

# 优化 CARS 结合 PSO-SVM 算法农田土壤 重金属砷含量高光谱反演分析

袁自然, 魏立飞\*, 张杨熙, 余 铭, 闫芯茹

湖北大学资源与环境学院, 湖北 武汉 430002

**摘 要** 土壤重金属污染是由于人类活动导致重金属物质大量残留在土壤中, 超过土壤环境承载力, 这种现象将造成土壤质量退化、生态环境恶化。高光谱遥感可以实现图谱合一, 能有效地识别出土壤中不同元素的异常情况。为实现农田土壤重金属高效、准确监测, 提出了一种特征提高型竞争性自适应重加权算法(CARS)选取特征波段的粒子群算法(PSO)优化支持向量机(SVM)农田土壤重金属砷(As)含量高光谱估测分析方法。利用 CARS 对暗室实测光谱值进行粗选; 利用一阶导数(FD)、高斯滤波(GF)、归一化(N)进行特征提高; 在特征精选阶段利用皮尔逊相关系数(PCC)求取预处理后的光谱指标与土壤重金属 As 之间的相关系数, 获取相关性大于 0.6 的波段作为特征波段; 最后利用 PSO 对 SVM 所选择的核函数  $\sigma$  和正则化参数  $\gamma$  进行优化, 以均方根误差(RMSE)作为适应度函数, 通过迭代最优适应度得到 SVM 最优参数值。选择江汉平原典型区域洪湖市燕窝镇的土壤为研究对象, 预测结果表明基于 PSO-SVM 算法其验证集的决定系数  $R^2$  为 0.982 3, 均方根误差 RMSE 为 0.521 6, 平均绝对误差 MAE 为 0.416 4。主要结论如下: PSO 算法优化 SVM 参数, 通过迭代更新个体极值和群体极值, 可以迅速获取全局最优解, 与支持向量机回归(SVMR)和随机森林回归(RFR)相比, 在预测精度有了较大的提高; 特征提高型 CARS 算法可以有效剔除无关信息, 提高相关性, 且选取波段少, 模型简单, 大大提高了效率; 可以实现土壤污染预警、满足精准农业需求、为后期重金属污染土地生态修复提供数据基础。

**关键词** 高光谱遥感; 土壤重金属; 粒子群算法; 特征波段; 竞争性自适应重加权算法

**中图分类号:** TP79 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2020)02-0567-07

## 引 言

重金属污染已经严重危害到人类的生活, 其难降解、易积累、毒性大等特征对农作物生长、产量和品质都有重大影响。农田土壤重金属污染会抑制土壤环境的优势细菌群落, 抑制微生物的生长, 尤其是它可以通过农作物作为媒介通过食物链, 最终威胁人类身体健康<sup>[1]</sup>。传统的土壤重金属检测方法有光度法、化学分析法、原子荧光光谱法、电感耦合等离子体发射光谱法、表面增强拉曼光谱法等<sup>[2]</sup>, 这类方法虽精度高, 但是仪器设备成本过大, 耗费大量的人力、物力, 且不具有推广性。高光谱分析利用连续的高分辨率光谱波段预测土壤中的重金属含量<sup>[3]</sup>。它可以实现大面积的快速测

定, 避免了复杂的采样步骤, 而且结合采样对比, 高光谱对重金属的测定结果也有很高的可信度。

国内外学者对土壤重金属高光谱反演已经有了很多研究, Gholizadeh 等<sup>[4]</sup>提出了一种基于支持向量机回归(SVMR)交叉验证的多变量校准方法, 建立了可见-近红外(Vis-NIR)区域反射光谱与土壤中锰(Mn)、铜(Cu)、镉(Cd)、锌(Zn)和铅(Pb)之间的关系。Angelopoulou 等<sup>[5]</sup>通过偏最小二乘法建模发现不同化学监测方法对于土壤重金属 Cd, Cr, Cu 和 Pb 反演结果影响不同, 表明土壤反射率完全可以预测土壤中重金属含量。

谭琨等<sup>[6]</sup>选择中国北方金属矿区和煤矿区的两个研究区域, 利用竞争自适应重加权算法结合偏最小二乘法(CARS-PLSR)选取特征波段, 结果表明与其他线性模型相比, 非线性

收稿日期: 2018-12-10, 修订日期: 2019-05-04

基金项目: 国家重点研发计划课题(2017YFB0504202), 湖北省技术创新重大专项(2018ABA078), 空间数据挖掘与信息共享教育部重点实验室开放基金项目(2018LSDMIS05), 农业部农业遥感重点实验室开放基金项目(20170007)资助

作者简介: 袁自然, 1995年生, 湖北大学资源与环境学院硕士研究生 e-mail: yuanziran11@163.com

\* 通讯联系人 e-mail: weilifei2508@163.com

性模型 CARS-PLS-SVM 通过适当的光谱特征提取，在实验中表现出较高的精度。但是目前反演精度受到原始光谱相关性低的限制，导致反演精度普遍不高，且现有特征选取方式和反演方法都有待提高。

本文以洪湖市燕窝镇土壤为研究对象，首先利用 CARS 智能选取特征波段，再利用一阶导数、高斯滤波、归一化进行特征提高，获取皮尔逊相关系数大于 0.6 的波段作为特征波段，采用 3 种回归方式 PSO-SVM, SVMR 和 RFR 作为对比，提高建模精度，找出该区域反演土壤重金属 As 的最优方法。

### 1 研究区概况

洪湖市(113°07'—114°05'E, 29°39'—30°12'N)隶属于湖北省荆州市，该区域土壤具有弱富铝化、粘化、酸性的特点，土壤性质兼有黄壤和棕壤的某些特征。燕窝镇为洪湖市辖镇，为传统粮棉油种植大镇，位于洪湖东北部，优良的气候条件和充足的水资源，使得该区域的农业和养殖业得到了快

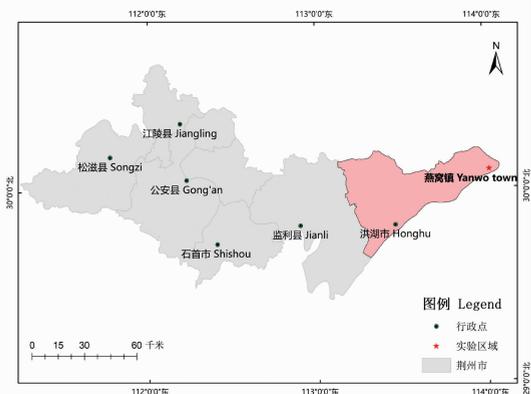


图 1 研究区区位图  
Fig. 1 The location of study area

速发展，但是大量的农田改造和过量施用农药化肥等致使该地区污染逐渐严重，洪湖水水质条件遭到严重破坏，很多耕地受到 As, Cd, Pb, Cd 等重金属污染的威胁，生态环境受到严峻挑战。

### 2 研究方法

#### 2.1 竞争性自适应加权算法

竞争性自适应重加权算法 (competitive adaptive re-weighted sampling)<sup>[7]</sup> 是种用于光谱特征筛选的新型算法，将每个光谱波段的集合作为单独个体，采用自适应重加权采样 (adaptive reweighted sampling, ARS)，利用线性模型偏小二乘法 (PLS) 作为适应度函数，和交叉验证不断优化计算，最终选择出最优子集，即回归模型精度最高子集，淘汰误差较大的变量，经过多次循环采样，选择出特征波段。

#### 2.2 基于 PSO 的 SVM 参数优化

支持向量机 (support vector machine, SVM) 是种常用非线性负荷预测模型<sup>[8]</sup>，粒子群优化 (particle swarm optimization, PSO) 算法是一种基于迭代寻优的群计算技术<sup>[9]</sup>。使用 PSO 算法对 SVM 的参数进行参数寻优，利用粒子群算法特征，不断更新粒子适应度，直到找到全局最优解，本文采用 PSO 优化 SVM 的参数<sup>[10]</sup>，其步骤流程如图 2 所示。

(1) 初始化 PSO 和 SVM 参数：设定群体规模  $n = 30$ 、最大迭代次数  $N = 300$ 、学习因子  $C_1 = 1.8$  和  $C_2 = 1.2$ ，粒子的初始位置和速度等。

(2) 分别利用每个粒子对所对应的实验样本的光谱和重金属含量进行预测，得到各粒子当前均方根误差 (RMSE)，将目前最优适应度与粒子本身最优适应度对比，如果目前最优适应度优于粒子本身最优适应度，则更新粒子速度和位置，从而更新个体极值。

(3) 检查是否满足极值条件，即最大迭代次数和精度要求，如果满足则将最优核函数  $\sigma$  和正则化参数  $\gamma$ ，代入 SVM 计算，否则执行步骤 2，继续搜索。

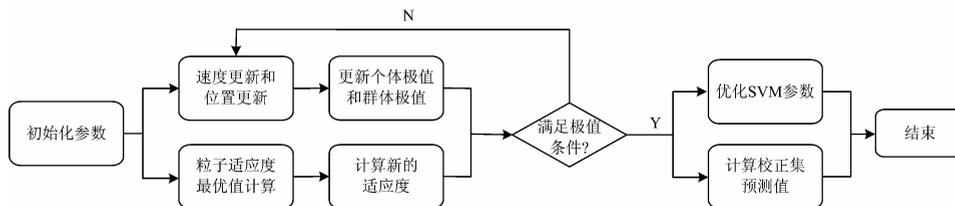


图 2 PSO 优化 SVM 参数流程图  
Fig. 2 Flow chart of optimizing the parameters of SVM based on PSO

#### 2.3 模型验证

选取了决定系数 (determination coefficients,  $R^2$ )、均方根误差 (root mean squared error, RMSE)、平均绝对误差 (mean absolute error, MAE) 等 3 个参数衡量评估模型精度<sup>[11]</sup>。 $R^2$  越接近 1，模型的拟合程度越高，RMSE 和 MAE 越小，模型的预测能力越高，模型鲁棒性越高。

### 3 实验部分

#### 3.1 土壤采集制备和理化分析

在洪湖市燕窝镇内进行土壤采样，该区域土壤为黄棕壤，定点采集耕层表层土壤土样 29 份，深度在 0~20 cm，将土样用聚乙烯袋密封干燥保存。然后将土样利用搪瓷盘摊开盛放，放于通风处自然风干，将风干土壤剔除石块、植物残

体等异物,用木棍研磨土壤样品,首先使用 2 mm 孔径筛筛选土壤,取出四分之一继续磨细使通过 0.15 mm 孔径筛作为最终土壤样品,研磨后的土壤样本分为两份,分别用于光谱采集和理化分析测试。理化分析的土样经硝酸,盐酸,高氯酸消煮后,采用硼氢化钾-硝酸银分光光度法进行实验测定,每份土样测量 3 次,从测量结果的算术平均值得出土壤中的重金属 As 含量。

### 3.2 土壤光谱测定

采用 SVC HR-1024 型地物光谱仪对土壤样品进行暗室光谱的测定,波长范围为 350~2 500 nm,光谱分辨率为:350~1 000 nm 为 1.5 nm,1 000~1 900 nm 为 3.8 nm,1 900~2 500 nm 为 2.5 nm,总波段数为 990。室内光谱测量具有易控制、数据结果质量较高等特点。将土壤放于通体漆黑的培养皿中,并利用钢尺刮平表层土壤。在暗室中,用 1 000 W 的卤素灯作为光源,照射方向与垂直方向呈 45° 夹角,光源距离土样表面约为 30 cm,光谱仪探头垂直于土壤表层上方,探头距离土样约为 10 cm。每次测量光谱之前,需对光谱仪进行白板校正。为消除测量数据误差和其不稳定性,每个土样旋转三次,每次转动 90°,以获得四个方向的反射率,有 4 条光谱曲线,其平均值作为各土壤样本的原始反射率光谱值,如图 3 所示。

### 3.3 数据处理和分析

#### 3.3.1 光谱预处理

由于仪器自身的原因和光谱采集过程中不可避免的受到测试环境、样品背景、观测角度、样品粗糙度、杂散光等因素的影响,光谱曲线边缘波段噪声比较大。为降低外界噪声对实际建模的干扰,每份土样都去除噪声较大的边缘波段 350~399 和 2 400~2 500 nm,保留 400~2 399 nm 波段用于建模分析,如图 4 所示。

#### 3.3.2 土壤 As 含量分析

对 29 个样点土壤重金属砷含量进行了统计分析,结果如表 1 所示。参考《土壤环境质量标准 GB15618—1995》,洪湖燕窝镇采集的土壤重金属元素 As 含量的平均值超过三级质量标准(为保障农业生产和植物生长的土壤污染临界值),表明该地区土壤的砷污染达到中度和重度污染。砷对于农作

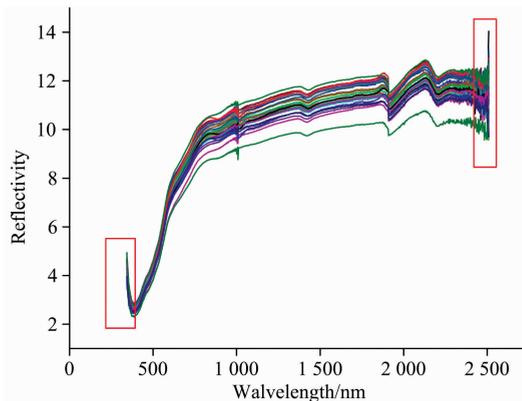


图 3 原始光谱

Fig. 3 Original spectra

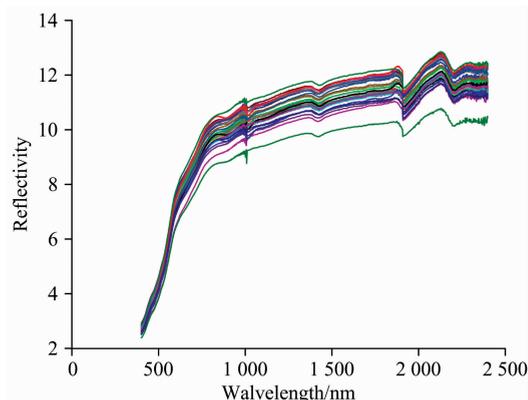


图 4 去除边缘噪声

Fig. 4 Edge noise removal

物的生长危害十分严重,可以通过植物根茎部分,利用植物生长吸收作用转移到植物各部位,初期表现为叶片枯萎,中期为根系停止发育,最后整株植物枯萎死亡。砷对于人体而言同样危害巨大,可以通过食物链累积,迅速达到临界值,它可以使红血球溶解,破坏正常细胞,并且具有遗传性、致癌性和致畸性等,所以对于重金属砷的监测具有重要意义<sup>[12]</sup>。

表 1 各数据集 As 值的描述性统计

Table 1 Statistical characteristics of As value

重金属类型	组别	样本数	最大值 /( $\mu\text{g} \cdot \text{g}^{-1}$ )	最小值 /( $\mu\text{g} \cdot \text{g}^{-1}$ )	均值	标准差	变异系数 /%
砷	全部	29	39.21	26.54	32.86	3.44	10
	校准集	21	39.21	26.54	32.86	3.82	11.6
	验证集	8	39.07	27.08	33.24	3.92	11.7

在进行建模之前,需要将样本分组,一组用于模型的构建,即校准集;另一组用于验证模型的预测能力,即验证集;如表 1 所示。采用浓度梯度法,最终选择 21 个校准样本,8 个验证样本,校准集和验证集的均值,标准差和变异系数均

较为接近。因此,这样划分比较合理,可以用于后续建模。对于建立一个满意的预测模型,校准采样算法和校准集的大小对整组数据集的代表性有着深刻的影响<sup>[13]</sup>。

### 4 结果与讨论

#### 4.1 特征提高型 CARS 特征波段选取

特征波段选取流程如图 5 所示,首先在特征粗选阶段利用 CARS 对暗室实测光谱进行粗选,设定蒙特卡罗采样为 100 次,从图 6 可以看出经过交叉检验(CV)所得到的 RM-

SECV 的变化趋势图, RMSECV 的变化曲线先由大到小至最小后再逐渐变大的变化,当采样循环次数为 60 时, RMSECV 值达到最小值,表明此时模型精度最高,而随着采样次数变大, RMSECV 上升,模型效果变差。因此,当采样次数为 60, RMSECV 值达到最小值时,所选择波段子集为最优。但在 CARS 算法过程中,采用偏最小二乘法为适应度函数,这种线性模型面对高维数据,会出现精度不高等问题,

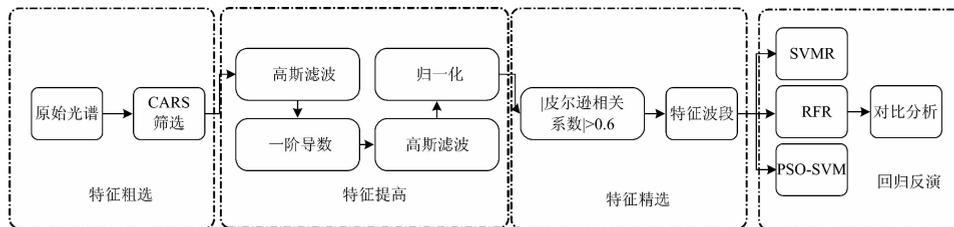


图 5 CARS 特征波段选择流程图

Fig. 5 Characteristic band selection flow chart of CARS

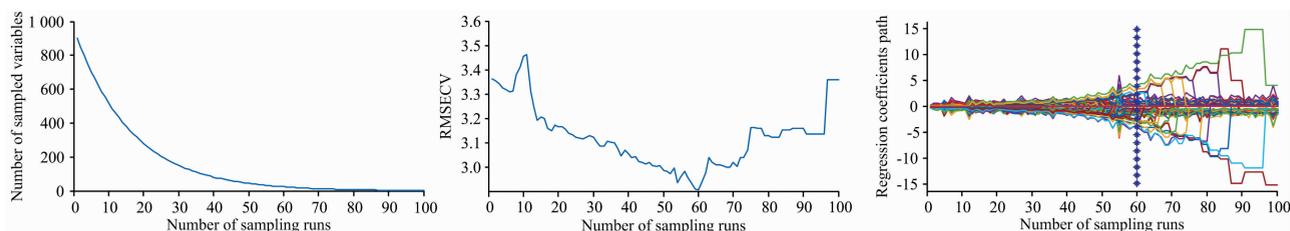


图 6 CARS 关键变量选择

Fig. 6 Key variables selection results of CARS

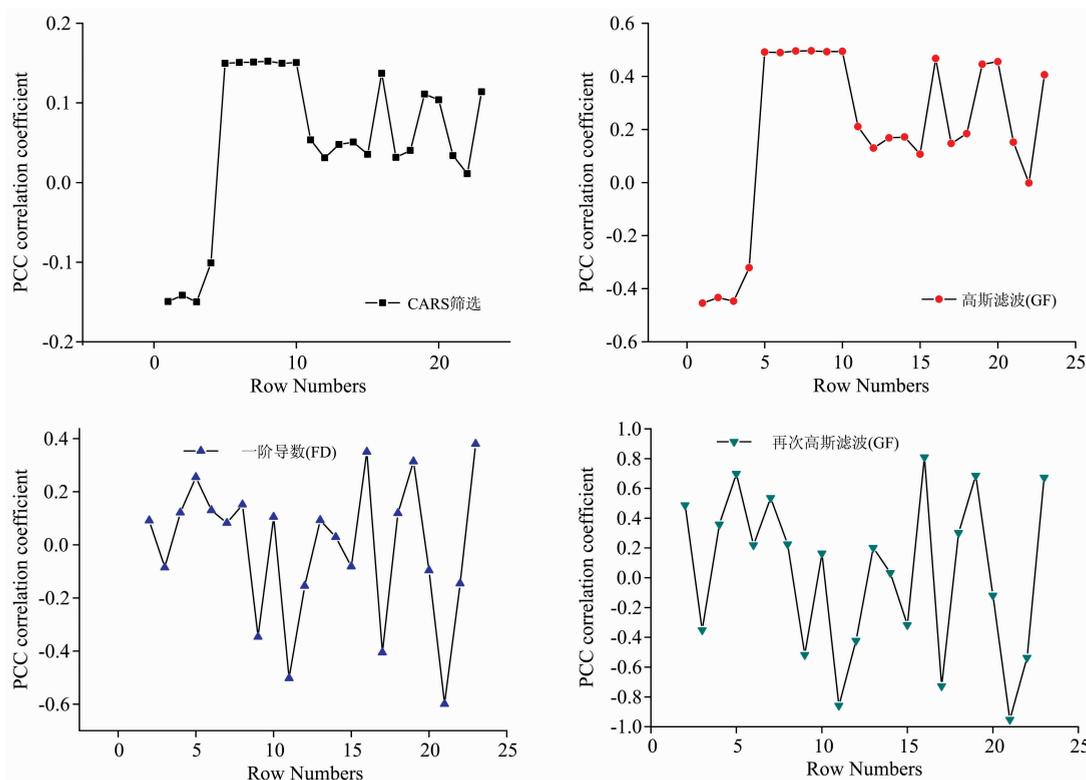


图 7 不同预处理情况下相关系数变化

Fig. 7 Change of correlation coefficient under different pretreatment conditions

且 CARS 算法未改变原始数据, 面对相关性较低数据, 往往无法使用, 所以有必要提高特征。在特征提高阶段, 从图 7 可以看出 CARS 筛选出来的原始数据与重金属 As 之间的相关性较低, 而经过高斯滤波、一阶导数, 再次高斯滤波后相关性大大提高。

在特征精选阶段, 利用皮尔逊相关系数 (pearson correlation coefficient, PCC) 求取预处理后的光谱指标与土壤重金属 As 之间的相关系数, 获取绝对值相关性大于 0.6 的波段作为特征波段。

#### 4.2 回归预测

特征提高型 CARS 特征波段选取选择了 7 个特征波段用于建模, SVM 模型的预测效果与核函数  $\sigma$  和正则化参数  $\gamma$  有密切的关系, 极大程度影响着预测精度, 如何通过优化获取最优参数是非常重要的步骤。因此选取了基于 PSO 优化 SVM 算法来选取最优参数获得最佳模型, 在 PSO 优化模型中, 粒子群的种群数  $N=30$ , 最大的迭代次数  $G_{\max}=300$ , 学习因子  $C_1$  和  $C_2$  为 1.8 和 1.2, 核函数形状参数  $\sigma$  取值范围为  $[0.1 \sim 1\ 000]$ , 正则化参数  $\gamma$  取值范围为  $[0.01 \sim 100]$ 。分别利用每个粒子对对应的实验样本的光谱和重金属含量进行预测, 得到各粒子当前均方根误差 (RMSE), 将目前最优适应度与粒子本身最优适应度进行对比, 如果目前最优适应度优于粒子本身最优适应度, 则更新速度和位置, 最终获得核函数为径向基核函数 (RBF), 核函数形状参数  $\sigma$  为 100, 正则化参数  $\gamma$  为 0.357 3。将最优核函数和正则化参数代入到 SVM 模型中, 用于土壤重金属 As 含量高光谱预测。从图 8 中可以看出, 适应度值逐渐向最优值接近, 一旦达到最优值后就趋于稳定, 利用 PSO-SVM 模型可以有效预测土壤重金属 As 含量, 其中  $R^2$ , RMSE, MAE 达到 0.982 3, 0.521 6 和 0.416 4, 再利用 SVMR 和 RFR 回归作为对比, 从图 9 可以看出验证集基本上呈现线性均匀分布, 表现出较好的预测水平和稳定性, 满足了实际预测要求。其中 SVMR 表现最差, 偏离较大, 而 PSO-SVM 离 1:1 线偏离最小, 拟合程度最高。

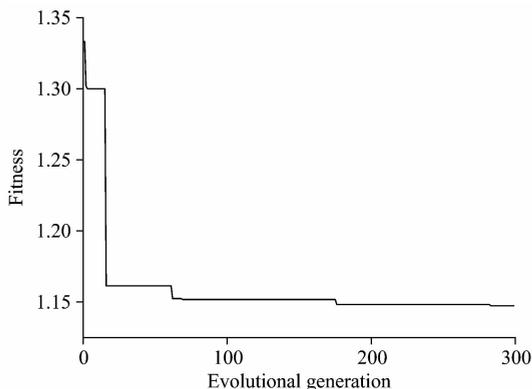


图 8 适应度曲线变化

Fig. 8 Fitness curve

对比传统化学分析法, 利用土壤反射率高光谱反演土壤重金属 As 含量的方法, 有效率高、易操作、环境友好型等优

点, 对于前期生态环境监测和后续生态修复提供数据支持具有重要意义。

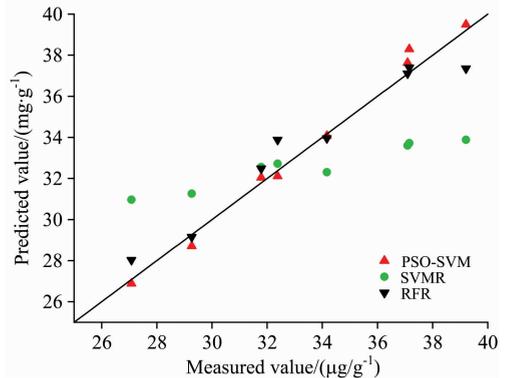


图 9 实测值与预测值关系

Fig. 9 Relationship between measured and predicted values

研究中分别利用三种方法建立了高光谱重金属 As 含量预测模型, 从表 2 可以看出 SVMR 预测效果较差, 而 PSO-SVM 验证集的反演精度表现最优。PSO 优化 SVM 的回归模型精度上有了极大地提高, 利用优化得到的核函数  $\sigma$  可以保证模型反演精度, 正则化参数  $\gamma$  可以使模型具有较好的泛化能力, 为预测其他重金属元素提供了参考依据。

表 2 算法精度对比

Table 2 Accuracy comparison of algorithms

建模方法	验证集		
	$R^2$	RMSE	MAE
SVMR	0.383 8	3.074 0	2.636 9
RFR	0.941 3	0.944 8	0.697 8
PSO-SVM	0.982 3	0.521 6	0.416 4

## 5 结 论

以江汉平原典型区域洪湖市燕窝镇污染土壤为研究对象, 采集总计 29 个土壤样本, 21 为校准集, 8 个为验证集, 基于三种不同模型构建方法, 重点讨论了利用土壤反射率光谱可以较好的反演土壤重金属 As 含量, 主要结论如下:

(1) 特征提高型 CARS 算法, 可以有效剔除无关冗余信息, 极大地提高了光谱与重金属 As 含量之间相关性, 且选取波段少, 模型简单, 大大提高了效率。

(2) 对比支持向量机回归 (SVMR)、随机森林回归 (RFR) 等建模方式, 采用 PSO-SVM 通过迭代更新个体极值和群体极值, 获得最佳参数, 具有较高的精度。PSO-SVM 的验证集的  $R^2$  为 0.982 3、RMSE 为 0.521 6、MAE 为 0.416 4。说明模型具有较高的稳定性和预测精度, 可以满足实际预测需求, 为今后大规模反演洪湖地区土壤重金属含量提供依据。此外, 区域构建高光谱模型也可为无人机、航空、航天遥感提供参考。

## References

- [ 1 ] Huang Y, Chen G, Xiong L, et al. *Asian Agricultural Research*, 2016, 8(1): 22.
- [ 2 ] Hu B, Chen S, Hu J, et al. *Plos One*, 2017, 12(2): e0172438.
- [ 3 ] YUAN Zhong-qiang, CAO Chun-xiang, BAO Da-ming, et al(袁中强, 曹春香, 鲍达明, 等). *Wetland Science(湿地科学)*, 2016, 14(1): 113.
- [ 4 ] Gholizadeh A, Luboš Borůvka, Saberioon M M, et al. *Soil and Water Research*, 2016, 10(4): 218.
- [ 5 ] Angelopoulou T, Dimitrakos A, Terzopoulou E, et al. *Water Air & Soil Pollution*, 2017, 228(11): 436.
- [ 6 ] Tan K, Wang H, Zhang Q, et al. *Journal of Soils and Sediments*, 2018, 18(5): 2008.
- [ 7 ] ZHANG Dong, TASHPOLA · Tiyp, ZHANG Fei, et al(张 东, 塔西甫拉提·特依拜, 张 飞, 等). *Acta Optica Sinica(光学学报)*, 2016, 36(3): 282.
- [ 8 ] WANG Wen-jun, LI Zhi-wei, WANG Can, et al(王文俊, 李志伟, 王 璨, 等). *Spectroscopy and Spectral Analysis(光谱学与光谱分析)*, 2019, 39(5): 1579.
- [ 9 ] Angelopoulou T, Tziolas N, Balafoutis A, et al. *Remote Sens.*, 2019, 11: 676.
- [10] Chen T, Chang Q, Clevers J G, et al. *Environmental Pollution*, 2015, 206: 217.
- [11] Wei L, Yuan Z, Zhong Y, et al. *Appl. Sci.*, 2019, 9: 1943.
- [12] CHEN Yi-xian, JIANG Xiao-san, WANG Yong, et al(陈怡先, 姜小三, 王 勇, 等). *Acta Scientiae Circumstantiae(环境科学学报)*, 2018, 38(4): 1642.
- [13] Sun W, Zhang X, Sun X, et al. *Geoderma*, 2018, 327: 25.

# Hyperspectral Inversion and Analysis of Heavy Metal Arsenic Content in Farmland Soil Based on Optimizing CARS Combined with PSO-SVM Algorithm

YUAN Zi-ran, WEI Li-fei\* , ZHANG Yang-xi, YU Ming, YAN Xin-ru

Faculty of Resources and Environmental Science, Hubei University, Wuhan 430002, China

**Abstract** Heavy metal pollution in soil is caused by human activity factors that bring heavy metals into the soil, resulting in deterioration of soil quality and ecological environment. Heavy metals in the soil tend to accumulate, are difficult to be degraded, are highly concealed for long periods of time, and can be enriched by atmospheric circulation and food chains, ultimately threatening human life and health. Hyperspectral remote sensing technology presents a combination of image and spectrum, and can effectively identify the abnormal conditions of different elements in the soil. At present, traditional soil monitoring techniques mainly rely on laboratory-based chemical detection methods such as photometry, chemical analysis, and atomic fluorescence spectroscopy. This kind of method can test the heavy metal content of farmland soil, but the precision depends on a large amount of manpower, material resources and equipment, and its detection efficiency and promotion are lacking. In order to achieve efficient and accurate monitoring of heavy metals in farmland soils. A method of hyperspectral estimation of heavy metal arsenic (As) content in farmland soils based on particle swarm optimization (PSO) and support vector machine (SVM), which use characteristic-enhanced competitive adaptive reweighted sampling (CARS) was proposed. In the characteristic rough selection stage, the measured spectral values from the darkroom are roughly selected by CARS. In the characteristic improvement stage, First Derivative (FD), Gaussian Filtering (GF), Normalization (N) are used to improve features. In the carefully chosen stage, Pearson Correlation Coefficient (PCC) is used to obtain the correlation coefficient between different pre-treated spectral indices and soil heavy metal As. The band whose correlation coefficient has an absolute value greater than 0.6 is selected as a feature band. Finally, PSO is used to optimize the kernel parameter sigma and the normalization parameter gamma used by the SVM. The root mean square error (RMSE) is used as the fitness function, and the optimal parameters of SVM are obtained by iterating the optimal fitness. The soil of Yanwo Town in Honghu City, a typical area of Jiangnan plain, was selected as the research object in this paper. The prediction results showed that the decision coefficient ( $R^2$ ) of the verification sets based on PSO-SVM algorithm is 0.982 3, the root mean square error (RMSE) is 0.521 6, and the mean absolute error (MAE) is 0.416 4. The main conclusions are as follows: the PSO algorithm is used to optimize the SVM parameters, and the global optimal solution can be obtained

quickly by iteratively updating the individual extremum and the group extremum. Compared with the support vector machine regression (SVMR) and random forests regression (RFR), the prediction accuracy has been greatly improved; The characteristic enhanced CARS algorithm can effectively eliminate irrelevant information and improve correlation. And it selects fewer bands, simplifies the model so that efficiency is greatly improved; It can realize early warning of soil pollution, meet the needs of precision agriculture and provide data basis for ecological restoration of heavy metal contaminated land in the later period.

**Keywords** Hyperspectral remote sensing; Soil heavy metal; Particle swarm optimization; Characteristic bands; Competitive adaptive reweighted sampling

(Received Dec. 10, 2018; accepted May 4, 2019)

\* Corresponding author

---

## 敬告读者——《光谱学与光谱分析》已全文上网

从 2008 年第 7 期开始在《光谱学与光谱分析》网站([www.gpxygpx.com](http://www.gpxygpx.com))“在线期刊”栏内发布《光谱学与光谱分析》期刊全文,读者可方便地免费下载摘要和 PDF 全文,欢迎浏览、检索本刊当期的全部内容;并陆续刊出自 2004 年以后出版的各期摘要和 PDF 全文内容。2009 年起《光谱学与光谱分析》每期出版日期改为每月 1 日。

《光谱学与光谱分析》期刊社