

基于高光谱反演的复垦区土壤重金属含量经验模型优选

陈元鹏¹ 张世文² 罗 明¹ 鄢文聚¹ 鞠正山¹ 李少帅¹

(1. 自然资源部国土整治中心, 北京 100035; 2. 安徽理工大学地球与环境学院, 淮南 232001)

摘要:以工矿复垦区为实验区域,基于 ASD FieldSpec 4 高光谱遥感数据,结合实测的土壤重金属含量,利用回归分析与特征选择方法,开展了基于高光谱数据的土壤重金属含量反演研究与实验并进行了经验模型优选。通过对光谱曲线进行一阶微分、对数一阶微分以及对数倒数的一阶微分等数学变换有效提高了光谱数据与土壤重金属含量的相关性。在此基础上采用偏最小二乘回归(Partial least squares regression, PLSR)、随机森林回归(Random forest regression, RFR)、支持向量机回归(Support vector machine regression, SVMR)3种回归分析模型开展土壤重金属含量反演实验,结果表明偏最小二乘回归(PLSR)对研究区内土壤中重金属含量的反演最为有效,尤其对区域内主要障碍因子镉(Cd)元素含量的反演效果最佳,验证集决定系数 R^2 为 0.76。基于粒子群算法(Particle swarm optimization, PSO)、遗传算法(Genetic algorithm, GA)、Relief F 算法 3 种特征选择方法对偏最小二乘回归(PLSR)模型进行优化,结果表明粒子群算法(PSO)可有效降低特征波段变量维度,进一步提高模型反演精度,使决定系数 R^2 由 0.76 提高至 0.84。综上,基于高光谱数据,采用偏最小二乘回归(PLSR)与粒子群算法(PSO)相结合的方法,可有效对工矿复垦区土壤中的重金属含量进行测度,可为复垦区土地的质量和生态指标监测提供理论方法和技术支持。

关键词: 工矿复垦区; 土壤重金属; 高光谱反演; 经验模型; 偏最小二乘回归; 粒子群算法

中图分类号: X833 **文献标识码:** A **文章编号:** 1000-1298(2019)01-0170-10

Empirical Model Optimization of Hyperspectral Inversion of Heavy Metal Content in Reclamation Area

CHEN Yuanpeng¹ ZHANG Shiwen² LUO Ming¹ YUN Wenju¹ JU Zhengshan¹ LI Shaoshuai¹

(1. Land Consolidation and Rehabilitation Center, Ministry of Natural Resources, Beijing 100035, China

2. College of Earth and Environmental Science, Anhui University of Science and Technology, Huainan 232001, China)

Abstract: Taking industrial and mining reclamation land as the research object, based on the ASD FieldSpec 4 hyperspectral remote sensing data, combined with the field survey data of soil heavy metal attributes, using regression analysis and feature selection methods, the retrieval research and experiment of soil heavy metal content based on hyperspectral data were carried out, and the selection and comparison of empirical models were conducted. The correlation between soil heavy metal concentration and spectral data was effectively improved by the first derivative and logarithmic reciprocal etc. On this basis, three regression analysis models, including partial least squares regression (PLSR), random forest regression (RFR) and support vector machine regression (SVMR) were used to carry out the inversion experiment of heavy metal content in soil. The results showed that the partial least squares regression (PLSR) had the highest precision for the retrieval of heavy metal concentration in the reclaimed soil, especially for the cadmium (Cd) concentration, which was the main obstacle factor in the area. The determination coefficient (R^2) of fit for the set was 0.76. Particle swarm optimization (PSO), genetic algorithm (GA) and Relief F were used to optimize the partial least squares regression (PLSR) model. The results indicated that PSO can effectively reduce the dimension of characteristic band variables and further improve the model inversion. And the R^2 of fit was increased from 0.76 to 0.84. In conclusion,

收稿日期: 2018-10-18 修回日期: 2018-11-27

基金项目: 国家重点研发计划项目(2018YFD0800701)和土地整治重点实验室开放课题(2018-KF-02)

作者简介: 陈元鹏(1988—),男,高级工程师,博士,主要从事遥感技术在土地利用与复垦中的应用研究,E-mail: cyp520918@163.com

通信作者: 罗明(1962—),女,研究员,博士,主要从事土地整治技术政策和矿地复垦标准制定研究,E-mail: luoming@mail.lcrc.org.cn

based on hyperspectral data, the combination of partial least squares regression (PLSR) and particle swarm optimization (PSO) can effectively measure the concentration of heavy metals in the soil of industrial and mining reclamation area, and it can provide theoretical methods and technical support for the detection of land quality and ecological indicators in the reclamation area.

Key words: mining reclamation area; soil heavy metal; hyperspectral inversion; empirical model; partial least squares regression; particle swarm algorithm

0 引言

作为地球生命的重要载体、国土资源的基本元素,土壤不仅是农业生产的根基,也是国家粮食安全的重要保证,同时是水资源、生态环境优化的重要基础^[1]。加强国土资源数量、质量、生态“三位一体”监测与管护,对土壤进行快速准确的调查识别至关重要。然而,传统土壤调查、评价、测量方法成本高、效率低,无法实现大范围的快速检测^[2],与之相比,光谱技术快速、便捷,对土地无污染、无破坏,优势明显,仅利用单一光谱便可获取多种土壤理化属性信息,同时还可以在野外直接进行原位光谱采集^[3-4],节省大量的人财物力,增强了大范围土壤属性数据测度的技术可行性,是土壤属性信息获取的重要技术手段。而光谱检测中的高光谱遥感技术更具优势,其光谱分辨率高、波段连续性强,能获取连续的地物光谱曲线,使地物信息更加精细地体现在光谱曲线上,实现更准确的反演和监测。随着技术的不断进步发展,高光谱遥感目前已被广泛应用于土壤资源调查、分类、评价、制图等工作中^[5]。

《土地复垦条例》(国务院令第 592 号)明确规定“复垦土地需要开展五年监测”。但基于传统方法的长期跟踪监测需要消耗大量的人财物力,为改变这一现状迫切需要引入高光谱遥感技术,以更加高效、便捷、成本节约的优势,对土地复垦项目开展土壤属性数据的检测和监测,以新理论、新技术来突破传统方法的弊端,为土壤调查监测提供一个新的研究和方向。现阶段,高光谱数据已能够准确反映地物光谱的细微特征,变换后的光谱在消除背景和噪声影响,放大提取光谱的吸收和反射特征等方面具有较优效果^[6],对挖掘光谱信息,构建精度高、稳健性好、泛化能力强的模型具有重要作用^[7-10]。不少学者利用高光谱遥感数据针对不同区域的 As、Cr、Cd、Cu、Fe、Zn、Ni 等重金属含量进行建模反演,获得了较好的反演精度。但目前利用高光谱数据反演工矿复垦区土壤重金属含量的研究相对较少,且现有研究中对于反演模型、特征选择方法的研究也有待进一步深入。

本文基于 ASD FieldSpec 4 高光谱遥感数据,结合土壤重金属属性数据,利用机器学习方法包括线

性(偏最小二乘回归)与非线性(随机森林与支持向量机)回归方法,开展基于高光谱数据的土壤重金属含量反演研究与实验,采用 3 种特征选择方法包括粒子群算法、遗传算法、Relief F 算法优化经验模型,以减少模型自变量数量并提高建模精度。

1 研究区概况

研究区为历史遗留工矿废弃地复垦项目区,位于四川省泸州市古蔺县,地理坐标为 27°59′43″~28°7′24″N,105°55′41″~106°4′47″E,区域内海拔 340~1 207 m,中亚热带气候,年平均气温 17.1~18.5℃、平均降雨量 748.4~1 184.2 mm^[11-13]。该项目分为 5 个复垦区,项目区域东西宽 3.4 km,南北长 2.8 km,总面积 297 hm²。研究区地理区位及遥感影像数据如图 1 所示。

2 数据获取与分析

2.1 土壤样本数据采集

在研究区域内,综合考虑复垦方向和复垦措施,采用网格布点、分层抽样,共采集 42 个土壤样品,实际有效采集了表层土壤样品 41 个,剖面土壤样品 10 个,样点分布如图 2 所示。

土壤样品采集时间为 2017 年 8 月。表层土壤采样深度为 0~25 cm。剖面土壤采集 0~25 cm、25~50 cm、50~75 cm 3 个层次,采集时间在上茬作物玉米成熟或收获以后,下茬作物尚未施用底肥和种植以前,该时间段能够反映采样地块的真实养分状况和供肥能力。野外采样采用铁锹挖采样坑,挖完坑后,用竹片去除与金属采样器接触的土壤,再采集样品。

在采样点的布设上,以 GPS 定位点为中心,向四周辐射确定 3~5 个分样点,组合成一个混合样。每个分样点的采土部位、深度均一致。采样地块为长方形时,采用“S”形布设分样点;采样地块近似正方形时,采用“X”形或棋盘形布设分样点。采样时均避开了沟渠、林带、田埂、路边、旧房基、粪堆及微地形高低不平无代表性地段。采集各分样点时将土壤掰碎,挑出根系、秸秆、石块、虫体等杂物,充分混合后,四分法留取 1.0~1.5 kg 装入样品袋。

土壤样品在日光下干燥。在样品干燥过程中采

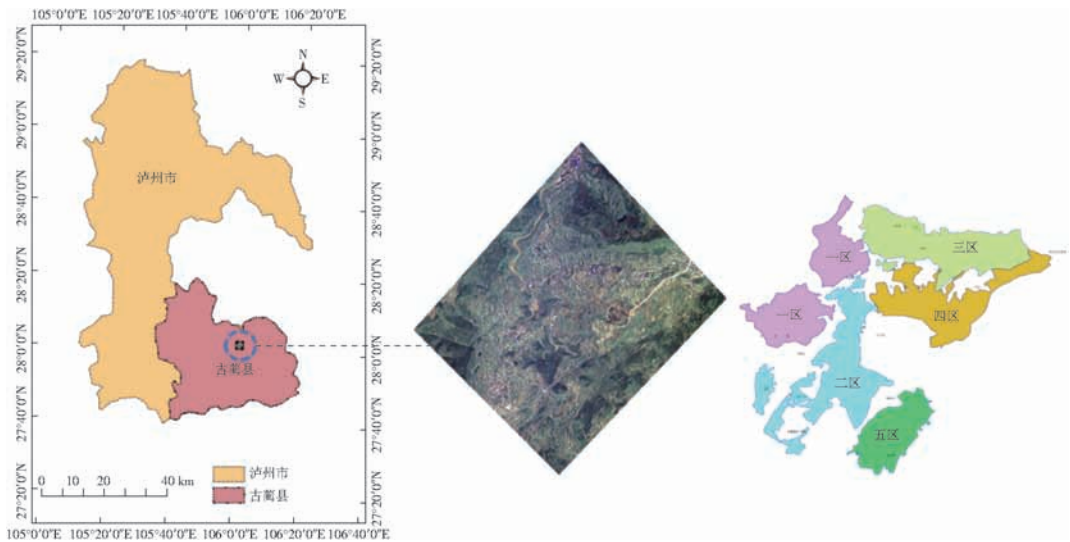


图1 研究项目区地理位置和遥感影像

Fig.1 Location of study area and remote sensing image

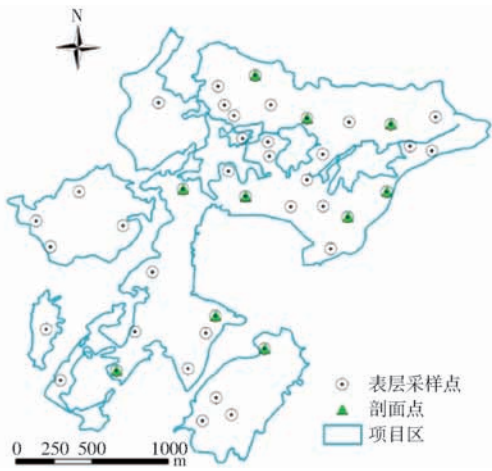


图2 样点分布图

Fig.2 Map of samples distribution

用揉搓法,以免胶结。干燥后的样品在加工(过筛)前均用木槌轻轻敲打,以便使土壤样品恢复至自然粒级状态。

2.2 样本土壤理化指标分析

所有土壤样本经风干、研磨并通过 2 mm 孔径筛,采用四分法分成 2 份,分别供化学分析和光谱测试用。采用内标法定量处理样土数据,结果经仪器校对、回收率校正,土壤理化性质的测试分析方法见表 1,表层土壤样品测定结果见表 2、3。

表1 土壤样品分析方法检出限

元素	方法	检出限/($\mu\text{g}\cdot\text{g}^{-1}$)
镉(Cd)	等离子体质谱法	0.5
铬(Cr)	等离子体质谱法	0.5
砷(As)	原子荧光法	0.01
镍(Ni)	火焰原子吸收法	1.10
pH 值	离子计	0.001

表2 表层土壤元素含量特征值

指标	均值/ ($\text{mg}\cdot\text{kg}^{-1}$)	标准差/ ($\text{mg}\cdot\text{kg}^{-1}$)	变异系 数/%	区域背景值/ ($\text{mg}\cdot\text{kg}^{-1}$)	富集 系数
镉(Cd)	1.05	0.97	92.39	0.19	5.52
铬(Cr)	165.63	54.93	33.16	82.38	2.01
砷(As)	12.85	5.57	43.35	7.06	1.82
镍(Ni)	64.45	16.26	25.23	22.64	2.84
pH 值	6.55	1.11	16.95		

表3 项目区土壤本底元素含量特征值

指标	均值/ ($\text{mg}\cdot\text{kg}^{-1}$)	标准差/ ($\text{mg}\cdot\text{kg}^{-1}$)	变异系 数/%	区域背景值/ ($\text{mg}\cdot\text{kg}^{-1}$)	富集 系数
镉(Cd)	0.36	0.30	83.33	0.19	1.89
铬(Cr)	87.65	23.12	26.38	82.38	1.06
砷(As)	6.91	3.17	45.88	7.06	0.98
镍(Ni)	33.45	6.86	20.51	22.64	1.48
pH 值	6.09	1.87	30.71		

由表 2、3 可知：①复垦区土壤环境指标中重金属元素镉(Cd)、铬(Cr)、砷(As)、镍(Ni)有明显的富集,镉(Cd)、铬(Cr)、砷(As)、镍(Ni)含量分别是区域背景值的 5.52、2.01、1.82、2.84 倍,其均值含量均是区域背景值的 1.5 倍以上,特别是镉(Cd),高达 5 倍。与区域土壤背景值相比,项目所在地区土壤本底元素镉(Cd)、铬(Cr)、砷(As)、镍(Ni)含量分别是区域背景值的 1.89、1.06、0.98、1.48 倍,其中砷(As)含量没有超过区域背景值。②复垦区内监测指标的变异系数均较高,镉(Cd)、铬(Cr)、砷(As)、镍(Ni)变异系数分别是 92.39%、33.16%、43.35%、25.23%,特别是镉(Cd)变异系数最大。这主要是先前采矿活动对环境扰动较大,不确定性

因素多,致使镉(Cd)分布发生突变。而项目区土壤本底元素中也是镉(Cd)的变异系数最大,为 83.33%,而铬(Cr)、砷(As)、镍(Ni)变异系数分别是 26.38%、45.88%、20.51%。

2.3 样本土壤光谱数据获取

土壤光谱数据测量采用美国 ASD 公司生产的 ASD FieldSpec 4 地物光谱仪。在室内条件下用标准白板和卤素光源获取土壤样本的反射光谱曲线,光谱波段范围为 350 ~ 2 500 nm。该光谱仪在不同的波段间有不同的间隔和分辨率,其中 350 ~ 1 000 nm 之间的采样间隔为 1.4 nm,光谱分辨率为 3 nm,1 000 ~ 2 500 nm 范围内的采样间隔为 2 nm,采样光谱分辨率为 10 nm。经过间隔为 1 nm 光谱重采样,两者的输出波段数均为 2 151。由于光谱很容易受到外界光线的干扰,因此测试选择在无光的暗室中进行,选择 50 W 的卤素灯为光源、25°裸光纤镜头接收反射波段。测量时将样本放入直径为 100 mm、高 2 mm 的透明玻璃培养皿中,光源距离样品 50 cm,光线与样品呈 15°,探头距样本 10 cm 位于光源对面,探头光纤末端位于土壤样本正上方。为保证测量的准确性,以及最大限度的提高仪器的优化性能,在测量前,先开机预热 30 min,再按照暗电流(DC)采集、仪器优化(OPT)、白板校正(WR)顺序对仪器进行调整和校准。为防止测量过程中有其他因素影响光谱信息收集的准确性,每隔 10 个样本进行一次白板校正。在光谱采样过程中由于光谱波段两侧产生不稳定的噪声区,为此,实验去除了 350 ~ 399 nm 和 2 351 ~ 2 500 nm 范围内的波段数据,获得 1 950 个波段数据。

2.4 光谱数据预处理方法

光谱特性由土壤性质决定,但土壤是一种复杂的混合物,具有极其复杂的多孔体系,由不同的矿物质、水分、气体和土壤有机质组成,影响土壤光谱反射特性的主要因素包括土壤有机质、重金属元素含量、含水率以及土壤质地和母质等。对本次研究采集的 41 个土壤样本光谱反射率测定数据进行制图,利用 Savitzky-Golay(S-G)卷积平滑法^[14]以 10 为窗口平滑处理,结果如图 3 所示。

由图 3 可知,虽然土壤样本光谱反射率有所不同,但整体趋势呈现共同特征,反射率介于 0.2 ~ 0.8 之间,各样本土壤光谱曲线在波长范围内呈相似波动性,光谱反射率随波长增加而增大,在 400 ~ 780 nm 可见光波段范围内反射率增加较快,在近红外的 780 ~ 2 350 nm 波段范围内反射率增加减缓,800 ~ 1 400 nm 范围内的近红外光谱区的光谱反射率高于可见光波段,在 1 400 ~ 1 500 nm、1 900 ~

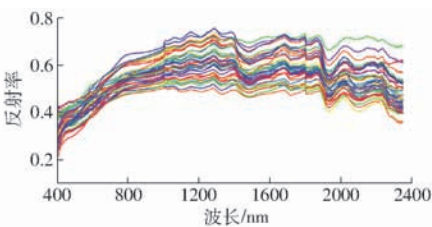


图 3 土壤样本光谱曲线

Fig. 3 Spectracurves of soil samples

2 000 nm和 2 200 ~ 2 300 nm 3 个范围内存在明显的水分吸收区间。综上,光谱数据的波形、波动性、吸收峰等特征与已有的研究结论相符,表明光谱数据的采集情况良好。

为更好地反映土壤重金属含量与光谱反射率之间的对应关系,消除噪声干扰,分离重叠样本,更加准确地寻找特征波段,对光谱曲线进行一阶微分变换、对数一阶微分变换以及对数倒数的一阶微分变换,结果如图 4 ~ 6 所示。由图可知,经过一阶微分变换后的光谱,有效地消除了基线和背景的干扰,提高了部分波段的分辨率和灵敏度,使重叠样本得到分离,进一步提取了原始数据中差异性不显著的光谱信息,使光谱特征波段峰值更加明显。

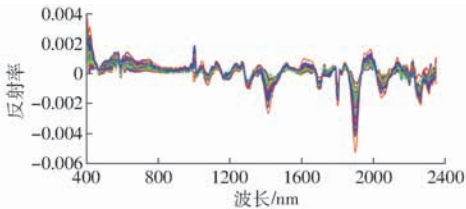


图 4 土壤样本点光谱一阶微分变换

Fig. 4 First order differential transformation of soil sample spectra

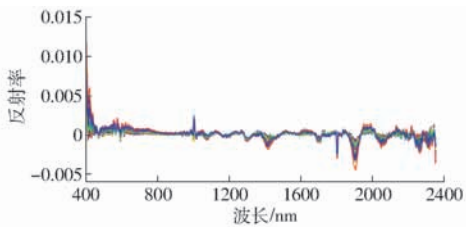


图 5 土壤样本点光谱对数一阶微分变换

Fig. 5 Logarithmic first order differential transformation of soil sample spectra

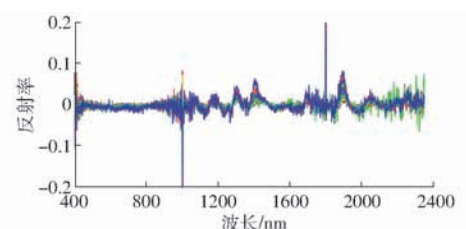


图 6 土壤样本点光谱对数倒数一阶微分变换

Fig. 6 Logarithmic inverse first order differential transformation of soil sample spectra

2.5 土壤光谱相关性分析

为进一步分析土壤光谱反射率与重金属元素之间的关联,探究土壤重金属含量与原始、一阶微分变换、对数一阶微分变换以及对数倒数一阶微分变换光谱数据的相关性,运用相关系数法对其进行分析,相关系数计算公式为

$$r_i = \frac{\sum_{n=1}^N (R_{ni} - \bar{R}_i)(y - \bar{y})}{\sqrt{\sum_{n=1}^N (R_{ni} - \bar{R}_i)^2 \sum_{n=1}^N (y - \bar{y})^2}} \quad (1)$$

式中 r_i ——土壤重金属元素含量与高光谱反射率的相关系数

R_{ni} ——第 n 个土壤样本第 i 波段的光谱反射率

$\bar{R}_i$ ——土壤样本在 i 波段反射率的均值

y ——第 n 个土壤样本重金属含量

$\bar{y}$ ——土壤样本重金属含量均值

表 4 为土壤重金属元素与光谱最大相关系数

绝对值和对应波段。由表 4 可知,只经过 S-G 平滑处理的光谱曲线相关系数绝对值在 0.16~0.28 之间;一阶微分变换后的光谱曲线相关系数绝对值在 0.36~0.47 之间;对数一阶微分变换后的光谱曲线相关系数绝对值在 0.40~0.55 之间;对数倒数一阶微分变换后的光谱曲线相关系数绝对值在 0.46~0.54 之间。土壤重金属元素含量与光谱曲线的相关性逐步增强,说明土壤光谱经过预处理变换后,均能在一定程度上消除背景因素或系统噪声等影响,对于提高建模精度能够起到一定作用。

相关系数绝对值中,不低于 0.5 的情况分别有:光谱对数一阶微分变换下的镉(Cd)元素相关系数、光谱对数倒数一阶微分变换下的铬(Cr)元素相关系数、光谱对数倒数一阶微分变换下的镍(Ni)元素相关系数。所以本文选择对数一阶微分、对数倒数一阶微分变换后的光谱进行建模。

表 4 土壤重金属与光谱最大相关系数绝对值和对应波段

Tab.4 Absolute value of maximum correlation coefficient between heavy metal elements and spectra and corresponding band

土壤重金属	S-G 平滑		一阶微分		对数一阶微分		对数倒数一阶微分	
	最大值	波长/nm	最大值	波长/nm	最大值	波长/nm	最大值	波长/nm
砷(As)	0.28	2 347	0.36	460	0.46	1 052/2 302	0.46	2 221
镉(Cd)	0.16	2 317	0.47	1 038	0.55	1 038	0.47	2 303
铬(Cr)	0.28	1 016	0.38	1 828	0.41	472/1 718	0.50	406
镍(Ni)	0.26	1 401	0.36	466	0.40	907/2 008	0.54	899

3 模型选择与评价

利用 41 个土壤样点光谱的对数一阶微分、对数倒数一阶微分变换后光谱数据进行重金属的反演建模,分别选择线性和非线性回归模型开展建模实验。

3.1 线性回归模型

偏最小二乘回归(Partial least squares regression, PLSR)集成了主成分、典型相关和线性回归分析方法的优点,其提供一种多对多的线性回归建模的方法。PLSR 是将因子分析和回归分析结合的方法。

(1)因子分析。将 X 和 Y 作如下分解

$$X = TP^T + E$$

$$Y = UQ^T + F$$

式中 X ——样本光谱矩阵

Y ——样本待检测属性矩阵

T —— X 的得分矩阵

U —— Y 的得分矩阵

P —— X 的载荷(即主成分矩阵)

Q —— Y 的载荷(即主成分矩阵)

E ——用 PLSR 模型拟合 X 时所引进的残差

矩阵

F ——用 PLSR 模型拟合 Y 时所引进的残差矩阵

(2)回归分析。将 T 和 U 作线性回归,得

$$U = TB$$

$$Y_1 = T_1BQ = X_1P^TBQ$$

式中 B ——关联系数矩阵

预测中,由未知样品矩阵 X_1 和校正得到的 P ,求出未知样品 X_1 的矩阵 T_1 。

3.2 非线性回归模型

3.2.1 随机森林回归

随机森林(Random forest regression, RFR)是一种基于决策树的高效机器学习算法,可用于对样本进行分类,也可用于回归分析。它属于非线性分类器,因此可挖掘变量之间复杂的非线性相互依赖关系^[15-19]。

3.2.2 支持向量机回归

支持向量机(Support vector machine regression, SVMR)是一种在分类和回归问题中应用较为广泛的模型,在分类和回归问题中均能得到较优效果,支持向量机在回归分析中,目标是在有限的噪声样本

基础上预测出未知的连续函数。支持向量机回归包括线性回归和非线性回归。对于非线性回归,只需要引入核函数,通过非线性映射将输入空间映射到高维的特征空间,在高维空间上进行线性回归。常用的核函数有:径向基核函数(RBF)、Sigmoid 核函数与多项式核函数等。本文选用 RBF 作为 SVM 的核函数^[20-27]。

3.3 光谱特征选择方法

基于高光谱数据进行回归分析,通常自变量较多,过多的样本数量或存在冗余信息导致建模精度降低,为此需要利用特征选择算法进行数据降维,减少自变量个数的同时保证建模精度不降低。为进一步降低数据维度、减少冗余信息、提高运算效率、提升模型稳定性与适应性,本文选取 3 种光谱特征选择方法来优化模型预测能力,实现模型预测准确性和稳定性的进一步提高^[28-30]。

3.3.1 粒子群算法

粒子群优化算法(Particle swarm optimization, PSO)模拟的是鸟群觅食行为,是一种基于集群智能的随机寻优算法,鸟群通过自身经验和种群之间的交流调整自己的搜寻路径,从而找到食物最多的地点。PSO 算法步骤如下:初始化一个规模为 m 的粒子群,设定初始位置和速度;计算每个粒子的适应值;对每个粒子,将其适应值和其经历过的最好位置的适应值进行比较,若较好,则将其作为当前最好位置;分别对粒子的速度和位置进行更新,以达到满足终止条件^[31-33]。PSO 算法涉及的参数包括:种群数量、迭代次数、惯性权重、学习因子、空间维数、位置限制和速度限制等。

3.3.2 遗传算法

遗传算法(Genetic algorithm, GA)是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型,是一种通过模拟自然进化过程搜索最优解的方法^[34-36]。GA 从一组随机产生的“种群”(初始可行解)开始搜索过程。其中可行解即被称为“染色体”,一个可行解一般由多个元素构成,那么每一个元素就被称为染色体上的一个“基因”。在算法执行过程中,染色体通常在编码后进行迭代计算,每次迭代都会生成若干条染色体。采用适应度函数分别计算每一条染色体的适应程度,并根据适应程度计算每一条染色体在下次进化中被选中的概率,对染色体进行优胜劣汰。在未达到算法终止条件时,通过选择、交叉和变异 3 个算子产生下一代染色体。重复此过程直到算法收敛从而得到最优解(最优染色体)。

3.3.3 Relief F 算法

Relief F 算法是一种考虑了多变量和特征互相依赖作用的过滤式(Filter)特征选择算法^[37],算法通过“假设间隔”对特征的分类能力进行评价,综合考虑类间距与类内距,若类间距大于类内距,则增加其权值,若类间距小于类内距,则降低其权值,通过类间距与类内距不断更新其权值,并根据计算的最终权值进行特征选择,权值更新公式为

$$W_f^i = W_f^{i-1} + \frac{1}{n} \left[\frac{\sum_{c \neq class(x_i)} diff_f(x_i, M(x_i))}{(n-1)m_{class(x_i)}} - \frac{diff_f(x_i, H(x_i))}{m_{class(x_i)}} \right]$$

(6)

假设间隔 θ 是指当保持样本类别不变时分类决策面可移动的最大距离,即

$$\theta = \frac{1}{2} (\|x - M(x)\| - \|x - H(x)\|)$$

(7)

式中 $diff()$ ——不同样本间的距离
 n ——样本数量 f ——评价特征
 $m_{class(x_i)}$ ——样本抽样次数
 i ——随机抽中的样本
 $H(x)、M(x)$ ——样本 x 的同类和异类最近邻样本点

4 结果分析

4.1 线性拟合结果

线性模型选择偏最小二乘回归(PLSR)模型,将样本按照 3:1 的比例划分进行建模验证(建模样本 30 个,验证样本 11 个),回归分析结果如表 5 所示,表中决定系数 R^2 和均方根误差(RMSE)均为验证集的评价系数。其中镉(Cd)元素与对数一阶微分变换后的光谱曲线建模拟合效果最好,决定系数 R^2 最大,达 0.76,均方根误差 RMSE 最小,为 0.62 mg/kg,说明基于室内处理的光谱在预测土壤中的重金属元素镉(Cd)时能够达到良好效果;而重金属元素镍(Ni)在与经对数一阶微分变换的光谱曲线建模拟合中,最优决定系数 R^2 仅为 0.29,均方根误差 RMSE 为 10.6 mg/kg,在与对数倒数一阶微分变换的光谱曲线建模拟合中 R^2 则更低,说明在研究区域中,高光谱数据无法满足精确建模预测土壤重金属元素镍(Ni)含量的基本要求。元素砷(As)、铬(Cr)与对数倒数一阶微分变换后的光谱建模,决定系数 R^2 分别为 0.54、0.65,但均方根误差 RMSE 差别较大,分别为 1.94、16.1 mg/kg,虽然决定系数 R^2 基本能够满足预测精度,但铬(Cr)的 RMSE 为 16.1 mg/kg,说明模型稳定性和泛化能力有待进一步增强。

表 5 PLSR 模型反演结果评价

Tab. 5 Assessment of PLSR inversion results				
金属元素	最优光谱变换方法	主成分数量	R^2	RMSE/ ($\text{mg}\cdot\text{kg}^{-1}$)
砷(As)	对数倒数一阶微分	15	0.54	1.94
镉(Cd)	对数一阶微分	15	0.76	0.62
铬(Cr)	对数倒数一阶微分	15	0.65	16.1
镍(Ni)	对数一阶微分	15	0.29	10.6

4.2 非线性拟合结果

非线性模型选择随机森林(RFR)和支持向量机(SVMR)模型,同样将样本按照3:1的比例划分进行建模验证(建模样本30个,验证样本11个),回归分析结果如表6、7所示。由表6、7可知,利用随机森林和支持向量机方法的反演结果中,依然是镉(Cd)元素的决定系数 R^2 最高,镍(Ni)元素的 R^2 最低,均方根误差RMSE中镉(Cd)元素分别为1.69、1.75 mg/kg,铬(Cr)元素分别为18.20、18.70 mg/kg,结果呈现的整体趋势与PLSR反演结果较为一致。与PLSR相比较,非线性模型对各重金属元素拟合精度偏低,究其原因或是因为建模样本数量过少,导致非线性建模效果不够理想,在大样本容量下,预测精度和模型的稳定性将会有所提高。

表 6 RFR 模型反演结果评价

Tab. 6 Assessment of RFR inversion results			
金属元素	最优光谱变换方法	R^2	RMSE/($\text{mg}\cdot\text{kg}^{-1}$)
砷(As)	对数倒数一阶微分	0.37	6.54
镉(Cd)	对数一阶微分	0.44	1.69
铬(Cr)	对数倒数一阶微分	0.35	18.20
镍(Ni)	对数一阶微分	0.17	12.60

表 7 SVMR 模型反演结果评价

Tab. 7 Assessment of SVMR inversion results			
金属元素	最优光谱变换方法	R^2	RMSE/($\text{mg}\cdot\text{kg}^{-1}$)
砷(As)	对数倒数一阶微分	0.40	7.87
镉(Cd)	对数一阶微分	0.48	1.75
铬(Cr)	对数倒数一阶微分	0.33	18.70
镍(Ni)	对数一阶微分	0.25	13.65

4.3 基于特征选择的模型优化

由线性(PLSR)和非线性(RFR、SVMR)模型分

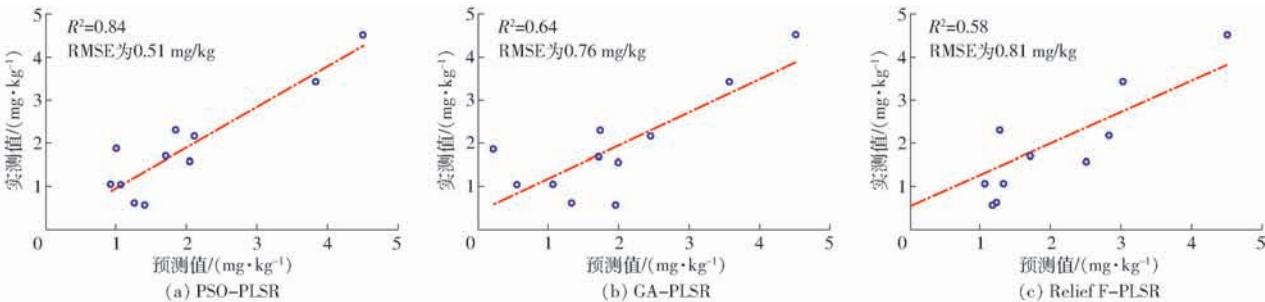


图 7 基于不同特征选择的 Cd 含量反演结果

Fig. 7 Cd content inversion results based on feature selection

析结果可知,在 PLSR 模型下镉(Cd)元素与对数一阶微分变换后的光谱曲线建模拟合效果最优,精度最高,决定系数 R^2 最大,达到0.76,所以选择PLSR模型下镉(Cd)元素的反演作为优化目标,以 PSO、GA、Relief F 3 种特征选择方法对高光谱数据进行特征波段选择后再进行 PLSR 建模,以验证特征选择、模型优化的效果。

PSO 方法参数设置:初始种群数量20、迭代次数20、惯性权重0.33;GA 方法参数设置:初始种群数量20,最大繁殖代数20、交叉概率0.6、变异概率0.03。基于以上两种方法选择的特征波段数量趋于一致,分别为535、537个。同时基于Relief F 方法,选择了535个特征变量。

以 PSO、GA、Relief F 3 种方法选择的特征波段为自变量,基于 PLSR 模型对镉(Cd)元素进行回归分析,反演结果如表8所示。由表8可知,基于 PSO 特征选择的反演结果最优, R^2 为0.84;其次为基于 GA 的特征选择的反演结果, R^2 为0.64;基于 Relief F 特征选择的反演结果 R^2 只有0.58。与未经过特征选择的反演结果比较,只有基于 PSO 特征选择方法的反演精度有所提高, R^2 由0.76提高到0.84,基于 GA、Relief F 特征选择方法的反演精度均出现了一定程度的降低, R^2 由0.76分别降至0.64和0.58。所以,在3种方法中,基于 PSO 优化的 PLSR 模型能够进一步提高对实验区内镉(Cd)元素反演的精度,并有效减少了特征波段的数量。

表 8 基于不同特征选择方法的 PLSR 模型反演结果评价

Tab. 8 Assessment of PLSR inversion results based on feature selection			
特征选择方法	变量数量	R^2	RMSE/($\text{mg}\cdot\text{kg}^{-1}$)
-	1 950	0.76	0.62
PSO	535	0.84	0.51
GA	537	0.64	0.76
Relief F	535	0.58	0.81

图7为镉(Cd)元素反演模型的预测值散点图与拟合结果。

5 结论

(1)以工矿复垦区为实验区域,进行了基于高光谱数据的土壤重金属含量反演研究和经验模型优选。实验结果表明,通过对光谱曲线进行一阶微分、对数一阶微分以及对数倒数的一阶微分等数学变换能够有效提高光谱数据与土壤重金属含量的相关性;针对实验区数据,PLSR对复垦区土壤重金属含量的反演精度最优,尤其对区域内主要障碍因子镉(Cd)元素含量的反演效果最佳,验证集的决定系数

R^2 达到0.76;采用的PSO、GA、Relief F 3种特征选择方法中,PSO在降低特征波段变量维度的基础上,能够有效提高PLSR模型对复垦区土壤重金属含量反演的精度,使决定系数 R^2 由0.76提至0.84。

(2)运用线性(PLSR)和非线性(RFR、SVMR)机器学习方法开展建模实验,采用基于随机搜索策略的3种特征选择方法(PSO、GA、Relief F)进行模型优化,通过对3种机器学习方法和3种特征选择方法分析优选,得到基于PSO-PLSR的土壤重金属含量高光谱反演方法,可以获得较佳建模精度。

参 考 文 献

- [1] 姜庆虎. 基于高光谱数据的滨湖土壤组分信息反演建模及优化[D]. 武汉:武汉大学,2014.
- [2] 滕靖,何政伟,倪忠云,等. 西范坪矿区土壤铜元素的高光谱响应与反演模型研究[J]. 光谱学与光谱分析,2016,36(11):3637-3642.
TENG Jing, HE Zhengwei, NI Zhongyun, et al. Spectral response and inversion models for prediction of total copper content in soil of Xifanping mining area[J]. Spectroscopy and Spectral Analysis, 2016, 36(11): 3637-3642. (in Chinese)
- [3] 纪文君,史舟,周清,等. 几种不同类型土壤的VIS-NIR光谱特性及有机质响应波段[J]. 红外与毫米波学报,2012,31(3):277-282.
JI Wenjun, SHI Zhou, ZHOU Qing, et al. VIS-NIR reflectance spectroscopy of the organic matter in several types of soils[J]. Journal of Infrared and Millimeter Waves, 2012, 31(3): 277-282. (in Chinese)
- [4] ROSSEL R A V, CATTLE S R, ORTEGA A, et al. In situ measurements of soil colour, mineral composition and clay content by VIS-NIR spectroscopy[J]. Geoderma, 2009, 150(3-4): 253-266.
- [5] 李曦. 基于高光谱遥感的土壤有机质预测建模研究[D]. 杭州:浙江大学,2013.
- [6] ZHANG Chao, SU Jinghua. Soil mapping via diffuse reflectance spectroscopy based on variable indicators: an ordered predictor selection approach[J]. Geoderma, 2018, 314: 146-159.
- [7] 张贤龙,张飞,张海威,等. 基于光谱变换的高光谱指数土壤盐分反演模型优选[J]. 农业工程学报,2018,34(1):110-117.
ZHANG Xianlong, ZHANG Fei, ZHANG Haiwei, et al. Optimization of soil salt inversion model based on spectral transformation from hyperspectral index[J]. Transactions of the CSAE, 2018, 34(1): 110-117. (in Chinese)
- [8] DENNISON P E, ROBERTS D A. Endmember selection for multiple endmember spectral mixture analysis using endmember average RMSE[J]. Remote Sensing of Environment, 2003, 87(2): 123-135.
- [9] 马孟莉,朱艳,李文龙,等. 基于分层多端元混合像元分解的水稻面积信息提取[J]. 农业工程学报,2012,28(2):154-159.
MA Mengli, ZHU Yan, LI Wenlong, et al. Extracting area information of paddy rice based on stratified multiple endmember spectral mixture analysis[J]. Transactions of the CSAE, 2012, 28(2): 154-159. (in Chinese)
- [10] 缪春华,张显峰,刘羽. 基于多端元光谱分解的干旱区植被覆盖度遥感反演[J]. 应用生态学报,2012,23(12):3243-3249.
MIU Chunhua, ZHANG Xianfeng, LIU Yu. Remote sensing retrieval of vegetation coverage in arid areas based on multiple endmember spectral unmixing[J]. Chinese Journal of Applied Ecology, 2012, 23(12): 3243-3249. (in Chinese)
- [11] 张玉芳,庞艳梅,刘琰琰,等. 近50年四川省水稻生产潜力变化特征分析[J]. 中国生态农业学报,2014,22(7):813-820.
ZHANG Yufang, PANG Yanmei, LIU Yanyan, et al. Potential productivity of rice in Sichuan Province in recent five decades[J]. Chinese Journal of Eco-Agriculture, 2014, 22(7): 813-820. (in Chinese)
- [12] 孙园园,徐富贤,孙永健,等. 四川稻作区优质稻生产气候生态条件适宜性评价及空间分布[J]. 中国生态农业学报,2015,23(4):506-513.
SUN Yuanyuan, XU Fuxian, SUN Yongjian, et al. Suitability evaluation of eco-climatic conditions for high quality rice production in Sichuan Province[J]. Chinese Journal of Eco-Agriculture, 2015, 23(4): 506-513. (in Chinese)
- [13] 肖科. 泸州市土地整理效益分析与评价[D]. 雅安:四川农业大学,2009.
- [14] 赵安新,汤晓君,张钟华,等. 优化Savitzky-Golay滤波器的参数及其在傅里叶变换红外气体光谱数据平滑预处理中的应用[J]. 光谱学与光谱分析,2016,36(5):1340-1344.
ZHAO Anxin, Tang Xiaojun, ZHANG Zhonghua, et al. Optimizing Savitzky-Golay parameters and its smoothing pretreatment for FTIR gas spectra[J]. Spectroscopy and Spectral Analysis, 2016, 36(5): 1340-1344. (in Chinese)
- [15] LIU Jiantao, FENG Quanlong, GONG Jianhua, et al. Land-cover classification of the Yellow River Delta wetland based on multiple endmember spectral mixture analysis and a random forest classifier[J]. International Journal of Remote Sensing, 2016, 37(8): 1845-1867.

- [16] 赵玉,王红,张珍珍. 基于遥感光谱和空间变量随机森林的黄河三角洲刺槐林健康等级分类[J]. 遥感技术与应用, 2016,31(2):359-367.
ZHAO Yu, WANG Hong, ZHANG Zhenzhen. Forest healthy classification of *Robinia pseudoacacia* in the Yellow River Delta, China based on spectral and spatial remote sensing variables using random forest[J]. Remote Sensing Technology and Application, 2016, 31(2): 359-367. (in Chinese)
- [17] VAHID E, SAEID H, AHMAD M, Y, et al. Land cover mapping based on random forest classification of multitemporal spectral and thermal images[J]. Environmental Monitoring and Assessment, 2015, 187(5): 167-175.
- [18] GHOSH A, SHARMA R, JOSHI P K. Random forest classification of urban landscape using Landsat archive and ancillary data: combining seasonal maps with decision level fusion[J]. Applied Geography, 2014, 48: 31-41.
- [19] 林楠,姜琦刚,杨佳佳,等. 基于资源一号 02C 高分辨率数据的农业区土地利用分类[J/OL]. 农业机械学报,2015, 46(1):278-284. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20150139&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2015.01.039.
LIN Nan, JIANG Qigang, YANG Jiajia, et al. Classifications of agricultural land use based on high-spatial ZY1-02C remote sensing images[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(1): 278-284. (in Chinese)
- [20] 王兴玲,李占斌. 基于网格搜索的支持向量机核函数参数的确定[J]. 中国海洋大学学报,2005,35(5):859-862.
WANG Xingling, LI Zhanbin. Identifying the parameters of the kernel function in support vector machines based on the grid-search method[J]. Periodical of Ocean University of China, 2005, 35(5): 859-862. (in Chinese)
- [21] 刘颖. 基于机器学习的遥感影像分类方法研究[M]. 北京:清华大学出版社,2014.
- [22] 张建华,孔繁涛,李哲敏,等. 基于最优二叉树支持向量机的蜜柚叶部病害识别[J]. 农业工程学报,2014,30(19):222-231.
ZHANG Jianhua, KONG Fantao, LI Zheming, et al. Recognition of honey pomelo leaf diseases based on optimal binary tree support vector machine[J]. Transactions of the CSAE, 2014, 30(19): 222-231. (in Chinese)
- [23] 许吉仁,董零红,杨源灏,等. 基于支持向量机的矿区复垦农田土壤-小麦镉含量高光谱估算[J]. 光子学报,2014, 43(5):0530001.
XU Jiren, DONG Jihong, YANG Yuanxuan, et al. Support vector machine model for predicting the cadmium concentration of soil-wheat system in mine reclamation farmland using hyperspectral data[J]. Acta Photonica Sinica, 2014, 43(5): 0530001. (in Chinese)
- [24] 丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报,2011,40(1):2-10.
DING Shifei, QI Bingjuan, TAN Hongyan. An overview on theory and algorithm of support vector machines[J]. Journal of University of Electronic Science and Technology of China, 2011, 40(1): 2-10. (in Chinese)
- [25] KWOK J T Y. Support vector mixture for classification and regression problems[C]//Proceedings of 14th International Conference on Pattern Recognition, 1998.
- [26] 梁栋,管青松,黄文江,等. 基于支持向量机回归的冬小麦叶面积指数遥感反演[J]. 农业工程学报,2013,29(7):117-123.
LIANG Dong, GUAN Qingsong, HUANG Wenjiang, et al. Remote sensing inversion of leaf area index based on support vector machine regression in winter wheat[J]. Transactions of the CSAE, 2013, 29(7): 117-123. (in Chinese)
- [27] MAXWELL A E, WARNER T A. Differentiating mine-reclaimed grasslands from spectrally similar land cover using terrain variables and object-based machine learning classification[J]. International Journal of Remote Sensing, 2015, 36(17): 4384-4410.
- [28] 叶勤,姜雪芹,李西灿,等. 基于高光谱数据的土壤有机质含量反演模型比较[J/OL]. 农业机械学报,2017,48(3): 164-172. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20170321&journal_id=jcsam. DOI: 10.6041/j.issn.1000-1298.2017.03.021.
YE Qin, JIANG Xueqin, LI Xican, et al. Comparison on inversion model of soil organic matter content based on hyperspectral data[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(3): 164-172. (in Chinese)
- [29] 曹引,冶运涛,赵红莉,等. 基于离散粒子群和偏最小二乘的水源地浊度高光谱反演[J/OL]. 农业机械学报,2018, 49(1):173-182. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20180122&journal_id=jcsam. DOI: 10.6041/j.issn.1000-1298.2018.01.022.
CAO Yin, YE Yuntao, ZHAO Hongli, et al. Satellite hyperspectral retrieval of turbidity for water source based on discrete particle swarm and partial least squares[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(1): 173-182. (in Chinese)
- [30] 任建强,吴尚蓉,刘斌,等. 基于 Hyperion 高光谱影像的冬小麦地上干生物量反演[J/OL]. 农业机械学报,2018, 49(4):199-211. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20180423&journal_id=jcsam. DOI: 10.6041/j.issn.1000-1298.2018.04.023.
REN Jianqiang, WU Shangrong, LIU Bin, et al. Retrieving winter wheat above-ground dry biomass based on Hyperion hyperspectral imagery[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(4): 199-211. (in Chinese)
- [31] ZHANG H L, SONG L L. Parameter identification in chaotic systems by means of quantum particle swarm optimization[J]. Acta Physica Sinica, 2013, 62(19): 750-754.

[32] HYMAN C, PALIINO R L. Possible role of the reticuloendothelial system in protein transport [J]. *Annals of the New York Academy of Sciences*, 2010, 88(1): 232 – 239.

[33] HUANG C, ZHANG D, SONG G. A novel mapping algorithm for three-dimensional network on chip based on quantum-behaved particleswarm optimization[J]. *Frontiers of Computer Science*, 2017, 11(4): 1 – 10.

[34] 陈永明,林萍,何勇. 基于遗传算法的近红外光谱橄榄油产地鉴别方法研究[J]. *光谱学与光谱分析*,2009,29(3):671 – 674. CHEN Yongming, LIN Ping, HE Yong. Study on discrimination of producing area of olive oil using near infrared spectra based on genetic algorithms[J]. *Spectroscopy and Spectral Analysis*, 2009, 29(3): 671 – 674. (in Chinese)

[35] 章海亮,罗微,刘雪梅,等. 应用遗传算法结合连续投影算法近红外光谱检测土壤有机质研究[J]. *光谱学与光谱分析*, 2017,37(2):584 – 587. ZHANG Hailiang, LUO Wei, LIU Xuemei, et al. Measurement of soil organic matter with near infrared spectroscopy combined with genetic algorithm and successive projection algorithm[J]. *Spectroscopy and Spectral Analysis*, 2017, 37(2): 584 – 587. (in Chinese)

[36] 廖阔,付建胜,杨万麟. 改进的 Relief F 算法用于雷达距离像目标识别[J]. *电子测量与仪器学报*,2010,24(9):831 – 836. LIAO Kuo, FU Jiansheng, YANG Wanlin. Modified Relief F algorithm for radar HRRP target recognition [J]. *Journal of Electronic Measurement and Instrument*, 2010, 24(9): 831 – 836. (in Chinese)

[37] 何涛,胡洁,夏鹏,等. 基于 Relief F 算法与遗传算法的肌电信号特征选择[J]. *上海交通大学学报*,2016,50(2):204 – 208. HE Tao, HU Jie, XIA Peng, et al. Feature selection of emg signal based on Relief F algorithm and genetic algorithm [J]. *Journal of Shanghai Jiaotong University*, 2016, 50(2): 204 – 208. (in Chinese)

~~~~~

(上接第 201 页)

[28] 王增丽,冯浩,余坤. 基于主成分分析的不同预处理麦秸改良土壤效果评价[J/OL]. *农业机械学报*, 2016,47(10): 150 – 155. [http://www.j-csam.org/jcsam/ch/reader/view\\_abstract.aspx? flag = 1&file\\_no = 20161020&journal\\_id = jcsam](http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20161020&journal_id=jcsam). DOI:10.6041/j.issn.1000-1298.2016.10.020. WANG Zengli, FENG Hao, YU Kun. Effect evaluation of different pretreated wheat straw utilizations on soil improvement based on principal component analysis [J/OL]. *Transactions of the Chinese Society for Agricultural Machinery*, 2016, 47(10):150 – 155. (in Chinese)

[29] 唐秀美,刘玉,刘新卫,等. 基于格网尺度的区域生态系统服务价值估算与分析[J/OL]. *农业机械学报*, 2017,48(4): 149 – 153. [http://www.j-csam.org/jcsam/ch/reader/view\\_abstract.aspx? flag = 1&file\\_no = 20170419&journal\\_id = jcsam](http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20170419&journal_id=jcsam). DOI:10.6041/j.issn.1000-1298.2017.04.019. TANG Xiumei, LIU Yu, LIU Xinwei, et al. Estimation and analysis of ecosystem service value based on grid scale [J/OL]. *Transactions of the Chinese Society for Agricultural Machinery*, 2017,48(4):149 – 153. (in Chinese)

[30] 崔文举,舒清态,刘满宾,等. 西双版纳热带林森林景观稳定性研究[J]. *云南地理环境研究*, 2010, 22(2):29 – 33. CUI Wenju, SHU Qingtai, LIU Manbin, et al. Study on stability of tropical forest landscape in Xishuangbanna [J]. *Yunnan Geographic Environment Research*, 2010, 22(2):29 – 33. (in Chinese)

[31] EETVEL V V, ANTROP M. A stepwise multi-scaled landscape typology and characterisation for trans-regional integration, applied on the federal state of Belgium[J]. *Landscape & Urban Planning*, 2009, 91(3):160 – 170.