中国农业科学 2020,53(21):4449-4459
Scientia Agricultura Sinica doi: 10.3864/j.issn.0578-1752.2020.21.013

开放科学（资源服务）标识码（OSID）：

基于数据挖掘技术的高光谱土壤质地分类研究

钟亮，郭熙，国佳欣，韩逸，朱青，熊杏

（江西农业大学国土资源与环境学院/江西省鄱阳湖流域农业资源与生态重点实验室，南昌 330045）

摘要：【目的】寻找红壤地区不同土壤质地类型的 Vis-NIR 光谱反射规律，通过光谱对土壤质地类别进行快速、准确的预测。【方法】以江西省奉新县北部为研究区，245 个土壤样本为研究对象，在国际制土壤质地 4 组和 12 级两种分类标准下，首先分析不同土壤质地类型的光谱反射率，然后采用 9 种数学变换方法和 5 种机器学习算法相互组合的数据挖掘模型，进行土壤质地的分类研究，最后对建模准确度最高的混淆矩阵和预测结果三角坐标分布图进行分析。【结果】（1）不同土壤质地之间的光谱反射率存在较多的交叉重叠现象，土壤质地与光谱反射率之间的规律较为复杂；（2）分数阶导数变换是整数阶导数的扩展，有助于土壤质地的分类，但原始光谱数据具有更加丰富的特征信息，更适合进行土壤质地分类建模；（3）在对非均衡数据集建模时，集成学习方法和神经网络方法都是不错的选择；（4）较难通过模型去区分土壤质地分界线附近的类别，其中在 4 组分类标准下最容易被预测错误成黏壤土组，在 12 级分类标准下最容易被预测错误成黏壤土和壤质黏土这两种土壤质地类型；（5）在 4 组分类标准中，进行归一化处理和 MLP 模型组合取得了 0.68 的最高预测准确度，其中黏壤土组的预测准确度能达到 0.84；再细分到 12 级分类后，分类效果最佳的组合来自于原始数据和 MLP 模型，其中壤质黏土分类准确度达到了 0.89。【结论】本研究结果可为南方红壤地区通过高光谱数据进行土壤质地分类提供参考依据。

关键词：红壤区；可见光近红外光谱；土壤质地；分类；数据挖掘技术

Soil Texture Classification of Hyperspectral Based on Data Mining Technology

ZHONG Liang, GUO Xi, GUO JiaXin, HAN Yi, ZHU Qing, XIONG Xing

(*College of Land Resources and Environment, Jiangxi Agricultural University/Key Laboratory of Poyang Lake Watershed Agricultural Resources and Ecology of Jiangxi Province, Nanchang 330045*)

Abstract:【Objective】 The aim of this study was to find the reflection law of Vis-NIR spectra of different soil texture types in red soil region, and to quickly and accurately predict the soil texture type by the spectrum. 【Method】 Taking the north of Fengxin County in Jiangxi Province as the research area, 245 soil samples were taken as the research objects. Under the 4 groups and 12 levels of international soil texture classification standards, the spectral reflectance of different soil texture types was analyzed first, then the data mining models combining 9 mathematical transformation methods and 5 machine learning algorithms were used to classify the soil texture, and finally analysis of the confusion matrix with the highest modeling accuracy and the triangular coordinate distribution map of prediction results. 【Result】 (1) There were many overlaps and overlaps in the spectral reflectance between different soil textures, and the law between the soil texture and the spectral reflectance was more complicated. (2) Fractional derivative transformation was an extension of the integer derivative, which was helpful for the classification of soil texture, but the original spectral data had more abundant feature information and was more suitable for the classification of soil texture. (3) Both

收稿日期：2020-02-22；**接受日期：**2020-03-18
基金项目：国家自然科学基金项目（41361049）、国家重点研发计划项目（2017YFD0301603）
联系方式：钟亮，E-mail：zhongliang1007@163.com。通信作者郭熙，E-mail：xig435@163.com

ensemble learning methods and neural network methods were good choices when modeling unbalanced data sets. (4) It was difficult to distinguish the categories near the boundary of soil texture by using the model. Among them, clay loam group was the most likely to be predicted wrongly under the four classification standards, and clay loam and loamy clay were the two most likely to be predicted wrongly under the 12 classification standards. (5) Among the four groups of classification standards, the highest prediction accuracy (at 0.68) was obtained by the combination of normalization treatment and MLP model, and the prediction accuracy of clay loam group could reach 0.84. After subdivision to 12 levels classification, the best classification result came from combination of original data and MLP model, and the classification accuracy of loamy clay was 0.89. 【Conclusion】 The results of this study could provide a reference for soil texture classification by using hyperspectral data.

Key words: red soil region; Vis-NIR spectroscopy; soil texture; classification; data mining technology

0 引言

【研究意义】土壤质地是土壤重要的物理性质之一，它与土壤保肥能力、保水状况、通气性及耕作的难易程度有着密切关系[1]。不同的土壤质地往往具有明显不同的农业生产性状，了解土壤的质地类型，对农业生产具有指导价值[2]。传统测定土壤质地的方法有比重计法、激光粒度仪法、吸管法和密度计法[3]，这些方法耗时耗力，容易出现人为误差，且无法解释区域土壤质地的确定问题[4]。近年来，随着光谱技术的发展，为快速获取土壤属性信息提供了新的途径[5]。土壤高光谱技术以其光谱分辨率高和波段信息丰富的特点，在估测土壤特性上具有强大的优势[6]，可节省大量的人力物力，在精准农业、数字土壤制图、土壤资源遥感调查等工作中起到至关重要的作用[7]。【前人研究进展】目前国内外分别有学者利用遥感影像[8-10]、土壤图像[11-13]、环境因子[1,14-17]和便携式 X 射线荧光光谱[18]进行土壤质地的预测研究，同时都表现出较好的精度。然而，众多学者利用光谱数据预测土壤质地时，现有的研究中大部分是进行土壤粒径的回归建模[19-23]，将得到的各粒径百分含量再推测出土壤质地的类别[4]，这样很难保证单独预测到的三种粒径含量总和为 100%[24-25]，不利于土壤质地类别的推测。因此，用光谱数据直接进行土壤质地的分类建模显得更加直接和准确[26-27]；同时，大部分的研究是寻找特征波段进行建模[5,28-29]，基于全谱建模的研究相对较少[30]；另外，在对光谱数据进行预处理时，常采用整数阶导数变换[31-32]，也有研究使用分数阶导数变换[33-34]。如今，数据挖掘技术因其能处理高维度数据，并且能够快速、准确地建立土壤属性与光谱反射率之间的关系模型，广泛应用在光谱与土壤属性的建模研究中[35-36]。【本研究切入点】以江西省奉新县北部为本研究区，245 个红壤样本为研究对象，在国际制土壤质地 4 组和 12 级两种分类标准下[3]，采用包含分数阶导数在内的 9 种数学变换方法以及 SVM、RF、MLP 等 5 种机器学习算法相互组合的数据挖掘模型，利用 Vis-NIR 光谱进行土壤质地分类的研究。【拟解决的关键问题】以明确高光谱数据预测红壤地区土壤质地类型的建模能力，并且寻找最优数学变换和机器学习算法的组合模型，以期为南方红壤地区通过高光谱数据进行土壤质地分类提供参考依据。

1 材料与方法

1.1 研究区概况与土壤样本采集

研究区位于江西省奉新县北部，总面积约 20 000 hm^2，坐标 115°03′—115°23′ E，28°40′—28°47′ N，属中亚热带湿润气候，年平均降雨量 1 612 mm，年平均气温 17.3℃，海拔介于 31—133 m 之间。研究区土地利用类型包括耕地、园地、林地和其他用地，分别占整个研究区面积的 25%、5%、55%和 15%。土壤类型主要为红壤。

土样的采集时间为 2018 年 7 月 23 日至 8 月 11 日，为了保证数据的代表性，将研究区划分为 1 km×1 km 的规则网格，在各网格内随机选取采样点，并且充分考虑地理环境因素，对个别网格进行采样点加密。在深度为 0—30 cm 内通过 5 点混合法进行土壤样本的采集，均匀混合后得到最终样本。采样点使用手持 GPS 仪器获取并记录位置，分布如图 1 所示，在研究区内共采集了 245 个样本，其中耕地 97 个、林地 92 个、园地 56 个。将采回的样本于实验室自然风干、研磨后过 2 mm 筛，并将其均匀分成两部分，分别用于土壤质地和土壤光谱测定。土壤质地采用激光衍射法（Beckman Coulter LS230，USA，测试粒径范围 0.04—2 000 μm）测定[37-39]。本研究采用国际制土壤质地分类标准，把土壤质地划分为 4 组 12 级，如图 2 所示，在国际制土壤质地分类三角坐标图中，3 个坐标轴分别为砂粒（2—0.02 mm）、粉粒（0.02—0.002 mm）、黏粒（＜0.002 mm）。

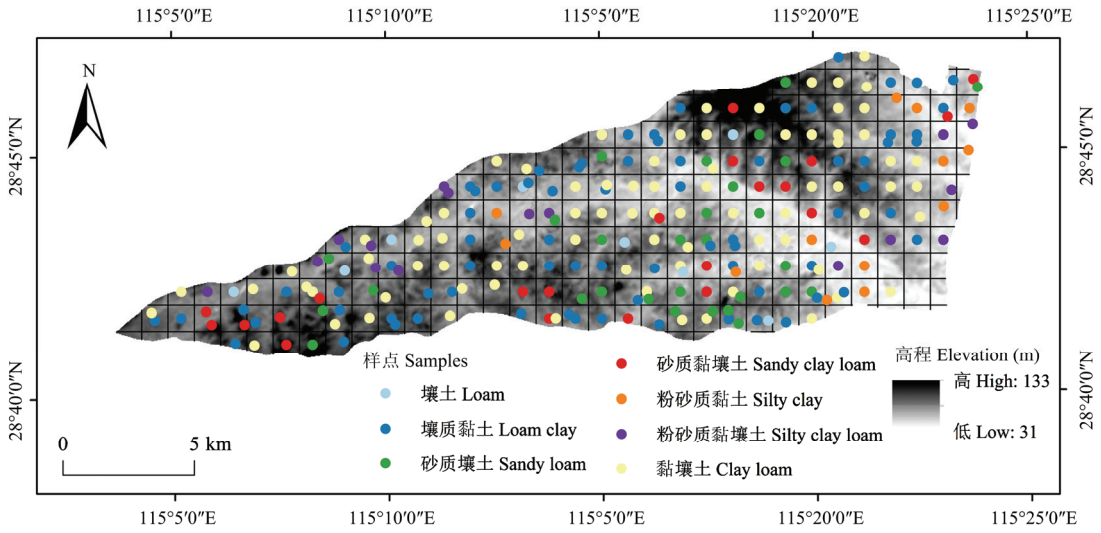

图 1　研究区采样点分布示意图

Fig. 1　Distribution diagram of sampling points in the study area

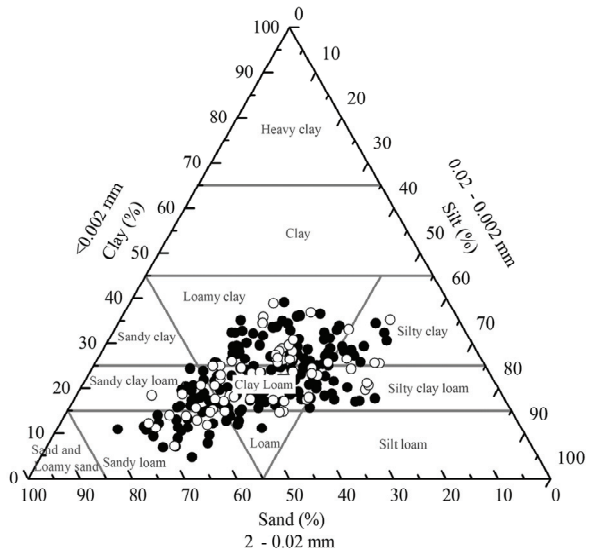

图 2　国际制土壤质地分类和土壤样本示意图

Fig. 2　Schematic diagram of international soil texture classification and soil samples

1.2　光谱数据采集与预处理

光谱测量采用美国 ASD 公司的 FieldSpec4 地物光谱仪，波长范围为 350—2 500 nm，输出得到 2 151 个波段。为避免干扰，在暗室环境下进行光谱的测量，每次测量前进行标准白板校正，对每个样本进行不同方向上的 5 次光谱测量，取 5 条光谱数据的平均值作为土壤样本的光谱曲线。将信噪比低的边缘波段 350—399 nm 和 2 451—2 500 nm 去除，使用 Daubechies6 小波进行三层分解，采用软阈法对光谱数据中的高频系数进行去噪处理[40-42]。为了降低数据维数和减少冗余度，本研究使用最邻近法重采样，对光谱数据每 10 nm 间隔取平均值，每个样本得到由 205 个波段组成的光谱曲线。

为寻找光谱数据预测土壤质地的最佳数学变换形式，本研究选取了包括原始光谱反射率（R）、归一化（Normalization）、标准化（Standardization）、0.5 阶导数（fractional order derivative，FOD（0.5））、1 阶导数（FOD（1））、1.5 阶导数（FOD（1.5））、2 阶导数（FOD（2））、倒数的对数（inverse-log reflectance，ILR）和对数的导数（log-derivative reflectance，LDR）共 9 种土壤光谱数学变换。这些数学变换有助于突出光谱特征，在一定程度上能够提高建模精度，在土壤光谱研究中已经得到广泛应用。其中分数阶导数变换采用 Grünwald-Letnikov 算法[43]通过 MatlabR2017b 编程实现。

1.3　模型建立与精度评价

1.3.1　建模方法　支持向量机（SVM）基于统计学习理论，通过非线性的核函数将数据映射到高维特征空间，以找出一个超平面作为决策边界，使模型在数据上的分类误差尽可能小。SVM 在解决小样本、非线性和高维度数据集时具有一定的优势[44]。

决策树（DT）是一种树形结构，通过计算特征的不纯度指标，选取不纯度最优的特征进行树的分枝，

在子节点上重复分枝过程，直至所有特征分枝完成[45]。决策树的缺点是容易过拟合，因此，需要对决策树进行剪枝来提高模型的泛化性，最常用的剪枝策略是限制树的深度。

集成学习是通过构建多个基评估器，采用某种方式集成所有基评估器的结果，以此来获取比单个模型更好的建模效果。装袋法（Bagging）和提升法（Boosting）是使用最广泛的两种集成学习算法，装袋法的核心思想是从训练集中有放回的随机选取若干样本构建多个相互独立的基评估器，然后对基评估器的预测结果通过平均或者多数表决原则来决定集成评估器的结果，其代表模型就是随机森林（RF）。提升法的基评估器是相关的，其核心思想是在迭代过程中提高前一轮错误评估的样本权值，一次次对难以评估的样本进行预测，从而构建一个强评估器，自适应提升算法（AdaBoost）是其代表模型。李勇等[46]的研究综述表明集成学习在不均衡数据集建模时有一定的优势。

多层感知器（MLP）是一种构建多隐含层的深度学习模型，通过学习构建的深层非线性网络结构，从数据集中挖掘潜在的特征规律，使用非线性的激活函数提升模型的表达能力，通过优化器更新和计算模型参数，经过多次迭代不断地学习使误差最小，具有从少数样本集值学习数据集本质特征的能力[47]。

1.3.2 精度评价 分别将 9 种光谱数学变换的全谱数据作为模型的输入，对土壤质地的 4 组分类和 12 级分类进行预测，以模型的预测准确度（预测正确的样本个数占样本总数的比例）作为精度评价指标，经过多次重复训练，选取各模型在不同参数调节下表现出的最高准确度作为建模的结果，以明确最佳的模型效果，其中 SVM 模型比较了不同核函数下的效果；DT、RF 和 AdaBoost 模型比较了不同树的深度下的效果；MLP 模型通过调节隐含层个数、每个隐含层的神经元数、迭代次数 3 个参数比较建模效果。混淆矩阵是机器学习中总结分类模型预测结果的情形分析表，可以更好地了解模型对各类别的区分情况，特别是在样本非均衡时，召回率（预测准确的类别个数占实际该类别总数的比例）能够对单个类别的预测情况进行分析[48]。

常规数据统计分析软件使用软件 ArcGIS 10.2、OriginPro 9.1 和 Microsoft Excel 2010，机器学习模型的构建在 Spyder 软件中通过 Python3.7 语言编写脚本调用 Sklearn 接口中的机器学习模块实现。

2 结果

2.1 土壤质地统计特征分析

首先随机打乱所有样本的顺序，然后将每一类别的样本按照 1、2、3、4 的顺序重复进行编号，选择编号为 2、3、4 的样本作为训练数据集，编号为 1 的样本作为验证数据集，共得到 180 个训练样本，65 个验证样本。如表 1 所示，根据国际制土壤质地分类标准的 4 组 12 级对所有样本进行分类，研究区土壤样本在 4 组分类中有壤土组 38 个、黏壤土组 119 个和黏土组 88 个，无砂土组样本。再细分到 12 级分类中共有 7 种土壤质地类型，分别为砂质壤土 29 个、壤土 9 个、砂质黏壤土 22 个、黏壤土 81 个、粉砂质黏壤土 16 个、粉砂质黏土 13 个和壤质黏土 75 个。

表 1 土壤质地统计结果

Table 1 Statistical results of soil texture

4 组分类 4 groups of classifications	12 级分类 12 levels of classifications	全部样本 All samples	训练样本 Training samples	验证样本 Validation samples
壤土组 Loam group	砂质壤土 Sandy loam	29	21	8
	壤土 Loam	9	6	3
黏壤土组 Clay loam group	砂质黏壤土 Sandy clay loam	22	16	6
	黏壤土 Clay loam	81	60	21
	粉砂质黏壤土 Silty clay loam	16	12	4
黏土组 Clay group	粉砂质黏土 Silty clay	13	9	4
	壤质黏土 Loamy clay	75	56	19
合计 Total		245	180	65

为比较不同质地土壤的光谱特征变化情况，对两种分类标准下各质地的原始光谱数据取其平均值进行分析。从图 3 中发现，在 600、900、1 100 和 2 100 nm 波长附近存在交叉现象，波长大于 1 600 nm 后黏壤土组和黏土组重叠明显。在图 4 中，也存在较多的交叉重叠现象，可以看出粉砂质黏壤土的光谱曲线一直低于粉砂质黏土，在 1 400—1 900 nm 之间壤土和黏壤土重叠非常明显，砂质壤土、砂质黏壤土和壤质黏土表现得也较为相近，说明土壤质地与光谱反射率之间的规律较为复杂，用光谱反射率去区分土壤质地相对困难，但对其研究是有应用价值的。另外，在 900 nm 左右有较为明显的氧化铁吸收谷，在 1 400、1 900 和 2 200 nm 附近存在明显的水分吸收谷[49]，由于样本经过了风干处理，对于建模的影响较小，本研究不作处理，保留光谱预处理后的 205 个波段进行建模。

2.2　建模结果比较

2.2.1　4 组分类建模结果比较　在 9 种数据预处理下分别采用 5 种机器学习模型对土壤质地的 4 组分类进行建模，模型在验证集上的分类准确度比较如表 2 所示。从表 2 中可以看出，所有模型的准确度都在 0.5 以上，倒数的对数变换在使用 SVM 模型时得到全局最低准确度 0.51，进行归一化处理后使用 MLP 模型达到 0.68 的全局最高准确度。原始数据在 5 种模型中的建模准确度都位于 0.6 以上，并且 DT 和 AdaBoost 两个模型在使用原始数据进行建模时都达到了各自方法的最高准确度；除标准化外的其他 8 种数学变换都

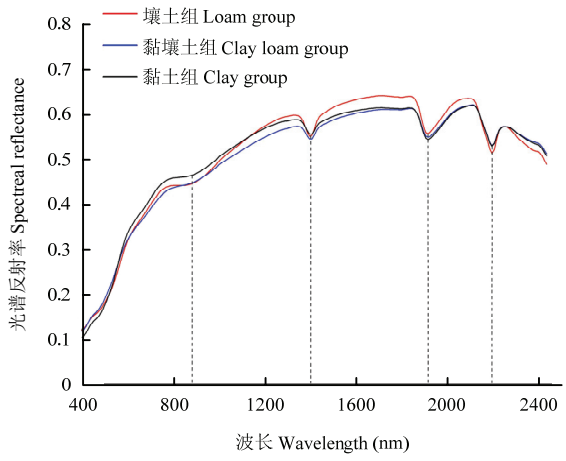

图 3　4 组分类土壤质地反射光谱曲线

Fig. 3　Reflection spectrum curve of soil texture in four groups of classifications

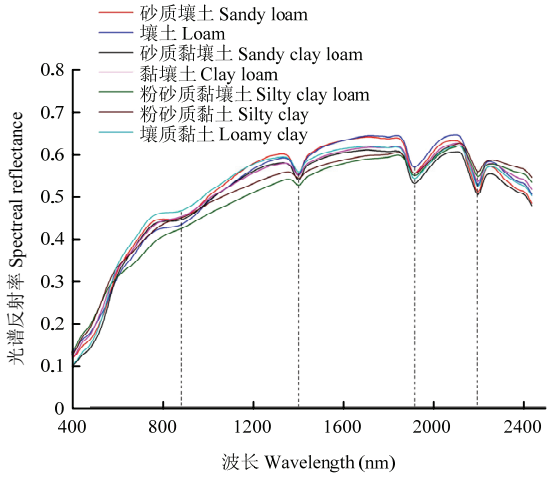

图 4　12 级分类土壤质地反射光谱曲线

Fig. 4　Reflection spectrum curve of soil texture in twelve levels of classifications

表 2　9 种数据处理和 5 种模型进行土壤质地 4 组分类的准确度比较

Table 2　Accuracy comparison of four groups of soil texture classification by nine data processing and five models

Method	SVM	DT	AdaBoost	RF	MLP
R	0.63	0.60	0.63	0.60	0.65
Normalization	0.57	0.54	0.63	0.60	0.68
Standardization	0.60	0.55	0.63	0.60	0.62
FOD（0.5）	0.65	0.52	0.55	0.58	0.66
FOD（1）	0.52	0.55	0.62	0.55	0.63
FOD（1.5）	0.57	0.57	0.60	0.65	0.65
FOD（2）	0.54	0.57	0.60	0.60	0.63
ILR	0.51	0.58	0.63	0.63	0.63
LDR	0.52	0.58	0.58	0.55	0.62

是 MLP 模型取得最高准确度，并且 MLP 模型在 9 种数学变换中准确度都大于 0.62，建模效果较好；SVM 和 RF 模型分别在 0.5 阶和 1.5 阶导数变换时达到最高准确度为 0.65；两种基于树模型的集成学习方法 AdaBoost 和 RF 在不同数学变换中建模准确度都大于或者等于单个 DT 模型，其中 AdaBoost 在多种数学变换中都优于 RF。

选取达到 0.68 最高准确度时的模型，建立其混淆矩阵如表 3 所示，预测结果分布如图 5 所示。矩阵中的每一列代表预测值，每一行代表的是实际的土壤质地类别，召回率可以知道各土壤质地类别的预测准确度，壤土组为 0.36（4/11）、黏壤土组 为 0.84（26/31）、黏土组为 0.61（14/23），原始光谱数据在进行归一化处理后对黏壤土组的预测效果最好，其次是黏土组，较难预测壤土组。同时，从表 3 中可以发现预测错误的样本绝大部分是样本数量多且与实际质地相似的类别，由于黏壤土组同时具有壤土组和黏土组的特性，所以壤土组和黏土组最容易预测错误成黏壤土组，共有 16（7+9）个样本预测错误，占样本总数的 25%。从图 5 中可以看出预测错误的类别容易出现在各类别的分界处，较难通过模型去区分土壤质地分界线附近的类别。

表 3 归一化处理和 MLP 模型混淆矩阵

Table 3 Normalization and MLP model confusion matrix

4 组分类 4 groups of classifications	壤土组 Loam group	黏壤土组 Clay loam group	黏土组 Clay group	合计 Total
壤土组 Loam group	4	7	0	11
黏壤土组 Clay loam group	1	26	4	31
黏土组 Clay group	0	9	14	23
合计 Total	5	42	18	65

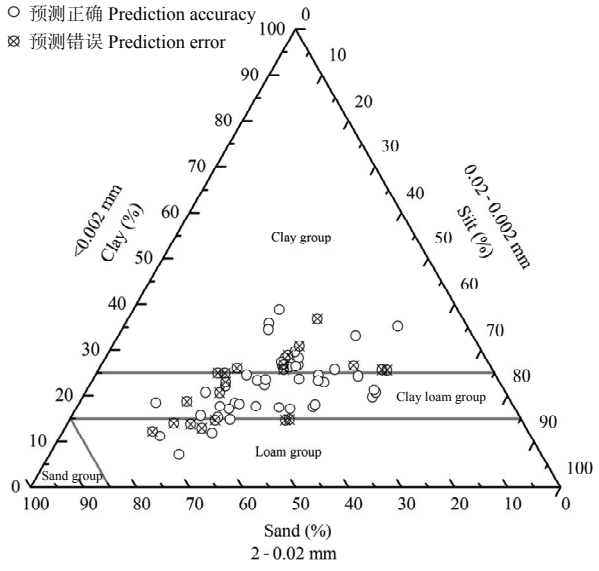

图 5 归一化处理和 MLP 模型预测结果分布图

Fig. 5 Normalized processing and MLP model prediction result distribution

2.2.2 12 级分类建模结果比较 将 4 组分类的土壤质地再细分到 12 级分类进行建模，模型在验证集上的分类准确度比较如表 4 所示。从表中可以看出，由于再将土壤质地类别进行细分，模型的准确度都在一定程度上有所降低，较难再用光谱数据对土壤质地进行区分。使用原始数据在 MLP 模型中达到 0.55 的全局最高准确度，0.40 的全局最低准确度来自于 SVM 在进行 1 阶导数或者对数的导数变换；两种集成学习方法和 MLP 模型使用原始数据建模都取得了最高的准确度；归一化和标准化处理的效果基本相当；在 5 种导数变换中，1.5 阶导数变换在除 SVM 外的其他 4 种建模方法中准确度都最高；除倒数的对数外的其他 8 种数学变换都是 MLP 模型取得最高准确度，并且在所有数学变换中的准确度都大于 0.49，模型表现依然较好；两种集成学习方法依然在多种数学变换中都优于 DT 模型，但在 0.5 阶导数变换时效果不好。从整体来看 4 组和 12 级两次分类，各种机器学习模型的建模效果趋势基本相同，各数学变换方法略微有所差异，但依然是原始光谱建模效果较好，分数阶导数普遍优于整数阶导数变换。

选取原始数据和 MLP 组合的 0.55 最高准确度时的模型，建立其混淆矩阵如表 5 所示，预测结果分布如图 6 所示。从召回率可以得到 7 种土壤质地类别的预测准确度，砂质壤土为 0.38（3/8）、壤土 0（0/3）、砂质黏壤土为 0.17（1/6）、黏壤土为 0.67（14/21）、粉砂质黏壤土 0（0/4）、粉砂质黏土为 0.25（1/4）、壤质黏土为 0.89（17/19），原始光谱数据和 MLP 的

表 4　9 种数据处理和 5 种模型进行土壤质地 12 级分类的准确度比较

Table 4　Accuracy comparison of soil texture classification of twelve levels by nine data processing and five models

Method	SVM	DT	AdaBoost	RF	MLP
R	0.48	0.46	0.52	0.51	0.55
Normalization	0.49	0.48	0.49	0.51	0.52
Standardization	0.46	0.48	0.51	0.51	0.52
FOD（0.5）	0.48	0.46	0.43	0.42	0.51
FOD（1）	0.40	0.42	0.48	0.46	0.49
FOD（1.5）	0.43	0.49	0.49	0.49	0.51
FOD（2）	0.42	0.42	0.46	0.48	0.49
ILR	0.45	0.49	0.51	0.46	0.49
LDR	0.40	0.46	0.49	0.48	0.49

表 5　原始数据和 MLP 模型混淆矩阵

Table 5　Raw data and MLP model confusion matrix

12 级分类 12 levels of classifications	砂质壤土 Sandy loam	壤土 Loam	砂质黏壤土 Sandy clay loam	黏壤土 Clay loam	粉砂质黏壤土 Silty clay loam	粉砂质黏土 Silty clay	壤质黏土 Loamy clay	合计 Total
砂质壤土 Sandy loam	3	0	0	3	0	1	1	8
壤土 Loam	1	0	0	2	0	0	0	3
砂质黏壤土 Sandy clay loam	0	0	1	2	0	1	2	6
黏壤土 Clay loam	1	0	0	14	0	3	3	21
粉砂质黏壤土 Silty clay loam	0	0	0	4	0	0	0	4
粉砂质黏土 Silty clay	0	0	0	2	0	1	1	4
壤质黏土 Loamy clay	0	0	0	2	0	0	17	19
合计 Total	5	0	1	29	0	6	24	65

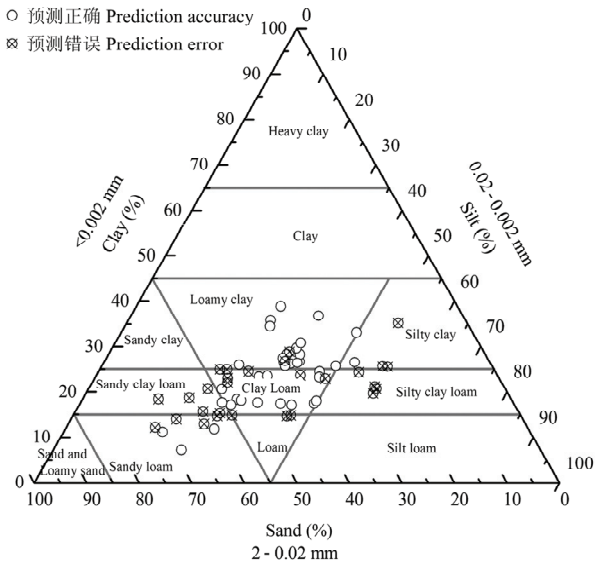

图 6　原始数据和 MLP 模型预测结果分布图

Fig. 6　Raw data and MLP model prediction result distribution

组合模型对壤质黏土的预测效果最好，达到 0.89，其次是黏壤土 0.67，较难分辨砂质壤土、砂质黏壤土和粉砂质黏土，完全不能区分壤土和粉砂质黏壤土。此时，从表 5 中可以发现预测错误的样本除了是与实际质地相似的类别外，还容易预测错误成样本数量较多的类别，如有 15 个样本预测错误为黏壤土，7 个样本错分成壤质黏土，被预测错误的比例分别占了验证集样本总数的 23%和 11%。结合图 6，除了仍然在土壤质地划分的边界处容易预测错误外，黏壤土和壤质黏土位于三角坐标图的中心，两种质地同时具有黏土和壤土的特性，最容易被错分成这两种质地类型。

3　讨论

通过光谱反射率预测土壤质地是高光谱技术的重要应用，在构建模型时，对原始数据进行各种数学变换以及选取合适的建模方法是研究此类问题的关键。目前较多研究表明，进行相应的数学变换可以提高模

型的精度[21,28,30,32]，国佳欣等[50]将分数阶导数应用到有机质的回归建模中，而在分类建模中使用较少，本研究在应用较多的整数阶导数的基础上选取了 0.5 阶和 1.5 阶两个中间段的分数阶导数，结果表明分数阶导数建模效果普遍优于整数阶导数。由此可见，在进行光谱数据预处理时，不应局限于整数阶导数变换，进行分数阶导数变换能将光谱内隐含的信息更好的表现出来，以提高建模的精度。但结合两次分类结果来看，原始数据由于具有了更加丰富的信息，在多种模型中的建模准确度相较于其他数学变换方法都最高，更适合预测土壤质地，这与王德彩等[31]结果一致。从建模方法来看，在两次分类中 MLP、AdaBoost 和 RF 模型效果都较好，其中 MLP 模型因其能更好地挖掘特征之间的内在规律而效果最佳，这也是神经网络模型在光谱建模的研究中广泛应用的原因[35]；AdaBoost 和 RF 模型是以树模型为基评估器的集成评估器，有着能够处理高维度数据、抗过拟合和泛化能力强的优点[51]，建模效果要比单个 DT 模型好。

各种土壤质地的光谱曲线形状基本相似，不同质地之间的区分不明显，且在数值上存在较多的交叉重叠现象，说明土壤质地与光谱反射率的规律较为复杂，用光谱反射率去区分土壤质地相对困难。这在预测结果中也得到了证明，土壤质地 4 组分类的预测精度最高仅为 0.68，在 12 级分类中的最高准确度降至 0.55，相较于曾庆猛等[26]的研究结果 4 组类 0.72 的准确度稍低，12 级类 0.22 的准确度高很多。同时，通过两次分类中取得最高准确度的混淆矩阵和预测结果三角坐标图发现，预测错误的样本大部分错分为与实际质地相似的类别，且容易出现在图中各类别的分界处，这是因为分界附近的各粒径百分含量很接近，质地类型相似，因此光谱反射率也会相差较小，较难通过模型去区分土壤质地分界线附近的质地类别。另外，还容易错分成样本数量多的类别，一方面可能是因为在样本数量不均衡的情况下，模型在训练时对样本数量多的类别能够学习到更多的特征规律，在验证集上容易将其他类别错分成样本数量多的类别[52]；另一方面 4 组类的黏壤土组以及 12 级类的黏壤土和壤质黏土都位于三角坐标图的中心，与多种质地边界相邻，同时具有黏土和壤土的特性，所以最容易被预测错误成这两种质地类型，这与 CHAWLA 等[53]提到非均衡数据集错误分类经常发生在类边界附近相同。

本研究的不足之处在于样本各土壤质地类别存在一定的不均衡问题，但在实际的采样过程中很难保证样本的绝对均衡性和代表性。因此，为了提高建模的精度，可以发掘更好的数学变换方法，寻找特征波段，尝试降维处理，使用更好的模型，如当前最热门的深度学习模型，同时还可以考虑不同土地利用类型下的土壤质地分类，这些也是今后需要进一步深入研究的方向。

4 结论

基于 245 个红壤样本的 Vis-NIR 光谱，在国际制土壤质地 4 组和 12 级两种分类标准下，采用的 9 种数学变换方法和 5 种机器学习算法相互组合的数据挖掘模型，进行土壤质地的分类研究。基于土壤质地 4 组分类时，归一化处理和 MLP 模型组合取得了 0.68 的最高准确度，其中黏壤土组的预测准确度能达到 0.84；再细分到 12 级分类后，分类效果最佳的组合来自于原始数据和 MLP 模型，其中壤质黏土分类准确度达到了 0.89。因此，本文的研究结果表明光谱分析方法快速进行土壤质地分类是可行的，同时为非均衡数据集分类建模在方法和思路上提供一定的参考。

References

[1] GREVE M H, KHEIR R B, GREVE M B, BØCHER P K. Quantifying the ability of environmental parameters to predict soil texture fractions using regression-tree model with GIS and LIDAR data: The case study of Denmark. *Ecological Indicators*, 2012, 18: 1-10.

[2] SHAHRIARI M, DELBARI M, AFRASIAB P, PAHLAVAN-Rad M R. Predicting regional spatial distribution of soil texture in floodplains using remote sensing data: A case of southeastern Iran. *Catena*, 2019, 182: 104149.

[3] 吴克宁, 赵瑞. 土壤质地分类及其在我国应用探讨. 土壤学报, 2019, 56(1): 227-241.
WU K N, ZHAO R. Soil texture classification and its application in China. *Acta Pedologica Sinica*, 2019, 56(1): 227-241. (in Chinese)

[4] 张娜, 张栋良, 李立新, 屈忠义. 基于高光谱的区域土壤质地预测模型建立与评价——以河套灌区解放闸灌域为例. 干旱区资源与环境, 2014, 28(5): 67-72.
ZHANG N, ZHANG D L, LI L X, QU Z Y. Establishment and evaluation of model for predicting soil texture based on hyperspectral data—Case study of Jiefangzha irrigation area in Hetao irrigation district. *Journal of Arid Land Resources and Environment*, 2014, 28(5): 67-72. (in Chinese)

[5] 乔天, 吕成文, 肖文凭, 吕凯, 水宏伟. 基于遗传算法的土壤质地

高光谱预测模型研究. 土壤通报, 2018, 49(4): 773-778.
QIAO T, LÜ C W, XIAO W P, LÜ K, SHUI H W. Hyperspectral prediction modeling of soil texture based on genetic algorithm. *Chinese Journal of Soil Science*, 2018, 49(4): 773-778. (in Chinese)

[6] 于雷, 洪永胜, 周勇, 朱强, 徐良, 李冀云, 聂艳. 高光谱估算土壤有机质含量的波长变量筛选方法. 农业工程学报, 2016, 32(13): 95-102.
YU L, HONG Y S, ZHOU Y, ZHU Q, XU L, LI J Y, NIE Y. Wavelength variable selection methods for estimation of soil organic matter content using hyperspectral technique. *Transactions of the Chinese Society of Agricultural Engineering*, 2016, 32(13): 95-102. (in Chinese)

[7] 史舟, 王乾龙, 彭杰, 纪文君, 刘焕军, 李曦. 中国主要土壤高光谱反射特性分类与有机质光谱预测模型. 中国科学:地球科学, 2014, 44(5): 978-988.
SHI Z, WANG Q L, PENG J, JI W J, LIU H J, LI X. Classification of hyperspectral reflectance characteristics and prediction model of organic matter spectra of main soils in China. *Scientia Sinica (Terrae)*, 2014, 44(5): 978-988. (in Chinese)

[8] MÜLLER B BERNHARDT M, JACKISCH C, SCHULZ K. Estimating spatially distributed soil texture using time series of thermal remote sensing & ndash: A case study in central Europe. *Hydrology and Earth System Sciences*, 2016, 20(9): 3765-3775.

[9] SAYAO V M, DEMATTÊJÉA M. Soil texture and organic carbon mapping using surface temperature and reflectance spectra in Southeast Brazil. *Geoderma Regional*, 2018, 14: e174.

[10] ZHAI Y S, THOMASSON J A, BOGGESS J E, SUI R X. Soil texture classification with artificial neural networks operating on remote sensing data. *Computers and Electronics in Agriculture*, 2006, 54(2): 53-68.

[11] ZHAO Z Y, CHOW T L, REES H W, YANG Q, XING Z S, MENG F R. Predict soil texture distributions using an artificial neural network model. *Computers & Electronics in Agriculture*, 2009, 65(1): 36-48.

[12] CHUNG S O, CHO K H, KONG J W, JUNG K Y. Soil Texture classification algorithm using RGB characteristics of soil images. *IFAC Proceedings Volumes*, 2010, 43(26): 34-38.

[13] BARMAN U, CHOUDHURY R D. Soil texture classification using multi class support vector machine. *Information Processing in Agriculture*, 2020, 7(2): 318-332.

[14] WU W, LI A D, HE X H, MA R, LIU H B, LÜ J K. A comparison of support vector machines, artificial neural network and classification tree for identifying soil texture classes in southwest China. *Computers and Electronics in Agriculture*, 2018, 144: 86-93.

[15] ADHIKARI K, KHEIR R B, GREVE M B, BØCHER P K, MALONE B P, MINASNY B, MCBRATNEY A B, GREVE M H. High-resolution 3-D mapping of soil texture in denmark. *Soil Science Society of America Journal*, 2013, 77(3): 860-876.

[16] LIE M, GLASER B, HUWE B. Uncertainty in the spatial prediction of soil texture. *Geoderma*, 2012,170:70-79.

[17] 孙艳俊, 张甘霖, 杨金玲, 赵玉国. 基于人工神经网络的土壤颗粒组成制图. 土壤, 2012, 44(2): 312-318.
SUN Y J, ZHANG G L, YANG J L, ZHAO Y G. Mapping of soil particle composition based on artificial neural network. *Soils*, 2012, 44(2): 312-318. (in Chinese)

[18] SILVA S H G, WEINDORF D C, PINTO L C, FARIA W M, JUNIOR F W A, GOMIDE L R, MELLO J M D, JUNIOR A L D P, SOUZA I A D, TEIXEIRA A F D S, GUILHERME L R G, CURI N. Soil texture prediction in tropical soils: A portable X-ray fluorescence spectrometry approach. *Geoderma*, 2020, 362: 114136.

[19] BAO N S, LIU S J, ZHOU Y C. Predicting particle-size distribution using thermal infrared spectroscopy from reclaimed mine land in the semi-arid grassland of North China. *Catena*, 2019, 183: 104190.

[20] PENG Y, KNADEL M, GISLUM RÉ, SCHELDE K, THOMSEN A, GREVE M H. Quantification of SOC and clay content using visible Near-Infrared Reflectance-Mid-Infrared reflectance spectroscopy with Jack-Knifing partial least squares regression. *Soil Science*, 2014: 179(7): 325-332.

[21] 王德彩, 蔚霖, 张俊辉, 杨红震, 黄家荣, 孙孝林. 基于正交信号校正的 Vis-NIR 光谱土壤质地预测. 河南农业大学学报, 2017(3): 408-413.
WANG D C, WEI L, ZHANG J H, YANG H Z, HUANG J R, SUN X L. Prediction of soil texture using Vis-NIR spectra based on orthogonal signal correction. *Journal of Henan Agricultural University*, 2017(3): 408-413. (in Chinese)

[22] 王德彩, 邬登巍, 赵明松, 张甘霖. 平原区土壤质地的反射光谱预测与地统计制图. 土壤通报, 2012, 43(2): 257-262.
WANG D C, WU D W, ZHAO M S, ZHANG G L. Prediction and mapping of soil texture of a plain area using reflectance spectra and geo-statistics. *Chinese Journal of Soil Science*, 2012, 43(2): 257-262. (in Chinese)

[23] 沈掌泉, 单英杰, 王珂. 田间行走式测定的红外光谱数据与土壤质地之间的相关性研究. 光谱学与光谱分析, 2009, 29(6): 1526-1530.
SHEN Z Q, DAN Y J, WANG K. Study on relationship between on-the-go near-infrared spectroscopy and soil texture. *Spectroscopy and Spectral Analysis*, 2009, 29(6): 1526-1530. (in Chinese)

[24] AMIRIAN-CHAKAN A MINASNY B, TAGHIZADEH-MEHRJARDI

R, AKBARIFAZLI R, DARVISHPASAND Z, KHORDEHBIN S. Some practical aspects of predicting texture data in digital soil mapping. *Soil and Tillage Research*, 2019, 194: 104289.

[25] LARK R M, BISHOP T F A. Cokriging particle size fractions of the soil. *European Journal of Soil Science*, 2007, 58(3): 763-774.

[26] 曾庆猛, 孙宇瑞, 严红兵. 土壤质地分类的近红外光谱分析方法研究. 光谱学与光谱分析, 2009, 29(7): 1759-1763.
ZENG Q M, SUN Y R, YAN H B. NIR spectral analysis for soil textural classification. *Spectroscopy and Spectral Analysis*, 2009, 29(7): 1759-1763. (in Chinese)

[27] 胡晓艳, 宋海燕. 基于支持向量机和近红外光谱特性的土壤质地分类. 山西农业科学, 2017, 45(10): 1643-1645.
HU X Y, SONG H Y. Soil texture classification based on support vector machine and near infrared spectral characteristics. *Journal of Shanxi Agricultural Sciences*, 2017, 45(10): 1643-1645. (in Chinese)

[28] 白燕英, 魏占民, 刘全明, 郭桂莲，刘霞. 基于高光谱的河套灌区农田表层土壤质地反演研究. 地理与地理信息科学, 2013, 29(5): 68-71.
BAI Y Y, WEI Z M, LIU Q M, GUO G L, LIU X. Surface soil texture of field inverted using hyper-spectrum in Hetao irrigation. *Geography and Geo-Information Science*, 2013, 29(5): 68-71. (in Chinese)

[29] SALAZAR D F U, DEMATT JÉA M, VICENTE L E, GUIMARAES C C B, SAYAO V M, CERRI C E P, PADILHA M C D C, MENDES W D S. Emissivity of agricultural soil attributes in southeastern Brazil via terrestrial and satellite sensors. *Geoderma*, 2020, 361: 114038.

[30] 黄明祥, 程街亮, 王珂, 龚建华，李洪义，史舟. 海涂土壤高光谱特性及其砂粒含量预测研究. 土壤学报, 2009, 46(5): 932-937.
HUANG M X, CHENG J L, WANG K, GONG J H, LI H Y, SHI Z. Coastal soil hyperspectral characteristics and soil sand content prediction. *Acta Pedologica Sinica*, 2009, 46(5): 932-937. (in Chinese)

[31] 王德彩, 张俊辉. 基于 Vis-NIR 光谱的土壤质地 BP 神经网络预测. 天津农业科学, 2015, 21(8): 6-9.
WANG D C, ZHANG J H. Estimation of soil texture based on Vis-NIR spectroscopy and BP neural network. *Tianjin Agricultural Sciences*, 2015, 21(8): 6-9. (in Chinese)

[32] 李春蕾, 许端阳, 陈蜀江. 基于高光谱遥感的新疆北疆地区土壤砂粒含量反演研究. 干旱区地理, 2012, 35(3): 473-478.
LI C L, XU D Y, CHEN S J. Soil sand content retrieving of bare soil in north Xinjiang based on hyper-spectral remote sensing. *Arid Land Geography*, 2012, 35(3): 473-478. (in Chinese)

[33] TONG P J, DU Y P, ZHENG K Y, WU T, WANG J J. Improvement of NIR model by fractional order Savitzky-Golay derivation(FOSGD) coupled with wavelength selection. *Chemometrics and Intelligent Laboratory Systems*, 2015, 143: 40-48.

[34] 王敬哲, 塔西甫拉提·特依拜, 丁建丽, 张东，刘巍. 基于分数阶微分预处理高光谱数据的荒漠土壤有机碳含量估算. 农业工程学报, 2016, 32(21): 161-169.
WANG J Z, TASHPOLAT TIYIP, DING J L, ZHANG D, LIU W. Estimation of desert soil organic carbon content based on hyperspectral data preprocessing with fractional differential. *Transactions of the Chinese Society of Agricultural Engineering*, 2016, 32(21): 161-169. (in Chinese)

[35] XU Z, ZHAO X M, GUO X, GUO J X. Deep learning application for predicting soil organic matter content by Vis-NIR spectroscopy. *Computational Intelligence and Neuroscience*, 2019, 2019: 1-11.

[36] 纪文君, 李曦, 李成学, 周银，史舟. 基于全谱数据挖掘技术的土壤有机质高光谱预测建模研究. 光谱学与光谱分析, 2012(9): 91-96.
JI W J, LI X, LI C X, ZHOU Y, SHI Z. Using different data mining algorithms to predict soil organic matter based on visible-near infrared spectroscopy. *Spectroscopy and Spectral Analysis*, 2012(9): 91-96. (in Chinese)

[37] ESHEL G, LEVY G J, MINGELGRIN U, SINGER M J. Critical evaluation of the use of laser diffraction for particle-size distribution analysis. *Soil Science Society of America Journal*, 2004, 68(3): 736.

[38] 杨金玲, 张甘霖, 李德成, 潘继花. 激光法与湿筛-吸管法测定土壤颗粒组成的转换及质地确定. 土壤学报, 2009(5): 22-30.
YANG J L, ZHANG G L, LI D C, PAN J H. Relationships of soil particle size distribution between sieve-pipetie and laser diffraction methods. *Acta Pedologica Sinica*, 2009(5): 22-30. (in Chinese)

[39] 李学林, 李福春, 陈国岩, 谢昌仁, 王金平, 李文静. 用沉降法和激光法测定土壤粒度的对比研究. 土壤, 2011(1): 132-136.
LI X L, LI M C, CHEN G Y, XIE C R, WANG J P, LI W J. Comparative study on grain-size measured by laser diffraction and sedimentation techniques. *Soils*, 2011(1): 132-136. (in Chinese)

[40] 史舟. 土壤地面高光谱遥感原理与方法. 北京: 科学出版社, 2014: 61-63.
SHI Z. *Principle and Method of Hyperspectral Remote Sensing of Soil Surface*. Beijing: Science Press, 2014: 61-63. (in Chinese)

[41] HU Y G, JIANG T, SHEN A G, LI W, WANG X, HU J. A background elimination method based on wavelet transform for Raman spectra. *Chemometrics & Intelligent Laboratory Systems*, 2007, 85(1): 94-101.

[42] 马翠红, 刘立业. 基于小波分析的光谱数据处理. 冶金分析, 2012, 32(1): 34-37.
MA C H, LIU L Y. Spectral data processing based on wavelet analysis. *Metallurgical Analysis(China)*, 2012, 32(1): 34-37. (in Chinese)

[43] BENKHETTOU N, CRUZ A M C B D, TORRES D F M. A fractional calculus on arbitrary time scales: Fractional differentiation and fractional integration. *Signal Processing*, 2015, 107: 230-237.

[44] 丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法研究综述. 电子科技大学学报, 2011, 40(1): 2-10.
DING S F, QI B J, TAN H Y. An overview on theory and algorithm of support vector machines. *Journal of University of Electronic Science and Technology of China*, 2011, 40(1): 2-10. (in Chinese)

[45] 刘勇洪, 牛铮, 王长耀. 基于 MODIS 数据的决策树分类方法研究与应用. 遥感学报, 2005(4): 405-412.
LIU Y H, NIU Z, WANG C Y. Research and application of the decision tree classification using MODIS data. *Journal of Remote Sensing*, 2005(4): 405-412. (in Chinese)

[46] 李勇, 刘战东, 张海军. 不平衡数据的集成分类算法综述. 计算机应用研究, 2014, 31(5): 1287-1291.
LI Y, LIU Z D, ZHANG H J. Review on ensemble algorithms for imbalanced data classification. *Application Research of Computers*, 2014, 31(5): 1287-1291. (in Chinese)

[47] 孙志军, 薛磊, 许阳明, 王正. 深度学习研究综述. 计算机应用研究, 2012, 29(8): 2806-2810.
SUN Z J, XUE L, XU Y M, WANG Z. Overview of deep learning. *Application Research of Computers,* 2012, 29(8): 2806-2810. (in Chinese)

[48] XU J F, ZHANG Y J, MIAO D Q. Three-way confusion matrix for classification: A measure driven view. *Information Sciences*, 2020, 507: 772-794.

[49] 赵小敏, 杨梅花. 江西省红壤地区主要土壤类型的高光谱特性研究. 土壤学报, 2018, 55(1): 31-42.
ZHAO X M, YANG M H. Hyper-spectral characteristics of major types of soils in red soil region of Jiangxi Province, China. *Acta Pedologica Sinica*, 2018, 55(1): 31-42. (in Chinese)

[50] 国佳欣, 赵小敏, 郭熙, 徐喆, 朱青, 江叶枫. 基于 PLSR-BP 复合模型的红壤有机质含量反演研究. 土壤学报, 2020, 57(3): 636-645.
GUO J X, ZHAO X M, GUO X, XU Z, ZHU Q, JIANG Y F. Inversion of organic matter content in red soil based on PLSR-BP composite model. *Acta Pedologica Sinica*, 2020, 57(3): 636-645. (in Chinese)

[51] 方匡南, 吴见彬, 朱建平, 谢邦昌. 随机森林方法研究综述. 统计与信息论坛, 2011, 26(3): 32-38.
FANG K N, WU J B, ZHU J P, XIE B C. A review of technologies on random forests. *Statistics & Information Forum*, 2011, 26(3): 32-38. (in Chinese)

[52] 叶志飞, 文益民, 吕宝粮. 不平衡分类问题研究综述. 智能系统学报, 2009, 4(2): 148-156.
YE Z F, WEN Y M, LÜ B L. A survey of imbalanced pattern classification problems. *CAAI Transactions on Intelligent Systems,* 2009, 4(2):148-156. (in Chinese)

[53] CHAWLA N V, JAPKOWICZ N, KOTCZ A. Editorial: Special issue on learning from imbalanced data sets. *Acm Sigkdd Explorations Newsletter*, 2004, 6(1): 1-6.

（责任编辑 李云霞）