doi:10.6041/j.issn.1000-1298.2018.10.029

基于 SGA - RF 算法的农业土壤镉浓度反演研究

王轩慧[1,2]　陈建毅[2]　郑西来[1]　朱　成[3]　王轩力[4]　单春芝[5]

(1. 中国海洋大学海洋环境与生态教育部重点实验室, 青岛 266100; 2. 青岛农业大学理学与信息科学学院, 青岛 266109;
3. 中国联合网络通信有限公司济南软件研究院项目管理部, 济南 250000; 4. 山西工程技术学院信息工程与自动化系,
阳泉 045000; 5. 国家海洋局北海环境监测中心, 青岛 266033)

摘要: 在农业土壤重金属高光谱检测领域,土壤镉元素近红外光谱的高维、高冗余特性会严重影响高光谱反演模型的准确性和稳定性。为了解决上述问题,本文提出一种基于斯皮尔曼相关分析的遗传随机森林特征选择算法(SGA - RF)。该算法首先对初始特征集合使用基于斯皮尔曼相关分析的特征预选方法,筛选出大量冗余波段,保留与镉元素相关性最强的特征波段;其次在特征精选阶段,提出一种基于随机森林的适应度函数评估方法,该方法充分结合遗传算法强大的全局搜索能力和随机森林算法较高的反演能力,提高了对相似个体的区分能力,获得具有最小冗余度和最大区分性的最优特征波段子集。为了验证所提算法的有效性,选取青岛市大沽河流域具有代表性的 124 个土壤样品为实验对象,利用 SGA - RF 算法将原始 2 051 个波段优选至 37 个最具代表性的敏感波段,并与现有特征选择算法所建模型进行对比分析。试验结果表明,该特征选择方法与随机森林回归模型相结合具有较低的预测均方根误差(0.060 1),较高的相关系数(0.950 2)和预测相对分析误差(2.03)。作为应用可见/近红外光谱技术定量反演农业土壤镉浓度的重要步骤,SGA - RF 算法以较少的敏感波段达到了较高的反演效果,可为监测土壤重金属污染情况提供一定的理论依据。

关键词: 农业土壤; 镉浓度; 特征波长选择; 斯皮尔曼等级相关分析; 遗传算法; 随机森林

中图分类号: X53; X87　　**文献标识码**: A　　**文章编号**: 1000-1298(2018)10-0261-09

Inversion of Cadmium Content in Agriculture Soil Based on SGA - RF Algorithm

WANG Xuanhui[1,2]　CHEN Jianyi[2]　ZHENG Xilai[1]　ZHU Cheng[3]　WANG Xuanli[4]　SHAN Chunzhi[5]

(1. *Key Laboratory of Marine Environmental Science and Ecology*, *Ministry of Education*, *Ocean University of China*, *Qingdao* 266100, *China*
2. *Science and Information College*, *Qingdao Agricultural University*, *Qingdao* 266109, *China*
3. *Project Management Department*, *China Unicom Ji'nan Software Research Institute*, *Ji'nan* 250000, *China*
4. *Information Engineering and Automation Department*, *Shanxi Institute of Technology*, *Yangquan* 045000, *China*
5. *The Environmental Monitoring Center of North China Sea*, *Qingdao* 266033, *China*)

Abstract: In the field of hyperspectral detection on heavy metal pollution levels in agricultural soils, the accuracy and stability of hyperspectral inversion model for soil cadmium were seriously affected by the high dimensional and high redundancy characteristics in visible/NIR spectra. In order to solve the above problems, Spearman's rank correlation analysis-based genetic algorithm by using random forest (SGA - RF) was proposed to select the characteristic wavelength from hyperspectral data. On the first-layer of feature selection stage, Spearman correlation analysis-based feature selection method was applied to remove redundancy between all spectra features and retain the characteristic wavelength which was the most relevant to the cadmium content. On the second-layer of feature selection stage, a new fitness function based on random forest was proposed, which perfectly combined the strong global search ability of genetic algorithm and the high inversion ability of random forest. With the proposed fitness function to evaluate the viability of individuals, the distinguishing ability between similar individuals was improved

收稿日期: 2018 - 04 - 12　修回日期: 2018 - 05 - 18

基金项目: 国家自然科学基金重点项目(41731280)和国家自然科学基金项目(11701310)

作者简介: 王轩慧(1978—),女,博士生,青岛农业大学讲师,主要从事计算机技术与资源信息研究,E-mail: margaretxuan@aliyun.com

通信作者: 郑西来(1959—),男,教授,博士生导师,主要从事滨海水资源开发利用与保护研究,E-mail: zhxilai@ouc.edu.cn

and a subset of optimal spectra feature set with minimum redundancy and maximum differentiation were obtained. In order to verify the validity of the proposed algorithm, totally 124 representative soil samples collected from the Dagu River Basin were chosen as samples. The optimal feature subset which contained 37 sensitive wavelengths was chosen and used to build soil available cadmium content inversion model, and its performance was compared with that of current feature selection methods. Results indicated that the minimum numbers of wavelength features was selected and meanwhile the prediction performance had lower predictive root mean square error of 0.060 1, higher correlation coefficient of 0.950 2 and residual predictive deviation of 2.03. As an important step for the quantitative inversion of cadmium concentration by using visible/NIR spectra, the research could provide some theoretical basis for monitoring soil heavy metal pollution.

Key words: agriculture soil; cadmium content; characteristic wavelength selection; Spearman's rank correlation analysis; genetic algorithm; random forest

0 引言

农业土壤重金属污染已成为我国主要的环境问题之一[1-2]。土壤中的镉会在农作物中富集并进入食物链,从而对人类健康造成严重威胁[3-4]。因此,密切监测农业土壤中的重金属镉超标情况对于防止农产品污染,保障人体健康有着重要意义。

目前,常规化学方法对土壤重金属浓度的测定不仅周期长,成本高,而且会产生大量危害环境的废弃物[5]。相比之下,可见/近红外光谱技术具有成本低、效率高、环境污染少等优点[6-7],逐渐成为有效监测土壤重金属污染情况的重要手段[8-9]。CHEN 等[10]采用正交信号校正预处理方法和反向传播神经网络建立污水灌溉区的重金属镉反演模型。GHOLIZADEH 等[11]结合各种预处理方法与支持向量机回归建立土壤重金属含量与反射光谱之间的对应关系。TAYEBI 等[12]使用可见/近红外反射光谱和偏最小二乘(PLSR)模型来预测农业土壤中重金属铁的浓度,指出二阶微分预处理与 PLSR 结合建立的模型反演精度较高。

尽管上述方法成功地将可见/近红外光谱技术应用于土壤重金属高光谱反演领域,但由于原始光谱数据中存在大量冗余和不相关特征,从而严重影响了反演模型的性能。特别是农业土壤中重金属镉所对应的可见/近红外光谱波段非常少,因而直接使用原始光谱构建回归反演模型不仅准确率和稳定性偏低,而且运算效率较差。因此,只有通过使用特征选择方法才能选出具有最小冗余度且最具代表性的敏感波段,进一步有效地提高农业土壤镉回归反演模型的预测能力[13]。JIANG 等[14]利用遗传算法结合 PLSR 建立土壤镉含量反演模型。虽然上述方法取得了一定成果,但是该方法对相邻光谱波段的特征值进行了融合,这会导致部分潜在重要特征的丢失。XIA 等[15]采用相关系数法结合 PLSR 成功反演长江流域沉积物中的镉含量。但基于相关系数法的特征选择方法只适用于线性统计方法建立的模型,对于非线性相关及样本分布不均匀的情况,反演精度较低。

本文在重金属镉高光谱反演领域,针对上述特征选择技术的不足,提出一种基于斯皮尔曼等级相关分析的遗传随机森林(Spearman's rank correlation coefficient-based genetic algorithm using random forest,SGA-RF)特征选择算法。基于大沽河流域表层农业土壤可见/近红外高光谱数据,通过与传统遗传算法和基于 K 近邻的遗传算法进行比较,探讨 SGA-RF 算法在农业土壤重金属高光谱反演领域应用的可能性,以期为利用高光谱技术检测农业土壤中低含量重金属污染状况提供一定的理论依据。

1 SGA-RF 特征选择算法

SGA-RF 特征选择算法流程如图 1 所示。该算法由 2 个子算法组成,是一种基于 Wrapper[16] 型的特征选择方法。首先,为了消除冗余和无关信息,采用斯皮尔曼相关分析特征选择方法(算法 1)对 2 051 个预处理后的初始特征集合进行预选,得到了彼此无关联的特征波段子集作为算法 2 的输入。其

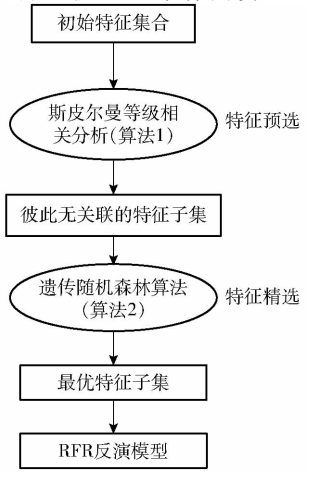

图 1 SGA-RF 特征选择流程图

Fig. 1 Flow chart of SGA-RF feature selection

次，在遗传随机森林算法（算法 2）中采用基于随机森林袋外误差的适应度函数来评价个体的优劣，最终选出适应度函数最小值所对应的特征子集为最优特征波段子集。最后，利用最优特征波段子集构建随机森林回归（Random forest regression，RFR）反演模型。

以斯皮尔曼等级相关分析作为 SGA－RF 算法的特征预选方法，主要目的是有效减少初始特征集维数并剔除原始光谱特征波段之间的冗余和无关信息。该算法包括：①数据准备阶段（步骤 1 ~ 步骤 2），通过计算两两特征波段之间的斯皮尔曼等级相关系数将强相关的特征波段存入二维矩阵 $\boldsymbol{X}$ 中。②算法的核心组成部分（步骤 3 ~ 步骤 8），通过遍历矩阵 $\boldsymbol{X}$ 构造所有相关特征的分组 $\boldsymbol{Y}$。③算法输出部分（步骤 9 ~ 步骤 11），通过对特征波段集合 Q 与 $\boldsymbol{Y}$ 中每组代表波段求并集得到最优光谱特征波段子集。算法的具体描述如算法 1 所示。

（1）算法 1——基于斯皮尔曼等级相关分析的特征选择算法

输入：①预处理后的光谱特征集合 M。②循环变量 $i \leftarrow 1$。③循环变量 $j \leftarrow 1$。④循环变量 $b \leftarrow 1$。⑤循环变量 $c \leftarrow 1$。⑥最大波长数 $L \leftarrow 2\,051$。

输出：所有相关光谱特征分组集合 Y（所有相关的分组）；最优光谱特征子集 F。

步骤 1：计算 M_i 与 M_j 两个特征波段之间的斯皮尔曼等级相关系数，即

$$\rho = |r(M_i, M_j)| = 1 - \frac{6\sum d^2}{N(N^2 - 1)} \tag{1}$$

式中　N——样本数量

d——两列成对变量的等级差分集合

在式（1）中，对 M_i 与 M_j 进行排序（同时为升序或降序），得到两个元素等级集合 m 和 n，其中 m_i 为 m_i 在 M_i 中的等级，n_i 为 n_i 在 M_j 中的等级。将集合 m、n 中的元素对应相减得到一个排行差分集合 d，其中 $d_i = m_i - n_i$，$1 \leqslant i \leqslant N$。如果 ρ 的绝对值（$|\rho|$）大于或等于 0.8 且标准偏差 p 小于 10%，则特征值 M_i 与 M_j 被认为是显著相关的，将 M_i 与 M_j 存入矩阵 $\boldsymbol{X}$ 中，$\boldsymbol{X}$ 为二维矩阵。

步骤 2：重复执行步骤 1，直至 j 达到最大波长数 L。

步骤 3：把 i 赋值给 j，重复执行步骤 1 ~ 步骤 2，直至 i 达到最大波长数 L。

步骤 4：把 $\boldsymbol{X}$ 中的第一行两个元素赋值给 Y_b。设置 *ismatch* 为 0，a 为 1。

步骤 5：取出矩阵 $\boldsymbol{X}_a$ 中第 a 行的第 1 个元素 X_{a1}。如果矩阵 $\boldsymbol{Y}_b$ 中不含有 X_{a1}，转入步骤 6；否则设置 *ismatch* 为 1，再判断矩阵 $\boldsymbol{Y}$ 中是否含有 X_{a2}，如果矩阵 $\boldsymbol{Y}$ 中不含 X_{a2}，将 X_{a2} 存入 $\boldsymbol{Y}_b$ 中，转入步骤 7。

步骤 6：取出矩阵 $\boldsymbol{X}_a$ 中第 a 行的第 2 个元素 X_{a2}。如果矩阵 $\boldsymbol{Y}_b$ 中含有 X_{a2}，设置 *ismatch* 为 1，并判断矩阵 $\boldsymbol{Y}_b$ 中是否含 X_{a1}，如果矩阵 $\boldsymbol{Y}_b$ 中不含有 X_{a1}，将 X_{a1} 存入 $\boldsymbol{Y}_b$ 中。

步骤 7：如果 *ismatch* 为 1，在 $\boldsymbol{X}$ 中剔除 $\boldsymbol{X}_a$，得到新的矩阵 $\boldsymbol{X}_a$ 与 $\boldsymbol{Y}_b$。

步骤 8：重复执行步骤 5 ~ 步骤 7，直到遍历完 $\boldsymbol{X}_a$ 的所有元素。$\boldsymbol{Y}_b$ 中存储一组显著相关的特征波段，$\boldsymbol{X}_a$ 为去掉与 $\boldsymbol{Y}_b$ 相关特征波长之后的剩余特征波长变量。若 $\boldsymbol{Y}_b$ 中含有不相邻波段，对该组进行标注。

步骤 9：重复执行步骤 4 ~ 步骤 8，直到遍历完 $\boldsymbol{X}$ 的所有元素，输出所有相关特征的分组 $\boldsymbol{Y}$，通过计算 $\boldsymbol{M}\backslash\boldsymbol{Y}$（$\boldsymbol{Y}$ 在 $\boldsymbol{M}$ 中的余集）得到特征波段集合 Q。

步骤 10：计算 $\boldsymbol{Y}_c$ 这一组内部所有斯皮尔曼相关系数的平均值和相应的标准偏差，保留与待测组分相关性最大的特征波段作为这一组的代表性特征波段。

步骤 11：重复执行步骤 10，直至处理完 $\boldsymbol{Y}$ 的所有分组，每组代表波段组成集合 P。

步骤 12：通过计算集合 $Q \cup P$，输出最优光谱特征子集 F。

采用算法 1 的特征选择方法，不仅能够剔除原始全光谱的所有相关特征值，而且大幅度地缩减了下一步特征精选（算法 2）的筛选范围，能够达到提高算法反演精度和执行效率的双重目的。对于波长特征选择，遗传算法是一种非常有效的方法。适应度函数[17]的设计在遗传特征选择算法中起着至关重要的作用，但是常规的适应度函数与具体的解决问题领域联系不够紧密，导致最终筛选出的特征子集建立的模型预测精度较低。为此提出了一种新型的适应度函数，综合考虑了特征子集的回归性能和算法的运行效率，该算法的具体描述如算法 2 所示。

（2）算法 2——遗传随机森林的特征选择算法

输入：①光谱特征集合 F。②每一个体的染色体长度（基因）$m \leftarrow$ 集合 F 中元素的数量。③染色体的基因位数 $p \leftarrow 1$。④初始群体包含的个体数 f。⑤最大迭代次数 *Generations* $\leftarrow 200$。⑥停滞代数 *StallGenLimit* $\leftarrow 100$。⑦交叉概率 $p_n \leftarrow 0.8$。⑧变异算子 $p_m \leftarrow 0.5$。

输出：最优光谱特征子集 E。

步骤 1（参数编码）：用一串含有 mp 个 0/1 字符（基因）的字符串（染色体串）来表示每种区间组合。若基因为 1 表示该特征被选中，基因为 0 表示

该特征未被选中。

步骤 2(种群初始化):随机产生 fm 个字符作为初始群体。

步骤 3(新的适应度函数):本算法中的适应度函数被定义为随机森林袋外误差和特征维度的加权和,即

$$F = \omega\alpha + (1-\omega)\frac{1}{m-\beta} \tag{2}$$

式中 α——RFR 袋外误差,是衡量 RFR 模型性能的重要指标

β——特征子集含有的特征值数量,直接决定算法的运行效率

ω——回归精度在适应度函数中所占的比重

m——染色体长度(原始特征值数量)

步骤 4(复制):采用精英保留策略。设置精英个数 c_1 为 2,按照式(2)计算所有个体的适应度,选出适应度最低的前两名个体不进行配对交叉自动推到下一代。

步骤 5(选择):采用锦标赛选择策略。当精英子代从当前种群移除后,对剩余 $m-2$ 个个体执行锦标赛选择算法。设置参加锦标赛的个体个数为 2,依据式(2)评估所有个体的适应度,从种群中选出 2 个最优个体作为父代个体。

步骤 6(交叉):对锦标赛策略生成的父代个体进行交叉操作。选用的交叉函数是算术交叉类型的异或运算,对两个父类染色体进行异或操作产生交叉子代。产生交叉子代的数量为 $c_2 = \mathrm{round}((m-c_1)p_n)$。

步骤 7(变异):变异操作采用均匀变异的方式,最终产生变异子代的数量 $c_3 = m - c_1 - c_2$。

步骤 8(生成子代个体):精英子代,交叉子代和变异子代的所有个体组成了新一代种群。利用式(2)的适应度函数评价新一代种群的适应性。

步骤 9:迭代步骤 4 ~ 步骤 8,直到满足停止条件。停止条件包含如下两类:①最大迭代次数达到设定值 *Generations*。②停滞代数达到设定值 *StallGenLimit*。

步骤 10:输出最优光谱特征子集 E。

传统的适应度函数一般只考虑回归模型中的交叉验证均方根误差和相关系数,提出的适应度函数综合考虑特征子集的回归性能和算法的运行效率。由式(2)可以得出,当随机森林的袋外误差取得最小值,并且选出的特征值数量最少时,本文算法设计的适应度函数取得最小值。

在变异操作算法中,首先生成符合某一范围内均匀分布的随机数集合(RD),每个随机数的值与染色体中每个基因座上的原有基因相关联。然后从左到右扫描染色体,将每个 RD 的值与变异算子 p_m 进行比较。如果位置 i 处的 RD 小于 p_m,则位置 i 处的基因被翻转;否则,该基因不被翻转。

2 实验

2.1 数据来源

在青岛市大沽河流域采集具有代表性的农田表层土壤样品 124 个,摊开并置于阴凉处,待自风干后,去除土样中的石子和动植物残体等明显杂物,用研钵研磨,再过 100 目筛子保存于自封袋中待测。准确称取 0.250 0 g 样品于 150℃下以“HNO_3 - HF - H_2O_2”消解体系和密封高压釜消解罐法进行消解,利用电感耦合等离子体发射光谱仪(ICP - OES, Optima 8000, PerkinElmer, USA)来检测 124 个样品的镉含量,并保留每份土样剩余部分用于高光谱数据采集。

2.2 光谱采集及预处理

本次实验选取 ASD FieldSpec 3 型便携式光谱仪测量土壤样品的光谱反射率。将土壤样品放入直径 10 cm、厚度约为 2 cm 的培养皿中。在 50 W 卤素灯作为光源的暗室中测量土壤样品的光谱,光源距样品 35 cm,天顶角为 30°,光谱仪探头安装在样品垂直上方 15 cm 处。每次实验开始之前和每测量 10 组样品之后使用具有 100% 反射率的标准化白板进行校正。每个土样测量时转动培养皿 3 次,每次转动 90°,每个方向上取样 10 次,4 个方向上共得到 40 条光谱测量值,算术平均后作为该样品的最终光谱反射率。

使用 Kennard - Stone[18] 算法从原始光谱数据集(包含 124 个样本)中选出 100 个样本作为校验集,剩余的 24 个样本作为预测集用于模型验证。剔除具有低信噪比的 350 ~ 399 nm、2 451 ~ 2 500 nm 两个边缘波段,在 400 ~ 2 450 nm 波段上对原始光谱数据采用多种预处理方法。基于随机森林回归模型,不同预处理方法所建模型的预测性能如表 1 所示。

由表 1 可得,一阶微分处理方法所建立的 RFR 模型结果最优,此时预测集相关系数、相对分析误差最大,分别为 0.897 3 和 1.26,均方根误差达到最小值 0.079 4。可见,一阶微分预处理最适于获取含量较少组分的光谱信息,与文献[19]研究结果一致。因此,后续特征选择与建模均在一阶微分分析基础上进行。

大沽河流域 124 个土壤样本的原始可见/近红外光谱曲线如图 2a 所示。从图 2a 可以看出,所有样本的反射率曲线特征基本相似。与文献[20]报

表 1　不同光谱预处理方法的随机森林建模结果

Tab. 1　Prediction results of RFR modeling by using different pre-processing methods

预处理方法	相关系数	均方根误差	预测相对分析误差
None	0. 622 9	0. 430 5	0. 22
D1	0. 897 3	0. 079 4	1. 26
D2	0. 445 7	0. 437 9	0. 41
SNV	0. 368 9	0. 422 0	0. 37
S－G	0. 351 5	0. 469 9	0. 47
D1 + SNV	0. 368 9	0. 467 6	0. 40
D2 + SNV	0. 406 6	0. 459 6	0. 44
S－G + SNV	0. 422 8	0. 419 7	0. 36
SNV + D1	0. 166 3	0. 485 7	0. 34
SNV + D2	0. 299 3	0. 453 4	0. 39
S－G + SNV + D1	0. 422 8	0. 419 7	0. 36
S－G + SNV + D2	0. 270 7	0. 474 3	0. 42

注：D1 表示对光谱矩阵求一阶导，D2 表示对光谱矩阵求二阶导，SNV（Standard normal variate transformation）表示对光谱矩阵进行标准正态变量变换，S－G（Savitzky－Golay）表示对光谱矩阵进行 Savitzky－Golay 卷积平滑。

道的一致，大多数土壤光谱在波长 1 400 nm、1 800 ~ 2 000 nm 和 2 200 ~ 2 400 nm 附近都会出现明显的水分吸收峰。图 2b 为最优预处理方式（一阶微分）后的光谱曲线。从图 2b 可以看出，一阶微分预处理不仅能够加强原始光谱的 3 个强吸收峰，而且能够明显增强原始光谱中 1 000 nm 和高于 2 200 nm 的弱吸收峰。这表明一阶微分预处理方法能够增强样品之间的光谱特征差异，较适应于农业土壤镉元素的高光谱响应。

3　结果与讨论

3.1　基于斯皮尔曼等级相关分析的特征选择算法结果分析

一般来说，特征值之间的相关性是通过相关系数来衡量的。由于斯皮尔曼等级相关系数对数据条件的要求没有皮尔逊相关系数严格，并且可以使用单调函数来描述变量之间的相关性，因此 SGA－RF 算法选用斯皮尔曼等级相关分析作为特征预选方法剔除冗余特征。对于初始的 2 051 个特征波段应用斯皮尔曼等级相关系数分析，在校验集上筛选的特征波长变量分布如图 3 所示。

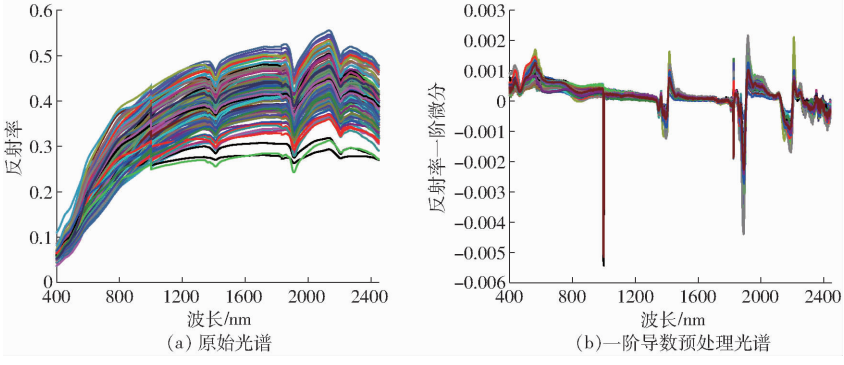

图 2　124 个土壤反射光谱

Fig. 2　Soil reflectance spectra of 124 soil samples

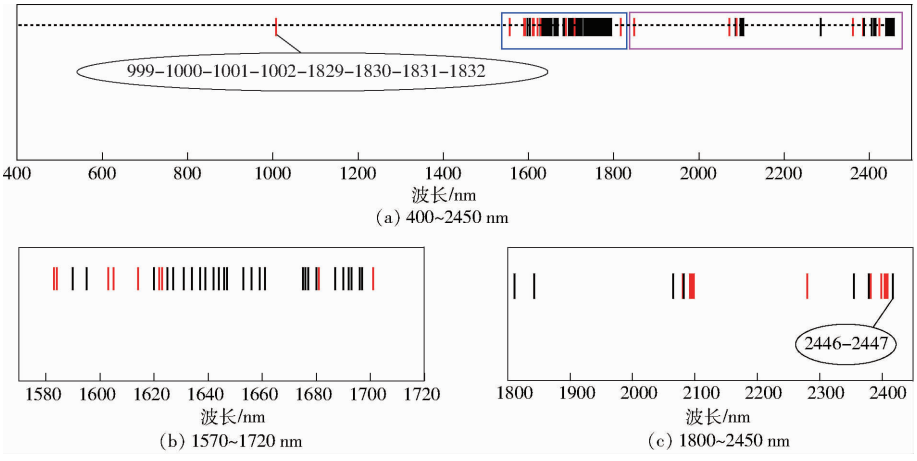

图 3　算法 1 筛选的波长变量分布图（不相邻波段之间的相关性用红色竖线表示）

Fig. 3　Distribution diagrams of spectral features selected by algorithm 1 (there was also a correlation between nonadjacent bands, which were shown on red long string)

算法 1 将原始全光谱变量从 2 051 个减少到 108 组特征波段，每组内部选出与镉浓度相关性最强的特征值为该组代表波段，所有代表波段在 400 ~ 2 450 nm 的分布如图 3a 所示。为了更加清晰地观

察不相邻波段之间是否存在相关性,将图 3a 中蓝色框内的局部波长变量放大于图 3b 中,将图 3a 中紫色框内的局部波长变量放大于图 3c 中。其中,对特征值的标注已在算法 1 中步骤 8 完成。在图 3c 中以 2 446 nm 为代表波段的这一组被标识为黑色,表明该组内部所有成员均为相邻波段。在图 3a 中以 999 nm 为代表波段的这一组被标识为红色,表明该组内部包含不相邻波段。以特征波段 999 nm 与 1 829 nm 的相关性为例,算法 1 判定二者具有相关性;但是采用传统遗传算法的区间分割法,二者不具有相关性,原因是 999 nm 和 1 829 nm 肯定不会被划分到同一个子区间内部。因此,传统遗传算法按照固定区间大小均分原始光谱,会导致部分潜在重要特征值的丢失。

从图 3a 可以看出,预选出的 108 个特征波段主要位于 1 550 ~ 1 850 nm 和 2 300 ~ 2 450 nm 的范围内。分析图 3b 可得,1 550 ~ 1 720 nm 范围内共有 9 组特征波段内部含有不相邻波段。由图 3c 可得,1 800 ~ 2 450 nm 范围内共有 7 组特征波段内部含有不相邻波段。108 组相关特征中共有 17 组内部存在不相邻波段之间具有相关性的现象。可见,相邻波段之间存在高相关性的概率较大,但是不相邻波段之间也有可能存在相关性。因此,基于斯皮尔曼等级相关分析的特征选择方法克服了传统遗传算法只对相邻波段特征值进行融合的局限性,采用一种更加合理的方法缩减遗传算法的自变量个数,使得遗传算法能够在冗余度最小且最具代表性的敏感波段中进行全局搜索,最终达到优化反演模型,提高预测能力的目的。

3.2 系数对适应度函数的影响

适应度函数的表达式(式(2))中包含一个权重参数 ω,取值范围为[0,1],该参数表示袋外误差在整个适应度函数中所占比重。为了找到适应度函数取最小值时所对应的权重参数 ω,需要考察不同权重对反演结果的影响。表 2 给出了不同权重参数 ω 所对应的适应度以及相对应的相关系数、均方根误差和预测相对分析误差。

从表 2 可以看出,当权重参数 ω 为 0.3,适应度函数取得最小值。因此,最优权重参数 ω 为 0.3。

3.3 特征选择结果分析

在算法 2 中,遗传算法种群规模被定义为 100,运行总代数为 200,基因位数设为 1。

SGA - RF 算法每一次迭代过程中,首先由 RFR 模型计算出每个个体的袋外误差,然后根据式(2)评估当前种群中每个个体的适应度,适应度越低的个体存活到下一代的概率越高。RFR 模型使用 100 个决策树来构建森林,并随机选取 $\rho = \sqrt{A}$ 作为分裂属性集(A 为属性个数)。基于上述参数设置,图 4 为 SGA - RF 算法的收敛过程。

表 2 不同权重参数 ω 所对应的适应度以及相对应的相关系数、均方根误差和预测相对分析误差

Tab. 2 Values of fitness function (f_i) and corresponding R, $RMSE$ and RPD according to several ω values

ω	适应度	相关系数	均方根误差	预测相对分析误差
0.1	0.021 59	0.898 1	0.073 0	1.93
0.2	0.028 00	0.908 8	0.084 2	1.32
0.3	0.020 15	0.950 2	0.060 1	2.03
0.4	0.020 86	0.915 0	0.080 1	1.45
0.5	0.023 14	0.909 9	0.078 9	1.86
0.6	0.024 51	0.928 3	0.070 6	1.98
0.7	0.022 98	0.935 7	0.071 3	1.88
0.8	0.256 4	0.921 6	0.082 2	1.76
0.9	0.279 1	0.889 4	0.086 6	1.45

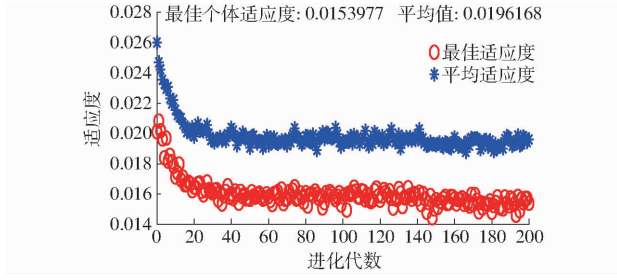

图 4 SGA - RF 算法特征选择过程适应度函数进化曲线

Fig. 4 Evolution curves of fitness function during feature selection process of SGA - RF

由图 4 可得,从第 109 代到第 200 代,适应度在 0.014 ~ 0.016 5 之间,变动幅度很小。最佳个体的适应度为 0.015 397 7,平均值约为 0.019 616 8。因此,可以判断 SGA - RF 算法没有出现过早收敛现象,原因是 SGA - RF 算法设置了正确有效的适应度,同时采用精英保留策略和锦标赛选择算法确保了该算法的必然收敛。

采用 SGA - RF 特征波段选择方法最终选出 37 个敏感波段,它们在 400 ~ 2 450 nm 波段的分布图如图 5 所示。图中黑色曲线为某个土壤样本的原始光谱曲线,红色空心圆圈表示特征波段所在位置。图 5b 和图 5c 分别放大了图 5a 中矩形框和正方形框内的敏感波段密集区域。

由图 5 可知,重金属镉的敏感波段主要位于 1 600 ~ 1 800 nm 和 2 350 ~ 2 450 nm 的范围内。已有研究表明,位于这 2 个范围的吸收峰主要与有机化合物中的 O—H、N—H 和 C—H 基团有关[21]。这可能与重金属镉的有机化合物吸附有关,吸附机理有待进一步深入研究。

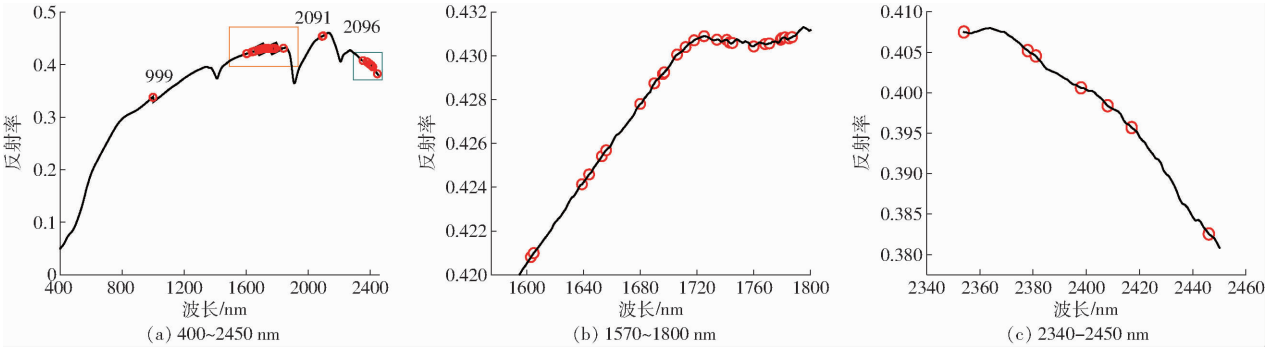

图 5　基于 SGA－RF 算法的镉可见/近红外光谱特征波段分布图

Fig. 5　Visible/NIR spectra distribution diagrams for soil Cr based on variables selected by SGA－RF (shown in red hollow circle markers)

3.4　与其他特征选择算法的比较与分析

为了验证算法 1，分别对不采用算法 1 的遗传随机森林算法（GA－RF）和采用算法 1 做预选的 SGA－RF 算法进行建模性能比较，结果如表 3 所示。两种算法均采用 RFR 建立回归模型。

表 3　评估算法 1 对本文算法预测效果的影响

Tab. 3　Impact of algorithm 1 on predictive effect by using SGA－RF

方法	选中特征值的数量	运行时间	相关系数	均方根误差	预测相对分析误差
SGA－RF	37	0.32 s	0.950 2	0.060 1	2.03
GA－RF	965	3 d	0.906 4	0.071 3	1.27

由表 3 可以看出，与 GA－RF 算法相比，SGA－RF 算法不仅建模时间短，特征值数量少，而且模型性能明显较高。因为算法 1 在有效剔除初始特征集冗余信息的同时能够保留与镉元素相关性最强的重要特征，大幅度减少特征精选阶段的输入特征值维数，不仅减少了计算复杂度，而且提高了模型反演能力。

为了进一步验证 SGA－RF 特征选择算法的有效性，本文算法与其他 3 类特征选择方法进行对比，结果如表 4 所示。这 3 类特征选择方法分别为：全光谱不做特征波段选择方法、单一特征选择方法和以 K 最近邻法构建适应度函数的遗传算法（SGA－KNN）。单一特征选择方法包含两种方法：原始遗传算法与偏最小二乘算法结合的特征波段选择方法（GA－PLSR）和斯皮尔曼等级相关分析（Spearman Rank）方法。这 3 类特征选择算法均采用 RFR 作为回归模型，并以选出特征波段的数量、建模时间、相关系数、均方根误差和预测相对分析误差作为模型的评价指标。

依据文献[21]，GA－PLSR 算法采用 2～20 nm 不同波段间隔对原始光谱进行重采样，每个子区间取平均数作为新的光谱自变量，以交叉验证均方根误差（RMSECV）最小值作为最优特征谱区筛选的标准，最终得到 12 nm 为最优光谱区间划分间隔。GA－PLSR 算法使用 PLS－Toolbox[22] 重复运行 50 次，以 PLSR 算法的 RMSECV 作为适应度函数。SGA－KNN 算法采用文献[23]中的基于 K 最近邻分类法（KNN）的适应度函数，其中参数 K 选取 3。上述 2 种算法的种群规模、运行总代数和 SGA－RF 算法保持一致。

表 4　本文特征选择算法与其他特征选择算法性能比较

Tab. 4　Comparison of performance between proposed and other feature selection methods

类别	变量筛选算法	特征数量	时间/s	校验集		
				相关系数	均方根误差	预测相对分析误差
全光谱	N/A	2 051	3.13	0.894 5	0.079 3	1.08
单一特征选择	GA－PLSR	38	0.34	0.539 0	0.129 4	0.65
	Spearman Rank	108	0.38	0.883 1	0.088 1	1.19
不同适应度函数	SGA－KNN	39	0.34	0.649 0	0.123 6	0.49
本文算法	SGA－RF	37	0.32	0.950 2	0.060 1	2.03

由表 4 可知，与其他 3 类特征选择算法相比，本文算法的建模时间最短，相关系数和预测相对分析误差最高，取得最佳预测效果。全光谱和斯皮尔曼等级相关分析建模的预测相对分析误差均在 1.4 以下，说明这 2 种方法建立土壤镉含量预测模型的性能较差，原因是全光谱中含有大量的冗余和无关信息；斯皮尔曼等级相关分析只能去除特征波段之间的冗余，无法剔除那些与待测组分无关的特征值，导

致模型精度较低。GA－PLSR 算法和 SGA－KNN 算法建模的预测相对分析误差均在 1.0 以下，说明以 GA－PLSR 算法和 SGA－KNN 算法作为特征波段选择方法无法建立土壤镉含量预测模型。原因为：①GA－PLSR 算法不仅忽略了不相邻波段之间的相关性，同时也没有考虑特征波段与待测组分之间的相关性，而且该算法用于评估每个染色体的 PLSR 模型为线性模型。②由于 KNN 算法对参数 K 的取值，距离的度量和多重共线性都非常敏感，导致该算法对特征的回归能力较差，并不适用于构造遗传算法的适应度函数。③SGA－RF 算法首先采用斯皮尔曼等级相关分析进行特征预选，去掉所有特征波段之间的冗余信息；其次，结合遗传算法的全局搜索能力和 RFR 的较高反演能力，优选出适应度最低的特征子集。其中，在适应性函数中采用非线性模型 RFR 评估每个个体的性能，而 RFR 对于参数设置与多重共线性问题均不敏感。

在 SGA－RF 算法特征筛选结果的基础上，以 RFR 作为回归模型，模型校验集和预测集样本的估测值和实测值之间的关系如图 6 所示。

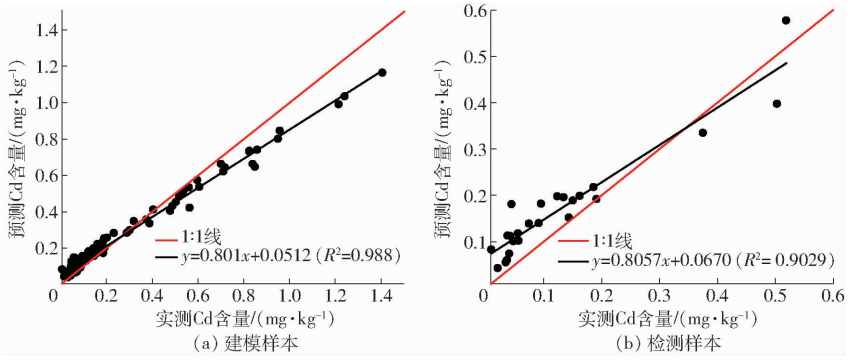

图 6 SGA－RF 算法结合 RFR 模型建模样本、检验样本的实测值与估测值比较

Fig. 6 Comparison of measured Cd contents and estimated values of modeling and testing samples through SGA－RF algorithm with RFR model

由图 6 可得，采用 SGA－RF 特征选择算法在校验集和预测集均取得了较好的反演效果，验证了该算法与随机森林回归模型相结合能够有效解决农业土壤中痕量级重金属含量的反演问题。

4 结论

（1）基于大沽河流域农业土壤样本真实数据集，SGA－RF 算法能够将 2 051 个原始特征波段优选至 37 个。以选出的 37 个敏感波段作为自变量建立大沽河流域农业土壤镉高光谱反演模型，该模型具有较低的预测均方根误差（0.060 1），较高的相关系数（0.950 2）和预测相对分析误差（2.03）。建模结果证实了该算法在降低计算复杂度的基础上提高了土壤镉含量反演模型的精度。

（2）斯皮尔曼等级相关分析将原始 2 051 波段缩减到 108 组特征子集，通过分析每组特征子集的组内相关性，发现在不相邻波段之间也存在相关性。说明基于斯皮尔曼等级的特征预选方法能够克服传统遗传算法的局限性，最大限度地缩减遗传算法的自变量个数，是 SGA－RF 算法的重要组成部分。

（3）以相关系数、均方根误差和预测分析相对误差作为评价标准，本文算法分别与其他 4 类特征选择方法进行了对比。比较结果说明，SGA－RF 算法可以有效剔除冗余波长变量，提取最具代表性的特征波长变量，减少建模时间，提高预测效果。

（4）SGA－RF 算法有效解决了土壤重金属镉高光谱反演领域建模时间长，预测精度差的问题，可用于其他农业土壤的重金属镉污染监测与识别。

参 考 文 献

1 REHMAN Z U, KHAN S, BRUSSEAU M L, et al. Lead and cadmium contamination and exposure risk assessment via consumption of vegetables grown in agricultural soils of five-selected regions of Pakistan [J]. Chemosphere, 2017, 168:1589－1596.

2 LUCE M S, ZIADI N, GAGNON B, et al. Visible near infrared reflectance spectroscopy prediction of soil heavy metal concentrations in paper mill biosolid- and liming by-product-amended agricultural soils[J]. Geoderma, 2017, 288:23－36.

3 MARIN A, ANDRADES M, INIGO V, et al. Lead and cadmium in soils of La Rioja Vineyards, Spain: pb and cd in Vineyards [J]. Land Degradation and Development, 2016, 27(4):1286－1294.

4 SUN Weichao, ZHANG Xia, ZOU Bin, et al. Exploring the potential of spectral classification in estimation of soil contaminant elements [J]. Remote Sensing, 2017, 9(6):632.

5 SANTAN F B D, SOUZA A M D, POPPI R J, et al. Visible and near infrared spectroscopy coupled to random forest to quantify

some soil quality parameters [J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2018, 191:454 – 462.

6 XIE Xianli, PAN Xianzhang, SUN Bo. Visible and near-infrared diffuse reflectance spectroscopy for prediction of soil properties near a copper smelter [J]. Pedosphere, 2012, 22(3):351 – 366.

7 王敬哲,塔西甫拉提·特依拜,张东.基于分数阶微分的荒漠土壤铬含量高光谱检测[J/OL]. 农业机械学报,2017,48(5):152 – 158. http://www.jcsam.org/jcsam/ch/reader/view_abstract.aspx? file_no=20170518&flag=1. DOI:10.6041/j.issn.1000-1298.2017.05.018.
WANG Jingzhe, TASHPOLAT·Tiyip, ZHANG Dong. Spectral detection of chromium content in desert soil based on fractional differential[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(5):152 – 158. (in Chinese)

8 VAHID K, FARAMARZDOULATI A, SAEED Y, et al. Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods[J]. Geoderma,2018, 318:29 – 41.

9 FRANK R, MICHAEL D, INGO M, et al. Prediction of soil parameters using the spectral range between 350 and 15 000 nm: a case study based on the permanent soil monitoring program in Saxony, Germany[J]. Geoderma,2018, 315: 188 – 198.

10 CHEN Tao, CHANG Qingrui, CLEVERS J G P W, et al. Rapid identification of soil cadmium pollution risk at regional scale based on visible and near-infrared spectroscopy[J]. Environmental Pollution, 2015,206: 217 – 226.

11 GHOLIZADEH A, BORŮVKA L, SABERIOON M M, et al. Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features[J]. Soil and Water Research, 2015, 10(4): 218 – 227.

12 TAYEBI M, NADERI M, MOHAMMADI J, et al. Comparing different statistical models for assessing Fe- contaminated soils based on VNIR/SWIR spectral data[J]. Environmental Earth Sciences, 2017,76(21):1 – 15.

13 SHI Tiezhu, LIU Huizeng, CHEN Yiyun, et al. Spectroscopic diagnosis of arsenic contamination in agricultural soils [J]. Sensors, 2017,17(5):1036.

14 JIANG Qinghu, LIU Minxia, WANG Jun, et al. Feasibility of using visible and near-infrared reflectance spectroscopy to monitor heavy metal contaminants in urban lake sediment [J]. Catena,2018,162:72 – 79.

15 XIA Xueqi, MAO Yuqian, JI Yufeng, et al. Assessment of cadmium contamination in the sediments of Changjiang(Yangtze) River by reflectance spectroscopy[J]. Environmental Science & Technology, 2007,41(10):3449 – 3454.

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

17 RAMMAL A, PERRIN E, VRABIE V, et al. Selection of discriminant mid-infrared wavenumbers by combining a naïve Bayesian classifier and a genetic algorithm: application to the evaluation of lignocellulosic biomass biodegradation [J]. Mathematical Biosciences, 2017,289:153 – 161.

18 KENNARD R W, STONE L A. Computer aided design of experiments [J]. Technometrics, 1969, 11: 137 – 148.

19 STENBERGA B, ROSSEL R A V, MOUAZEN A M, et al. Visible and near infrared spectroscopy in soil science[J]. Advances in Agronomy, 2010, 107:163 – 215.

20 BELLON-MAUREL V, MCBRATNEY A. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils-critical review and research perspectives [J]. Soil Biology and Biochemistry, 2011, 43: 1398 – 1410.

21 SHI Tiezhu, CHEN Yiyun, LIU Huizeng. Soil organic carbon content estimation with laboratory-based visible-near-infrared reflectance spectroscopy: feature selection [J]. Applied Spectroscopy,2014,68(8): 831 – 837.

22 Eigenvector research incorporated. Chemometrics Software[EB/OL][2018 – 02 – 03]. http://www.eigenvector.com.

23 OLULEYE B, LEISA A, LENG J, et al. Zernike moments and genetic algorithm: tutorial and application [J]. British Journal of Mathematics & Computer Science, 2014, 4(15): 2217 – 2236.