基于机器学习的沟谷地貌识别模型对比
——以黄土高原典型流域为例

范天程，汪珍亮，李云飞，贾云飞，袁 可，赵建林
（长安大学地质工程与测绘学院，西安 710054）

摘要：探索沟谷地貌空间分布与环境控制特征之间的联系并构建沟谷地貌准确提取模型，对大尺度范围沟谷提取具有重要意义。基于人工提取黄土高原典型流域沟谷地貌样本，结合不同时期的 Landsat8 OLI 影像数据和 DEM 数据，建立随机森林模型确定黄土高原沟谷地貌提取最佳影像时期和最佳组合特征，基于最优模型参数，对比其与支持向量机和人工神经网络沟谷提取模型效果，验证模型泛化能力。结果表明：(1)黄土高原沟谷提取的最佳影像时期为 12 月，最佳组合特征集为 Red、Blue、*H*、SWIR1、PNT、Coastal、GLCM4 和 NIR；(2)3 种方法提取测试区域的沟谷空间分布一致，从定性和定量角度进行比较，随机森林模型提取效果最好，验证样区平均总体精度为 80.48%，相较于支持向量机模型和人工神经网络模型分别提高 4.00 和 8.63 个百分比；(3)测试区域中，沟谷地貌面积约占总面积的 56.91%，且呈现西北至东南方向逐渐集中的特点。研究表明随机森林模型在黄土高原地区高精度沟谷地貌识别研究中综合表现最佳，可大范围推广。

关键词：沟谷分布；机器学习；遥感影像；地形特征；黄土高原
中图分类号：P208 **文献标识码**：A **文章编号**：1009-2242(2023)04-0205-09
DOI：10.13870/j.cnki.stbcxb.2023.04.026

Comparing the Performance of Machine Learning Models for Identifying Gully Landforms
—A Case Study of a Typical Watershed on the Chinese Loess Plateau

FAN Tiancheng, WANG Zhenliang, LI Yunfei, JIA Yunfei, YUAN Ke, ZHAO Jianlin
(*College of Geological Engineering and Geomantics, Chang'an University, Xi'an* 710054)

Abstract: Exploring the relationship between spatial distribution and environmental control characters of gully landforms and building accurate extraction model are of great significance for gully landforms extraction in large scale. Based on the artificial extraction of gully landform samples combing with Landsat8 OLI image data with different periods of and DEM data of a typical watershed on the Chinese loess plateau, the random forest model was established to determine the best period for gully landforms extraction and the best combination of gullying features. Then, combined with the optimal model parameters, results of random forest were compared with support vector machine and artificial neural network model to validate the model generalization ability. Our results showed that: (1) The best image period for gully extraction was in December, and the best combination feature set was Red, Blue, elevation (*H*), SWIR1, positive and negative terrain (PNT), Coastal, texture (GLCM4) and NIR; (2) The distribution of gully landforms in the testing area extracted by three methods had consistently spatial pattern. Based on qualitatively and quantitatively modelling performance, the random forest model presented the best extracting performance, with the average overall accuracy of 80.48%, which was higher by 4.00 percentage and 8.63 percentage compared with the support vector machine model and the artificial neural network model, respectively; (3) The gully landforms accounted for 56.91% of the total testing area and the distribution of gullies in testing area was gradually concentrated from northwest to southeast. The results show that the random forest model has the best comprehensive performance in the study of high-precision gully landforms identification on the Chinese Loess Plateau, and can be widely extended.

收稿日期：2022-11-17
资助项目：国家自然科学基金项目(41907048)；中央高校基本科研费专项(300102260206)
第一作者：范天程(1997—)，男，硕士研究生，主要从事地貌遥感研究。E-mail：2020226008@chd.edu.cn
通信作者：赵建林(1988—)，男，副教授，硕士生导师，主要从事土壤侵蚀与区域生态评价研究。E-mail：jianlin.zhao@chd.edu.cn

Keywords: gully distribution; machine learning; remote sensing image; topographical characters; Chinese Loess Plateau

黄土高原是世界上侵蚀最为严重的区域之一,强烈侵蚀过程造就了黄土高原"千沟万壑"的地貌特征,在所有侵蚀过程中,发生于沟谷地貌的沟蚀过程是该区域泥沙的主要来源。已有研究[1]表明,黄土高原地区当沟谷密度大于30%,小流域泥沙贡献超过75%。在世界范围内,沟谷侵蚀作为一种常见的自然现象,特别是在干旱半干旱地区,是导致土地严重退化的主要原因之一,造成土壤质量下降、农业生产力降低和水生生物量减少等不利影响[2]。准确高效地获取黄土高原地区沟谷地貌空间分布对当地水土保持、泥沙控制、环境保护以及流域管理等领域具有重要意义。

针对黄土高原地貌识别,国内学者开展了系列研究。其中,宏观上可将黄土高原地貌分为坡面和沟谷区域,基于尺度大小沟谷区可细分为细沟、浅沟、切沟、冲沟、坳沟和河沟等地貌类型。关于黄土高原侵蚀沟提取,早期研究主要基于单一高分辨率的DEM数据(5 m及以上),采用多向阴影法[3]、地形开度和差值图像阈值分割法[4]等方法进行地貌的提取和分割。后来基于高分影像数据的提取方法逐渐普及,但因其存在严重的"同物异谱,同谱异物"现象[5],多数学者在侵蚀沟提取研究中同时加入了地形因子,常用方法有流向检测法[6]、定向边缘检测法[7]等。虽然这些研究能够获得分辨率较高的侵蚀沟地貌,但对高分辨率地形数据依赖程度高,因此相关试验多数为小尺度区域(<10 km^2),在较大尺度区域内存在高分数据获取困难等问题,无法高效地推广至大尺度区域[8]。近年来机器学习发展迅速,广泛应用于不同领域大尺度范围研究[9-11],相关模型也被尝试应用到沟谷地貌的研究中。目前已有学者基于机器学习方法从侵蚀沟密度[12]、侵蚀沟易发性[13]角度在大范围流域对侵蚀沟展开研究,取得较好的结果。而在黄土高原地区基于机器学习方法的大范围沟谷地貌提取研究较少,目前主要集中在小范围内[14]。针对黄土高原大面积的沟谷识别研究,Liu等[15]提出一种结合随机森林和地形骨架的方法提取了黄土高原丘陵沟壑区(约15.4万 km^2)的侵蚀沟分布,但基于随机分布点的预测精度仅为78.80%。因此,上述研究中所涉及方法在黄土高原大尺度沟谷提取研究还存在较大的不确定性。同时,本文选取Google Earth Pro平台作为训练和测试样本的获取途径,作为一款免费提供高分辨率遥感影像的工具,其具备高效提取高质量样本的潜力[1]。

综上所述,在黄土高原沟谷识别的研究中,目前的研究存在着提取精度高但研究范围小与研究范围大但提取精度低的两难问题,直接将小范围高精度的研究方法应用到大区域内,需要考虑到数据获取难度和时间成本,难以实现大范围高精度的沟谷地貌提取。因此,本研究探索以30 m分辨率的Landsat8 OLI遥感影像和30 m DEM数据作为数据源,以黄土高原典型流域作为研究区,以冲沟、河沟等宏观沟谷地貌为提取对象,结合Google Earth Pro平台人工解译像元尺度的样本数据,融合地形和光谱特征,探索基于多种机器学习模型的黄土高原沟谷地貌提取方法研究,并分析其精度和控制特征。以期提供一种综合考虑尺度和精度的沟谷地貌空间分布制图方法,为整个黄土高原的沟谷地貌提取和泥沙治理提供可靠的方法和数据支撑。

1 研究区域与数据来源

1.1 研究区概况与样本分布

本研究主要在延河流域(36°21′—37°19′N,108°38′—110°29′E)开展,延河流域位于黄土高原中部,属于黄土丘陵沟壑区,地势西北高,东南低,地表破碎[16],是黄河中游水土流失最严重的区域之一[17],该流域土壤侵蚀程度剧烈,沟谷分布广泛,呈西北到东南方向逐渐集中的特点[18]。

本研究在延河流域东南方位选取沟谷提取训练区域(32.74 km^2)、测试区域(76.76 km^2)2个典型流域,以及6个沟谷提取验证样区(1 km^2×1 km^2)(图1)。其中,训练区域用于获取沟谷训练样本;测试区域用来对比不同模型的沟谷提取结果;验证样区用于提取结果的定量分析。沟谷样本是基于Google Earth Pro平台人工解译,解译过程(图1)为:首先,基于DEM数据在训练区域内随机分布3 000个像元;然后,基于该平台高分影像识别每一个像元的类别(沟谷、坡面(沟谷以外区域));最后,共识别沟谷像元1 398个、坡面像元650个和不确定像元952个。随机选取500个沟谷和500个坡面像元作为后续模型建立的样本数据集。

1.2 数据来源及预处理

1.2.1 *DEM数据及预处理* 本文采用的DEM数据来源于地理空间数据云平台(http://www.gscloud.cn),为ASTER GDEM 30m分辨率数字高程模型,在对DEM进行拼接、填洼和裁剪后,将结果投影至WGS_1984_UTM_Zone_49N坐标系,得到延河流域DEM数据,作为下一步分析的基础数据。

1.2.2 *Landsat8 OLI影像及预处理* Landsat8 OLI

影像数据来源于地理空间数据云平台,选择训练区域和测试区域 2016—2021 年 1—12 月云量小于 10%的遥感影像,且研究区域上方无云层遮挡。由于 Landsat8 OLI 影像数据已经经过系统的几何校正,故只对影像数据进行辐射定标、大气校正和裁剪。分析获取的遥感影像,相邻月份之间相差的天数最大为 38 天,最小为 22 天,其中 3 月、7 月、8 月和 11 月对应影像条带号为 126、行编号为 35,剩余月份对应影像条带号为 127、行编号为 35。由于获取的遥感影像成像年份相近,不考虑将其作为沟谷提取的影响因素。

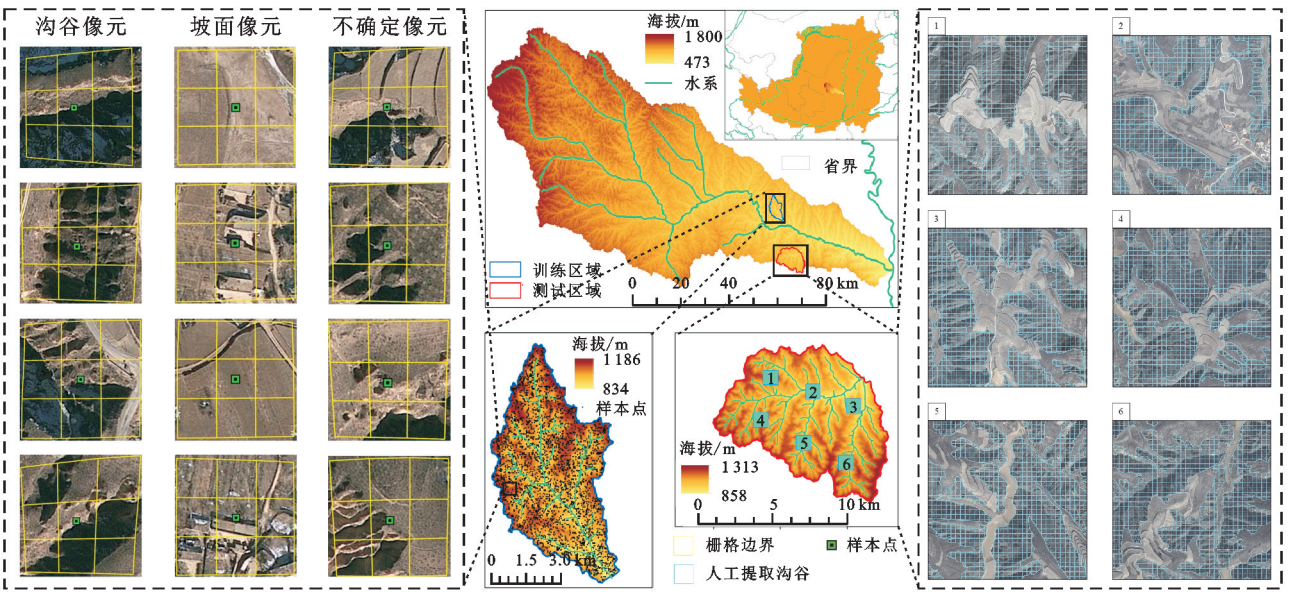

图 1　研究区域地理位置及沟谷样本点识别和分布

1.3　机器学习模型

1.3.1　随机森林　随机森林(random forest,RF)是一种由决策树组成的集成算法,适合处理高维数据且运算速度快。基于此,本文使用 RF 算法构建模型提取并研究。该模型构建时需要考虑 2 个重要的超参数[19]:RF 中树的总个数(N_{tree}值设置为 1 000)和变量节点数 M_{try},本文设置 M_{try}值为 $2*\sqrt{p}$,其中 p 为模型训练中的特征数量[20]。其预测结果是基于每一棵决策树的预测结果投票确定,本文以预测结果为沟谷的决策树个数占总数量 N_{tree}的比例作为某一像元被预测为沟谷的概率。

1.3.2　支持向量机　支持向量机(support vector machine,SVM)通过引入核函数将低维线性不可分数据映射到高维空间中,借助不同类域边缘的支持向量寻找最优分类超平面,隔离不同类别样本数据[21]。该方法在处理高维数据集上,泛化能力强,适用于本文研究的二分类问题[22]。本文选取适用性最强的径向基函数(radial basis function,RBF)进行研究,其模型为[23]:

$$K(X_i,X_j)=\exp(-\text{gamma}\parallel X_i-X_j\parallel^2) \quad (1)$$

式中:X_i和 X_j为低维空间中的特征向量;$\parallel X_i-X_j\parallel$为欧氏距离;gamma>0 为径向核参数。除此之外,惩罚系数(C)为对误差的宽容度,C 值过大导致模型出现过拟合,过小则出现欠拟合,合适的 C 值对预测结果影响较大[24]。本文将参数 gamma 和惩罚系数(C)取值设置为 $2^{-10}\sim2^{10}$,以 2 倍为间隔依次取值,采用网格搜索结合 5 折交叉验证方法确定最优参数[21]。模型预测结果是对高维空间中每个样本到分类超平面的距离进行 Sigmoid 压缩,得到每个像元被预测为沟谷的概率。

1.3.3　人工神经网络　人工神经网络(artificial neural network,ANN)是模仿大脑神经网络结构和功能而建立的一种数学模型,只要参数选取合适并且数据训练足够多,就能很好地拟合非线性问题。其网络结构包括网络层数以及输入、输出和隐藏层个数,可以表示为[25]:

$$y_m=f(\sum w_{ml}x_l+b_m) \quad (2)$$

式中:x_1和 y_m分别为输入因子和输出因子;w_{ml}、b_m和 f 分别为权重因子、偏差项和激活函数。已有研究[26]证明,1 个 3 层神经网络模型(隐藏层只含 1 个全连接层),可以逼近任意非线性函数。对于模型参数的选取,输入层神经元个数为最优特征的个数,输出层施加 Logistic 函数得到像元被预测为沟谷像元的概率,隐藏层的最优神经元个数采用试错法确定,其取值范围依据经验公式(3)确定,取 50 次平均值作为最后的评价精度确定最优参数。

$$m=\sqrt{l+n}+a \quad (3)$$

式中:m 为模型隐藏层神经元个数;l 为模型输入层神经元个数;n 为输出层神经元个数;a 为 1～10 的常数。

1.4　区域特征

1.4.1　特征初选　本文基于 DEM 数据和 Landsat 8 OLI 影像数据获取训练区域和测试区域特征,包括光谱

特征、植被特征、地形特征和纹理特征共26个特征，光谱特征为Landsat8 OLI影像预处理输出的前7个波段，同时基于波段运算得到4个植被指数(表1)。

地形特征是沟谷提取研究中的重要特征。高程、坡度等都特征对区域植被和降水量造成影响，进一步影响沟谷空间分布。因此，本文选取高程(H)、坡度(S)作为研究沟谷分布的特征。除此之外，还基于DEM数据获取正负地形(positive or negative terrains，PNT)、汇流累积量(flow accumulation area，FAA)和距离河流距离(d_r)3个特征，其中正负地形(PNT)反映的是地貌相对于周围地貌的相对高低情况，由于沟谷地貌下切明显，与周边坡面地貌具有较明显的高低落差，因此该指标适合于沟谷提取的研究。

纹理信息可以在一定程度上提高分类精度，参照侯蒙京等[27]的方法，本文利用灰度共生矩阵(grey level co-occurrence matrix，GLCM)计算每个波段的8种纹理特征(均值Mean、方差Variance、同质度Homogeneity、对比度Contrast、非相似性Dissimilarity、熵Entropy、角二阶矩ASM和相关性Correlation)得到56个纹理特征，利用主成分分析降维(PCA)，选取前10个主成分(GLCM1－GLCM10)作为本文研究的纹理特征。各特征信息描述见表1。

表1 分类特征

特征类型	特征名称	名称缩写	计算方法或描述
光谱特征	波段		依次为Coastal(海岸波段)、Blue(蓝波段)、Green(绿波段)、Red(红波段)、NIR(近红外波段)、SWIR1(短波红外1)和SWIR2(短波红外2)
植被特征	归一化植被指数	NDVI	$\frac{NIR-Red}{NIR+Red}$
	增强植被指数	EVI	$2.5*(\frac{Red-Green}{Red+6*Green-7.5*Coastal+1})$
	差值植被指数	DVI	$Red-Green$
	比值植被指数	RVI	$\frac{Red}{Green}$
地形特征	高程	H	与DEM数据一致
	坡度	S	基于DEM数据获取
	正负地形	PNT	由DEM中每个栅格像元值和其5×5栅格范围内像元均值的差值计算获得
	汇流累积量	FAA	基于D8单流向算法确定
	距离河流距离	d_r	由FAA＞300生成区域河网，计算得到每个像元至河网的欧氏距离
纹理特征	灰度共生矩阵	GLCM	GLCM1－GLCM10

1.4.2 特征筛选 相关研究[1]表明，沟谷分布与植被特征存在密切关系。因此为探讨沟谷和遥感影像月份之间的关系，本研究以基于1—12月每个月份的Landsat8 OLI影像数据和DEM数据获取的特征作为12组特征集。本文所选3种方法(RF、SVM、ANN)中仅RF算法具备特征筛选的能力[20]，因此采用RF算法确定沟谷提取的最优月份遥感影像和特征子集，具体过程为：

(1)以单月份为例，将1.1节获取的样本数据集按照7∶3比例随机划分为训练集和测试集，基于训练集和26个特征建立随机森林模型，依据平均准确率减少(mean decrease accuracy，MDA法)，对特征的重要性进行排序，该过程重复50次，统计每个特征排名在最后一位的频率；

(2)删除频率最高的特征，剩余特征进行下一轮筛选，再删除排名最后1位的特征，以此类推，直至剩下最后2位特征；同时，为降低模型运算时间，提高工作效率，综合考虑特征个数和分类精度，选取最优特征子集；

(3)循环上述步骤，分别得到1—12月具有最优组合的特征子集。

1.5 评价指标

1.5.1 模型评价 为评价具有最优参数的RF、SVM、ANN 3种模型沟谷提取性能，本文选取AUC值、Kappa系数、准确率、精确率、召回率和F1分数6种指标，在测试集上进行精度评价。选取的ROC曲线下面积(area under the curve，AUC)是衡量二分类预测效果的综合性指标，常用该指标比较不同算法构建的分类模型性能[23]；其余5种指标基于二分类混淆矩阵获取，具体计算公式为：

$$\text{Kappa}=\frac{TN+TP-Q}{TN+TP+FP+FN-Q} \tag{4}$$

$$Q=\frac{(TP+FN)*(TP+FP)+(TN+FN)*(TN+FP)}{TN+TP+FP+FN} \tag{5}$$

$$\text{准确率}=\frac{TP+TN}{TN+TP+FP+FN} \tag{6}$$

$$\text{精确率}=\frac{TP}{TP+FP} \tag{7}$$

$$\text{召回率}=\frac{TP}{TP+FN} \tag{8}$$

$$\text{F1 分数}=\frac{2*\text{精确率}*\text{召回率}}{\text{精确率}+\text{召回率}} \tag{9}$$

式中:TP 为实际为沟谷且预测为沟谷的像元数量;FP 为实际为坡面且预测为沟谷的像元数量;FN 为实际为沟谷且预测为坡面的像元数量;TN 为实际为坡面且预测为坡面的像元数量。最后,本文选取 6 种评价指标在 50 次预测的平均值来评价模型性能。通常情况下,评价指标数值越高,表示模型性能越好。

1.5.2　*精度评价*　沟谷提取定量分析是评价模型预测结果的关键。本文在测试区域内均匀选取 6 个 1 $km^2\times1$ km^2验证样区,基于 Google Earth Pro 影像数据人工识别沟谷地貌,结合 RF、SVM 和 ANN3 种方法沟谷提取结果,从沟谷和坡面地貌的生产者精度、用户精度和总体精度 3 个方面对不同模型的沟谷提取结果进行定量评价,具体计算公式分别为:

$$\text{生产者精度}=\frac{GG}{GG+NG} \tag{10}$$

$$\text{用户精度}=\frac{GG}{GG+GN} \tag{11}$$

$$\text{总体精度}=\frac{GG+NN}{GG+NN+GN+NG} \tag{12}$$

式中:GG 为实际面积为沟谷地貌,预测面积为沟谷地貌;GN 为实际面积为坡面地貌,预测面积为沟谷地貌;NG 为实际面积为沟谷地貌,预测面积为坡面地貌;NN 为实际面积为坡面地貌,预测面积为坡面地貌。其中,公式(10)和公式(11)中的 GG 换为 NN,NG 换为 GN,GN 换为 NG,即为坡面地貌的生产者精度和用户精度。

2　结果与分析

2.1　最优特征子集

基于 DEM 数据和 1—12 月遥感影像的特征集建立 RF 模型(图 2),随着特征的依次剔除,模型预测准确率呈现出先平缓后降低的变化趋势。以 12 月为例,特征集个数从 26 变为 8 的过程中,其准确率在 88.63%上下浮动,分类特征数从 7 开始,精度呈现较快的下降趋势。因此,最优分类结果的特征数被选定为 8,最优特征子集为 Red、Blue、H、SWIR1、PNT、Coastal、GLCM4 和 NIR,按此方法依次获取 1—12 月最优特征子集做下一步研究。

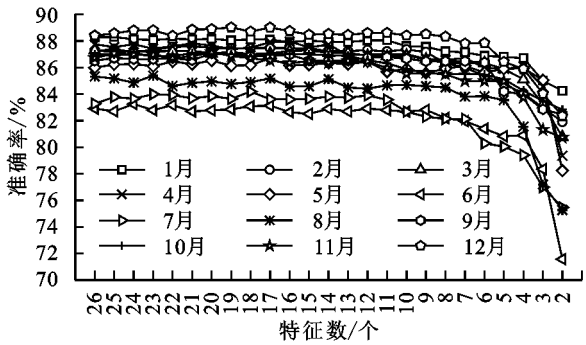

图 2　特征数与分类精度关系

2.2　不同月份 RF 模型精度对比

基于不同时期的最优特征子集建立 50 次 RF 模型,由表 2 可知,所有分类结果中,1—12 月测试集的准确率平均值和 Kappa 系数都呈现出先降低后增加的趋势,其中平均准确率最高的 3 个月份分别为 12,1,2 月,最低的 3 个月份分别为 6,7,8 月。基于 DEM 数据和 12 月遥感数据的最优特征子集平均准确率和 Kappa 系数最高,依次为 88.33%和 0.767,相比于准确率平均值和 Kappa 系数最低的 6 月,分别提高 5.52%和 0.111,具体特征子集见 2.1 节。可以发现,相比于夏季时期,冬季时期的影像表现出更好的分类效果。

表 2　不同月份最优特征子集精度

指标	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
准确率/%	88.050	87.160	86.910	86.720	85.950	82.810	83.570	84.490	86.380	86.880	86.430	88.330
Kappa 系数	0.760	0.743	0.738	0.734	0.719	0.656	0.672	0.690	0.726	0.737	0.727	0.767

2.3　模型参数寻优

由图 3a 可知,SVM 模型中,随着参数 gamma 和惩罚因子 C 交叉组合的不同,SVM 模型的预测准确率相差最大值可达到 25.84%,对比不同的参数组合,最终得出 SVM 模型最优参数组合 C 为 0.5,gamma 为 0.5,最高精度为 86.56%。由图 3b 可知,ANN 模型中,随着隐藏层个数的变化,模型准确率平均值在 88%左右浮动,当隐藏层个数为 9 时,模型准确率平均值达到相对最大值 88.45%,确定其为 ANN 模型最优参数。

2.4　模型精度与预测结果

2.4.1　*模型验证精度*　表 3 为 3 种模型测试集沟谷地貌的提取精度,结果表明 RF 和 ANN 模型的 AUC 值均达到 0.950 以上,Kappa 系数达到 0.760 以上,其余指标都在 0.86 以上,两者精度指标值均高于 SVM 模型(精确率除外)。

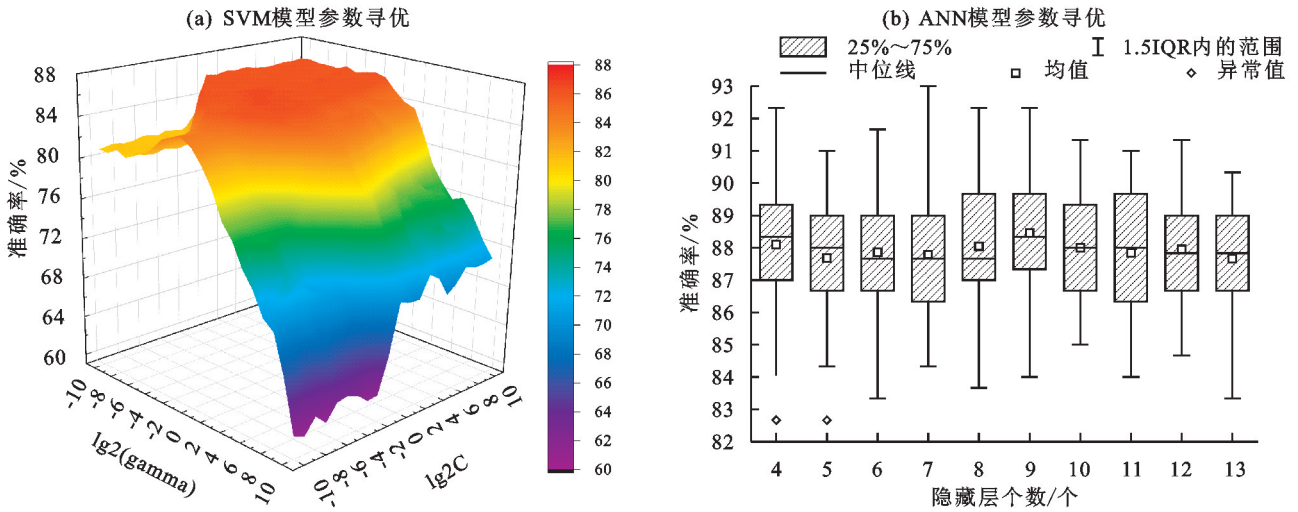

图 3 模型参数寻优

表 3 3 种模型的提取精度对比

方法	AUC 值	Kappa 系数	准确率/%	精确率	召回率	F1 分数
RF	0.952	0.767	88.33	0.894	0.864	0.882
SVM	0.912	0.724	86.56	0.885	0.832	0.854
ANN	0.953	0.769	88.45	0.877	0.882	0.884

2.4.2 沟谷预测结果及分类 由图 4 可知,RF 模型估算的沟谷概率值为 0~99.99%,均值为 52.91%;SVM 模型估算的沟谷概率为 0.08%~97.27%,均值为 53.81%;ANN 模型估算的沟谷概率为 0~100.00%,均值为 42.25%。从沟谷地貌预测结果来看,RF 模型预测的沟谷面积占比为 56.91%,SVM 模型预测结果为 50.84%,而 ANN 模型预测结果仅为 42.65%,然而三者提取的沟谷空间分布具有一定的一致。

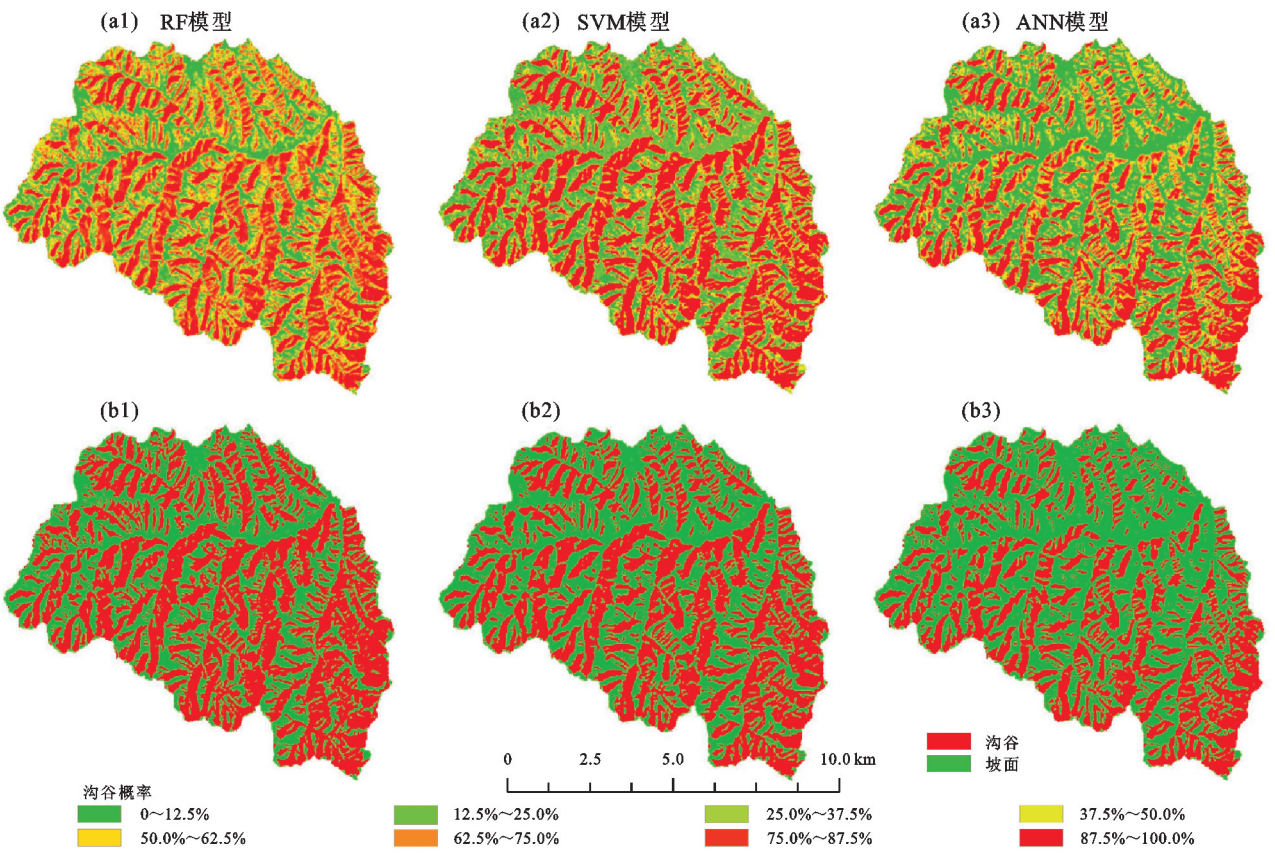

图 4 3 种模型估算沟谷概率

2.4.3 模型预测精度

(1)定量评价。3 种模型在 6 个验证样区的预测精度指标见图 5,从沟谷和坡面提取的平均生产者精度、平均用户精度和总体精度 3 个角度来看,RF 模型预测结果均高于 SVM 模型和 ANN 模型(第 4 个样本的平均生产者精度除外)。其中 RF 模型 6 个样本的总体精度依次为 81.31%,79.62%,83.12%,82.07%,79.39%,77.37%,平均值为 80.48%,高于 SVM 模型的 76.48%和 ANN 模型的 71.85%。总体来说,相比于 SVM 模型和 ANN 模型,RF 模型对沟

谷地貌提取的精度更高。

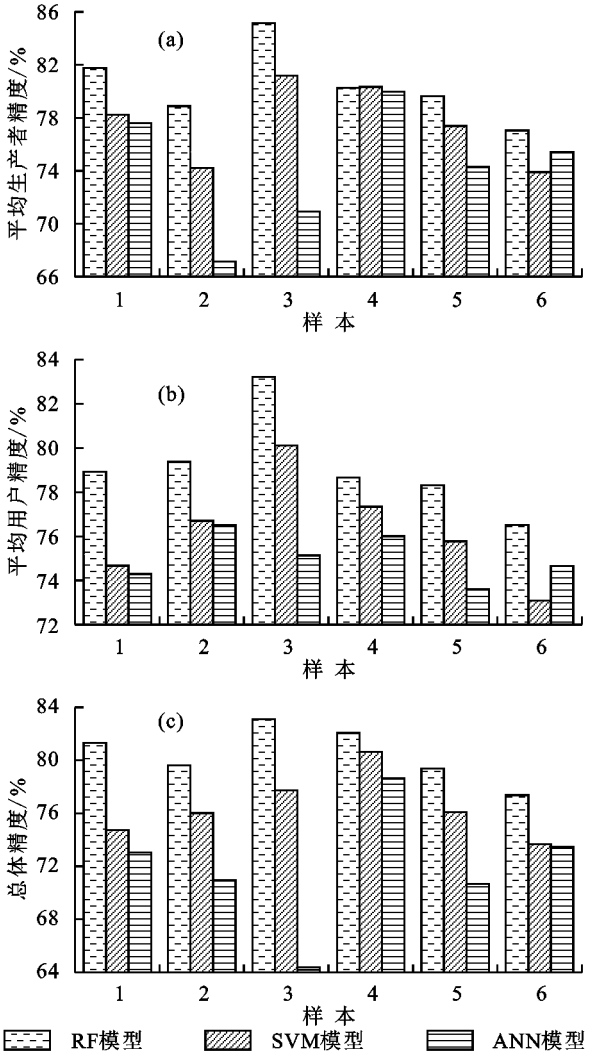

图 5　不同模型沟谷和坡面提取精度定量对比

(2)定性评价。图 6 为样本 1 基于 12 月重要性排名前 3 位的特征,可以看出,人工提取沟沿线在 NIR、CLCM4 和 Coastal 特征能够很好地区分沟谷和坡面地貌,沟谷地貌影像值普遍低于坡面地貌。

图 7 为基于 Google Earth 影像在验证样区对沟谷提取结果及其空间分布的定性评价。与 6 块验证样区的人工解译的矢量格式沟谷区域进行对比。通过进一步分析,基于 RF 模型的沟谷识别效果较好,沟谷轮廓与人工提取一致性较高,像元错分率较低。在 SVM 模型预测结果中,可以明显看到有部分沟谷被错分为坡面,从而导致测试区域的沟谷提取结果相比 RF 模型较差;在 ANN 模型预测结果中,这种现象更加明显,存在大量沟谷区域无法被正确分类,相比前 2 种模型提取结果最差。因此,综合考虑对 RF、SVM 和 ANN 3 种模型沟谷提取的定量和定性评价结果,RF 模型在各方面表现出更好的提取效果,表明该模型在沟谷提取问题上有更强的适用性。

3　讨 论

从影像优选结果来看,基于冬季遥感影像建立的沟谷提取模型效果最好,这与相关研究[28]结果一致,原因是冬季植被凋落导致植被覆盖率低,对沟谷提取影响较小。通过对 12 月特征集多个特征筛选发现,光谱特征对沟谷提取模型有显著影响,数量占比最大,地形特征则为正负地形(PNT)和高程(H)。光谱特征作为遥感影像信息的直接反映,具有高分辨化的特点[29],在模型中效果表现最佳;同时黄土高原沟谷地内侵蚀量大,高程上相对凹陷,而坡面以水流侵蚀为主,侵蚀量相对来说较小,这种侵蚀差异造成地形正负表达上的不同[30]。

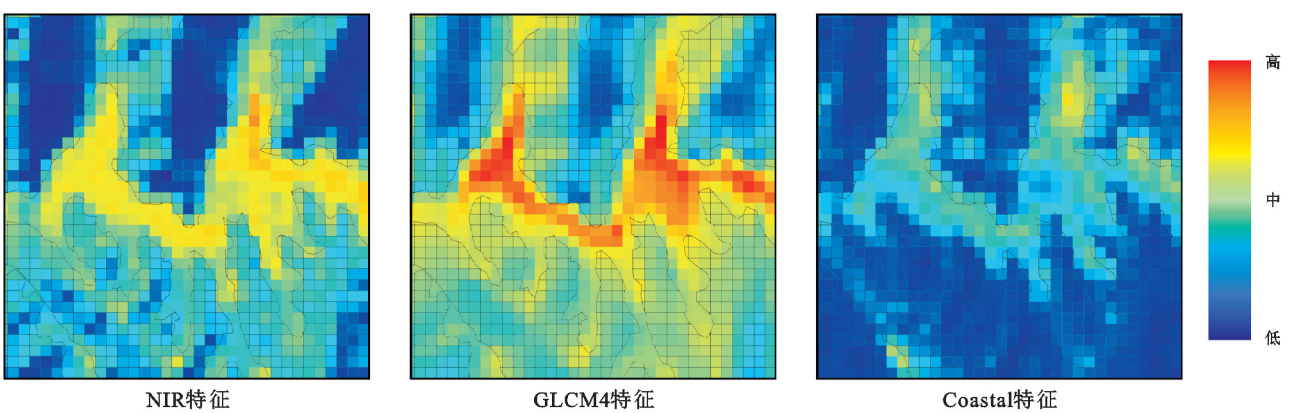

图 6　12 月重要性排名前 3 位

从验证样区沟谷提取结果来看,相比于 SVM 和 ANN 模型,RF 模型的建模方式更适合黄土高原流域沟谷提取,这主要是由于 RF 模型是一种以分类决策树为基分类器,将 Bagging 和随机特征选择结合起来的集成学习算法,其预测效果要优于单一分类算法的 SVM 和 ANN 模型。相关研究[22]表明,对于数据结构复杂和数据质量参差不齐的样本数据,集成学习算法通常优于单一分类方法。这更加验证本文基于不同模型进行沟谷提取结果的准确性。3 种模型在测试区域提取的沟谷空间分布基本一致,均呈现西北方向至东南方向逐渐集中的特征,与以往对延河流域沟谷地貌分布研究[18]结果一致。验证样区的定量分析结果表明,模型迁移过程中精度保持在 80.48%左右,与以往黄土高原沟谷区(15.4 万 km^2)

沟谷提取相关研究[15]对比，提高 1.68%，说明 RF 算法适合黄土高原沟谷地貌高精度提取研究，可大范围推广。因此可以基于 RF 模型对黄土高原其他流域的沟谷空间分布进行预测，模型反演结果能综合反映流域沟谷整体空间分布格局，可以满足大范围和高精度的沟谷提取。

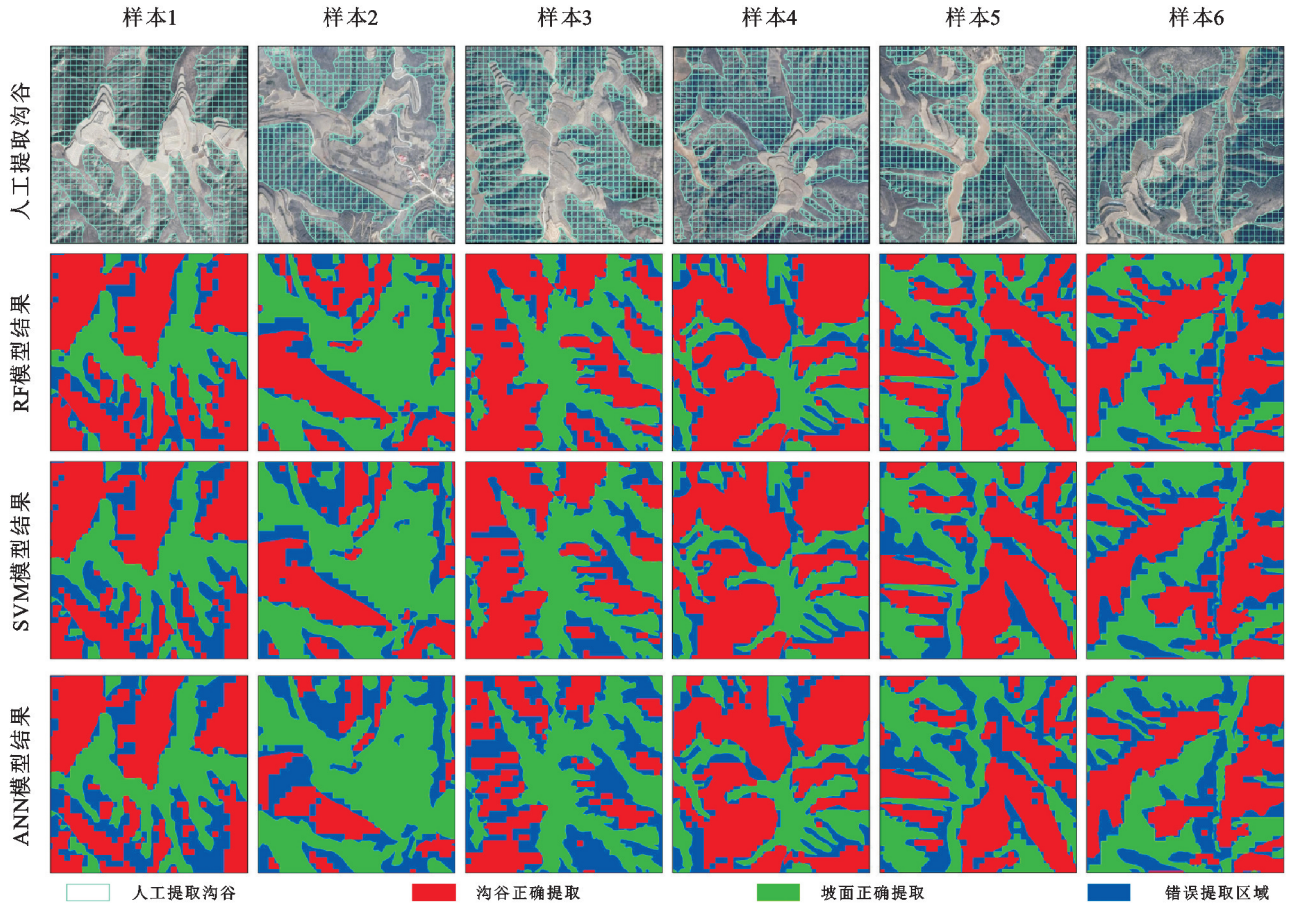

图 7 基于 Google Earth 影像对分类结果空间分布的验证

本文以像元为单位进行沟谷提取，对黄土高原典型流域像元单位沟谷概率进行预测和建模，取得较好的效果。但从空间分布结果来看，对于沟谷边界和一些分布零散的沟谷来说，基于 30 m×30 m 分辨率遥感影像进行研究的分类总体效果还有待提高，这是导致其错分率增加的主要原因之一。同时，从样本角度来看，Google Earth Pro 平台在研究区域提供的 0.3 m 分辨率的影像数据，能够满足沟谷像元识别的要求，但对大量样本解译时，需要花费大量时间，文中建立的模型仅使用 1 000个数据样本进行训练，容易使模型出现过拟合或欠拟合现象，比如，本文 ANN 模型中，测试集精度指标要高于 RF 模型和 SVM 模型，但从实际空间分布来看，ANN 模型效果差于 RF 模型和 SVM 模型，这主要是对样本数据的过度拟合限制其泛化性能。因此，在后续研究中，可以从 2 个方面进一步优化沟谷地貌提取和空间分布制图精度：(1)采用高分辨率影像数据(10 m 分辨率或 15 m 分辨率)获取沟谷特征信息，实现沟谷地貌的精细提取；(2)优化样本识别过程，基于 Google Earth Pro 平台寻找一种能快速获取大量样本信息的方法，解决模型过拟合问题，实现大尺度流域沟谷的高精度、自动化提取。

4 结 论

(1)基于不同时期遥感影像和 DEM 数据构建的沟谷识别 RF 模型，在测试集上准确率最高的 3 个月依次为 12，1，2 月，最低的 3 个月依次为 6，7，8 月，本研究表明，基于冬季影像获取的特征子集在黄土高原沟谷提取的问题上具有更强的优越性。

(2)使用 RF 模型对 12 月遥感影像和 DEM 数据的特征筛选结果表明，波段特征中 Coastal(海岸波段)、Blue(蓝波段)、Red(红波段)、NIR(近红外波段)、SWIR1(短波红外 1)重要性排名靠前；纹理特征中，主成分分析第 4 特征即 GLCM4 重要性排名靠前；地形特征中，高程(H)和正负地形(PNT)重要性排名靠前，这 8 种特征对模型贡献率最高，可为今后黄土高原沟谷提取研究中的特征选择提供一定的参考。

(3)结合最优特征子集和机器学习模型预测测试区域沟谷空间分布，3 种方法均表明测试区域沟谷分布呈现西北至东南方向逐渐集中的特征，说明利用机器学习模型预测沟谷及其空间分布的方法具备广泛的应用价值。通过对测试区域验证样区进行定量和定性分析，RF

模型总体精度最高,在沟谷提取中有更好的适用性和应用潜力,最适用于整个黄土高原沟谷地貌的提取。

参考文献:

[1] Zhao J L, Vanmaercke M, Chen L Q, et al. Vegetation cover and topography rather than human disturbance control gully density and sediment production on the Chinese Loess Plateau[J]. Geomorphology, 2016, 274(1):92-105.

[2] Zabihi M, Mirchooli F, Motevalli A, et al. Spatial modelling of gully erosion in Mazandaran Province, northern Iran[J]. Catena, 2018, 161:1-13.

[3] Yang X, Li M, Na J M, et al. Gully boundary extraction based on multidirectional hill-shading from high-resolution DEMs[J]. Transactions in GIS, 2017, 21(6):1204-1216.

[4] 王轲,王琤,张青峰,等.地形开度和差值图像阈值分割原理相结合的黄土高原沟沿线提取法[J].测绘学报,2015,44(1):67-75.

[5] 柳潇,吕新彪,吴春明,等.面向高空间分辨率遥感影像的山区地形校正方法[J].地球科学,2020,45(2):645-662.

[6] Dai W, Hu G H, Yang X, et al. Identifying ephemeral gullies from high-resolution images and DEMs using flow-directional detection[J]. Journal of Mountain Science, 2020, 17(12):3024-3038.

[7] Yang X, Dai W, Tang G A, et al. Deriving ephemeral gullies from VHR image in Loess Hilly Areas through directional edge detection[J]. ISPRS International Journal of Geo-Information, 2017, 6(11):e371.

[8] 陈靖涛,史明昌,罗志东,等.基于双向地形阴影法的黄土侵蚀沟自动提取技术[J].农业工程学报,2022,38(7):127-135.

[9] 李乐,马巍,勾蒙蒙,等.三峡库区典型流域硝态氮输出特征及归因分析[J].水土保持学报,2022,36(4):74-84.

[10] 王少丽,臧敏,王亚娟,等.年径流系数变化特征及预测模型研究[J].水土保持学报,2020,34(5):56-60,67.

[11] 李柳阳,朱青,刘亚,等.基于气象因子的长三角地区农田站点土壤水分时间序列预测[J].水土保持学报,2021,35(2):131-137.

[12] Vanmaercke M, Chen Y X, Haregeweyn N, et al. Predicting gully densities at sub-continental scales: A case study for the Horn of Africa[J]. Earth Surface Processes and Landforms, 2020, 45(15):3763-3779.

[13] Yang A N, Wang C M, Pang G W, et al. Gully Erosion susceptibility mapping in highly complex terrain using machine learning models[J]. ISPRS International Journal of Geo-Information, 2021, 10(10):e680.

[14] Ding H, Liu K, Chen X Z, et al. Optimized segmentation based on the weighted aggregation method for loess bank gully mapping[J]. Remote Sensing, 2020, 12(5):e793.

[15] Liu K, Ding H, Tang G A, et al. Large-scale mapping of gully-affected areas: An approach integrating Google Earth images and terrain skeleton information[J]. Geomorphology, 2018, 314:13-26.

[16] 贾云飞,李云飞,范天程,等.基于长时间序列 NDVI 的黄土高原延河流域及其沟壑区植被覆盖变化分析[J].水土保持研究,2022,29(4):240-247.

[17] 呼媛,鲁克新,李鹏,等.延河流域骨干坝拦沙量反推与未来可拦沙年限预测[J].水土保持学报,2021,35(3):38-45.

[18] 范天程,贾云飞,李云飞,等.基于遥感影像与逻辑回归模型的延河流域沟壑分布概率预测[J].水土保持研究,2022,29(4):316-321.

[19] 夏子书,白一茹,王幼奇,等.基于 GIS 和随机森林算法的宁东土壤饱和导水率分布与预测[J].水土保持学报,2021,35(1):285-293.

[20] 王飞,杨胜天,丁建丽,等.环境敏感变量优选及机器学习算法预测绿洲土壤盐分[J].农业工程学报,2018,34(22):102-110.

[21] 陈黔,李晓松,修晓敏,等.基于 Google Earth Engine 与机器学习的大尺度 30 m 分辨率沙地灌木覆盖度估算[J].生态学报,2019,39(11):4056-4069.

[22] 杨剑锋,乔佩蕊,李永梅,等.机器学习分类问题及算法研究综述[J].统计与决策,2019,35(6):36-40.

[23] 陈强.机器学习及 R 应用[M].北京:高等教育出版社,2020:314-315.

[24] 张鹏,马庆勋,吕杰,等.机器学习算法在森林地上生物量估算中的应用[J].测绘通报,2021(12):28-32.

[25] 杨丽萍,侯成磊,苏志强,等.基于机器学习和全极化雷达数据的干旱区土壤湿度反演[J].农业工程学报,2021,37(13):74-82.

[26] 李辉东,关德新,袁凤辉,等.BP 人工神经网络模拟杨树林冠蒸腾[J].生态学报,2015,35(12):4137-4145.

[27] 侯蒙京,殷建鹏,葛静,等.基于随机森林的高寒湿地地区土地覆盖遥感分类方法[J].农业机械学报,2020,51(7):220-227.

[28] Chen Y X, Jiao J Y, Wei Y H, et al. Accuracy assessment of the planar morphology of valley bank gullies extracted with high resolution remote sensing imagery on the Loess Plateau, China[J]. International Journal of Environmental Research and Public Health, 2019, 16(3):e369.

[29] 李斌兵,黄磊.基于面向对象技术的黄土丘陵沟壑区切沟遥感提取方法研究[J].水土保持研究,2013,20(3):115-119,124.

[30] Zhou Y, Tang G A, Yang X, et al. Positive and negative terrains on northern Shaanxi Loess Plateau[J]. Journal of Geographical Sciences, 2010, 20(1):64-76.