwedder J J R, Pasquini C, Filho H A D, Jose G E, Araujo M C U. *Analyst*, **2003**, 128: 1204 – 1207
13 CHU Xiao-Li, YUAN Hong-Fu, WANG Yan-Bin, LU Wan-Zhen. *Chinese J. Anal. Chem.*, **2001**, 29(4): 437 – 442
褚小立, 袁洪福, 王艳斌, 陆婉珍. 分析化学, 2001, 29(4): 437 – 442
14 Fink H, Panne U, Niessner R. *Anal. Chem.*, **2002**, 74(17): 4334 – 4342
15 Gottfried J L, Harmon R S, Jr. F C D L, Miziolek A W. *Spectrochim. Acta*, *B*, **2009**, 64: 1009 – 1019
16 Anderson R B, Morris R V, Clegg S M, Bell J F, Wiens R C, Humphries S D, Mertzman S A, Graff T G, Mcinroy G R. *Icarus*, **2011**, 215: 608 – 627
17 Unnikrishnan V K, Nayak R, Aithal K, Kartha V B, Santhosh C, Suri B M. *Anal. Methods*, **2013**, 5: 1294 – 1300
18 Ebinger M H, Norfleet M L, Breshears D D, Cremers D A, Ferris M J, Unkefer P J, Lamb M S, Goddard K L, Meyer C W. *Soil Sci. Soc. Am. J.*, **2003**, 67: 1616 – 1619
19 Haddad J E, Villot-Kadri M, Ismael A, Gallou G, Michel K, Bruyere D, Laperche V, Canioni L, Bousquet B. *Spectrochim. Acta*, *B*, **2009**, 79-80: 51 – 57
20 Martin M Z, Mayes M A, Heal K R, Brice D J, Wullschleger S D. *Spectrochim. Acta*, *B*, **2013**, 87: 100 – 107
21 Kramida A E, Ralchenko Y, Reader J. http://physics.nist.gov/asd

Quantitative Analysis of Soil by Laser-induced Breakdown Spectroscopy Using Genetic Algorithm-Partial Least Squares

ZOU Xiao-Heng, HAO Zhong-Qi, YI Rong-Xing, GUO Lian-Bo, SHEN Meng,
LI Xiang-You*, WANG Ze-Min, ZENG Xiao-Yan, LU Yong-Feng
(*Wuhan National Laboratory for Optoelectronics, Laser and Tera-Hertz Technology Division, Huazhong University of Science and Technology, Wuhan* 430074, *China*)

Abstract Laser-induced breakdown spectroscopy (LIBS) was used to detect the compositions of soil in the air, and the quantitative analysis model with genetic algorithm-partial least squares (GA-PLS) was established. A total of fifty-eight soil samples were split into calibration, monitoring and prediction sets. Eleven soil compositions including Mn, Cr, Cu, Pb, Ba, Al_2O_3, CaO, Fe_2O_3, MgO, Na_2O, and K_2O were quantitatively analyzed. The results demonstrated that, as a pretreatment method for optimizing the selection of spectral lines, GA could be effectively used to reduce the number of spectral lines for use in building PLS model, and hence simplify the quantitative analysis model. More importantly, for most of the soil compositions, GA-PLS could significantly improve the prediction ability compared with the conventional PLS model. Take Mn as an example, the root-mean-square error of prediction (RMSEP) was decreased from 0.0215% to 0.0167%, and the mean percent prediction error (MPE) was decreased from 8.10% to 5.20%. The research provides an approach for further improving the accuracy of LIBS quantitative analysis in soil.

Keywords Laser-induced breakdown spectroscopy; Genetic algorithm; Partial least squares; Soil compositions analysis

(Received 6 August 2014; accepted 29 October 2014)

This work was supported by the National Major Scientific Instrument and Equipment Development Project (No. 2011YQ160017), the 54-class General Financial Grant from the China Postdoctoral Science Foundation (No. 2013M542014), and the Fundamental Research Funds for the Central Universities of Ministry of Education of China (Nos. CXY13Q021, CXY13Q022)