基于混合式特征选择的高分五号影像农田识别

陈珠琳[1,2]，贾坤[1,2]，李强子[3]，肖晨超[4]，魏丹丹[4]，赵祥[1,2]，魏香琴[3]，姚云军[1,2]，李娟[3]

1. 北京师范大学 地理科学学部 遥感科学国家重点实验室, 北京 100875;
2. 北京师范大学 北京市陆表遥感数据产品工程技术研究中心, 北京 100875;
3. 中国科学院空天信息创新研究院, 北京 100101;
4. 自然资源部国土卫星遥感应用中心, 北京 100048

摘 要：精准农田识别是农作物估产和粮食安全评估的基础。遥感数据作为农田识别的重要数据源，可提供动态、快速的监测结果。高光谱数据在农田识别分类方面具有巨大的应用潜力，但其中的冗余波段影响了分类效率和分类精度。因此，本研究提出了一种适用于高光谱数据农田分类的混合式特征选择算法。首先，基于变量的重要性排序或约束程度，按步长逐步进行降维；其次，寻找分类精度骤减的转折点，并将其对应的变量作为特征子集；最后，利用序列后向选择SBS（Sequential Backward Selection）方法搜索最优分类特征子集。本研究利用GF-5高光谱数据，共研究了3种降维方法（随机森林RF（Random Forest）、互信息MI（Multi-Information）和L1正则化（L1 regularization））和3种分类算法（随机森林、支持向量机SVM（Support Vector Machine）和K近邻KNN（K-Nearest Neighbor））的组合在农田分类中的表现。结果表明，基于L1正则化法得到的特征子集自相关性较低，并且包含的红边和近红外波段有效提高了农田、森林和裸土的区分度。在不同分类模型比较中发现，SVM在高维空间中表现出非常好的抗噪能力，分类精度高于RF和KNN。而RF在低维空间中的泛化能力要高于SVM和KNN。相比于第一步降维得到的特征子集，使用SBS搜索得到的最优特征子集均提高了分类精度。最终，具有23维输入的L1-SVM-SBS分类模型得到了最高的总体分类精度（94.64%）和农田召回率（95.83%）。本研究为高光谱数据特征优选提供了一种新思路，筛选出了更具代表性的特征波段，提高了农田分类精度，对高光谱遥感分类研究具有参考价值。

关键词：农田识别，高分五号，特征选择，高光谱遥感，L1正则化，后向序列选择

引用格式：陈珠琳,贾坤,李强子,肖晨超,魏丹丹,赵祥,魏香琴,姚云军,李娟.2022.基于混合式特征选择的高分五号影像农田识别.遥感学报,26(7): 1383-1394

Chen Z L, Jia K, Li Q Z, Xiao C C, Wei D D, Zhao X, Wei X Q, Yao Y J and Li J. 2022. Hybrid feature selection for cropland identification using GF-5 satellite image. National Remote Sensing Bulletin, 26(7): 1383-1394[DOI: 10.11834/jrs.20220458]

1 引 言

在全球极端气候、自然灾害肆虐的环境下，粮食安全成为影响社会稳定和可持续发展的重要因素（Yu等，2019）。中国因受退耕还林还草、城市扩张等因素的影响，耕地数量逐年减少，使得粮食生产面临严峻挑战（Han和Song，2019）。农田面积统计是作物估产的基础，因此快速、准确的农作物种植面积监测成为国家粮食宏观决策的重要支持（李强子，2018；樊东东 等，2019）。

遥感技术具有动态、快速且大面积观测的特点，在农田识别中发挥着重要作用（Liu等，2011；Xu等，2019）。目前，各类遥感数据已广泛应用于农田分类研究（Yin等，2018；Hao等，2019）。例如，Kussul等（2017）基于Landsat 8和Sentinel-1A遥感数据，使用卷积神经网络算法实现了土地利用类型分类，其中农田识别精度高于90%。Ghorbanian等（2020）基于Google Earth

收稿日期：2020-10-26；预印本：2021-05-21

基金项目：国家重点研发计划(编号:2019YFE0127300, 2016YFB0501404)；国家自然科学基金(编号:42171318)

第一作者简介：陈珠琳，研究方向为资源环境遥感。E-mail：chenzhulin@mail.bnu.edu.cn

通信作者简介：贾坤，研究方向为植被定量遥感，土地覆盖遥感分类和生态遥感。E-mail：jiakun@bnu.edu.cn

Engine（GEE）平台实现了伊朗地区10 m空间分辨率的农田识别研究，生产精度和用户精度分别为91%和90.2%。尽管如此，多光谱遥感影像的波段较少，无法避免地物间“异物同谱”的现象，从而影响了农田识别精度。相比之下，高光谱遥感数据可提供可见光到短波红外范围内的上百个窄波段信息，有潜力提高相似地物的分类精度，在农田识别方面有着巨大的研究价值（贾坤和李强子，2013；Yuan等，2020）。目前，高光谱遥感也越来越多的应用于农业遥感分类，例如，Aneece和Thenkabail（2018）利用EO-1 Hyperion高光谱数据，基于GEE平台实现了5种主要作物（玉米、大豆、冬小麦、水稻和棉花）的分类，各类型识别精度到达75%—95%。刘晓双等（2018）使用河南省平镇地区的Hyperion高光谱数据，提出了一种结合光谱、纹理和空间信息的多特征地类提取方法，该方法将农田分类精度提高到了90%。

虽然高光谱数据可以获取更详细的地物波谱特征，但连续波段间的高相关性造成了不可避免的维数灾难。所以，在构建分类模型之前需要对数据进行降维处理，从而降低模型的复杂度，提高计算效率，并减小过拟合的风险（Ding等，2020）。降维方法通常分为两类，一类是特征提取FE（Feature Extraction），即将原始数据进行空间变化，使生成的低维数据包含大多数原始信息，比如主成分分析法（Alvarez-Meza等，2017）。但该类算法所得到的变量解释性较差，并且筛选结果仍需要所有变量参与运算。第二类称为特征选择FS（Feature Selection），即在原始数据中按照某种原则选取最具代表性的变量（Sylvester等，2018），该类方法保留了数据的原始特性，使变量更具解释性。

从选择策略的角度，FS算法可分为过滤法（Filter）、包装法（Wrapper）和嵌入式法（Embedded）（Li等，2017）。其中，过滤法仅依赖于数据的特征来评估其重要性，尽管更加高效，但最终结果不一定对目标是最优的（Sánchez-Maroño等，2007）。包装法依赖于学习算法的预测性，能评估所选特征的质量，虽然精度高，但效率低，不适合处理高维数据（González等，2019）。嵌入式法是过滤法和包装法的一种权衡，它将特征选择嵌入到了模型学习中，包含了与学习算法的交互，又比包装法更有效。因此，嵌入式法也是目前研究中最常用的方法（Bolón-Canedo和Alonso-Betanzos，2019）。然而，虽然嵌入式方法依据某种特定的算法对变量的重要性进行排序，但是这种排序并不能代表目标的最优特征子集。另外，由于嵌入式法基于特定的分类算法，所以大部分研究选择同时使用该特定的算法进行降维和分类处理（如同时使用随机森林算法的变量排序和分类功能）。但目前尚未有研究证明使用同种算法进行降维和分类处理为最佳选择。

针对上述问题，本研究提出了一种混合式特征选择算法，综合过滤法、包装法和嵌入式法的特点，改善特征选择和分类效果。本文对比了3种特征选择方法（随机森林，互信息和基于L1正则化的方法）和3种分类算法（随机森林，支持向量机和K近邻）的组合，并通过序列后向选择方法优选特征子集，确定最佳分类模型，以期为高光谱数据的农田分类研究提供参考。

2 材料与方法

2.1 研究区概况

研究区坐落于吉林省长春市（43°05′N—45°15′N；124°18′E—127°05′E）。该地区位于中国东北平原中部，地势平坦，海拔为137—160 m。气候类型属于温带季风气候，年降水量为600—700 mm，水热条件适合农作物生长。该地区的主要农作物为玉米，根据GF-5影像（图1）以及同期Google Earth高清图像，将研究区分为5种类别：农田、森林、裸土、水体和不透水面。本研究采用目视解译选取样本点，其中农田、森林、裸土和不透水面各选取了4000个像元，水体选择了1500个像元。在总样本中随机选取2/3的样本点作为训练数据，剩余样本则为验证数据，其中各类型样本点的个数如表1所示。

2.2 高光谱数据及预处理

GF-5卫星是中国高分重大专项中唯一一颗高光谱卫星，其空间分辨率为30 m，幅宽为60 km，共有330个波段（400—2500 nm），其中可见光和短波红外部分的光谱分辨率分别为5 nm和10 nm（董新丰等，2020）。与Hyperion（224个波段，光谱分辨率为10 nm）相比，GF-5高光谱数据不仅提

供了更详细的谱段信息，同时具有更高的信噪比（刘银年 等，2020）。本研究获取的GF-5高光谱数据成像于2019年7月4日，覆盖面积为3760 km^2。数据获取后，首先根据同一区域的Sentinel-2影像对其进行几何校正，然后根据每个波段的辐射定标系数进行辐射定标，最后进行FLAASH大气校正，得到地表反射率数据。由于影像存在无效波段，经剔除后共保留295个有效波段。

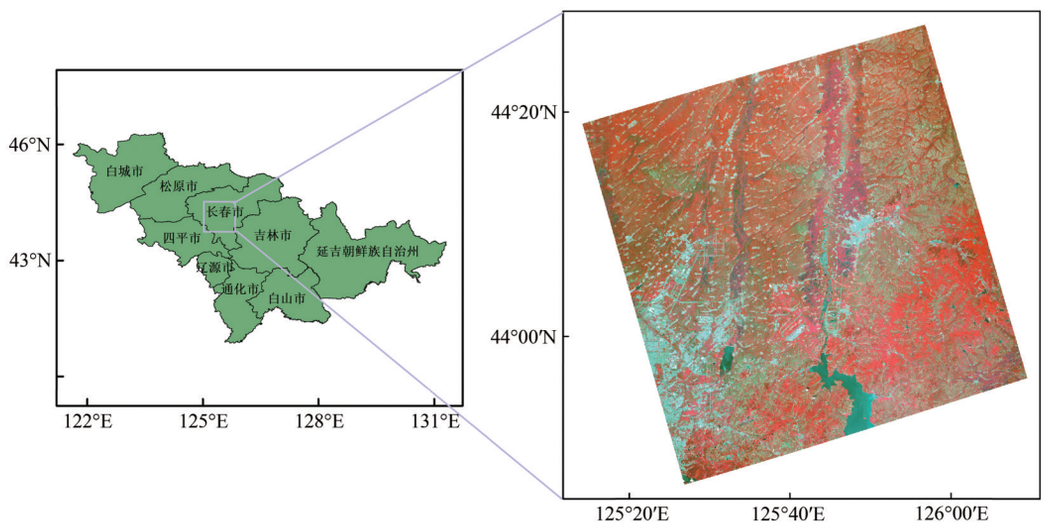


图1 研究区地理位置及高分五号伪彩色合成影像图
（R：Band 107（843.768 nm）；G：Band 65（664.115 nm）；B：Band 41（561.461 nm））

Fig. 1 Location of study area and false color image of GF-5 hyperspectral data
（R：Band 107（843.768 nm）；G：Band 65（664.115 nm）；B：Band 41（561.461 nm））

表1 训练集和验证集中样本的类型及个数

Table 1 Sample type and sample number in training dataset and validation dataset

/个

数据集	类型				
	森林	农田	水体	裸地	不透水面
训练集	3016	3002	1029	2984	2986
测试集	984	998	471	1016	1014

2.3 研究方法

图2为本研究技术路线图。首先，对高光谱数据进行预处理，得到地表反射率数据，然后采用目视解译法选择训练和验证样本。结合不同的分类算法，分别使用随机森林RF（Random Forest）、互信息MI（Multi-Information）和基于L1正则化（L1 regularization）的方法进行第一次降维，得到3个特征子集。然后，再结合分类算法与序列后向选择法SBS（Sequential Backward Selection）进行第二次降维，得到9种分类模型，通过比较其验证集的分类精度，优选出最佳分类模型。

2.4 混合式特征选择算法

本文提出的混合式特征选择方法结合两类算法，既弥补了过滤法或嵌入式算法无法搜索最优解的缺陷，又解决了包装法处理高维数据时表现出的低效率特征。该算法首先使用具有特征排序功能或约束功能的算法（过滤法或嵌入式法），得到变量的重要性排序或约束后的特征子集。然后从特征全集开始，并按照一定的步长减小特征个数（即去除相关性较小的变量），绘制分类精度变化曲线。寻找该曲线中分类精度急剧降低的突变点，将该点所对应的特征作为特征子集。最后，使用后向序列选择算法搜索最佳特征子集，得到最终分类结果。

在混合式特征选择算法中，特征排序以及使用的分类方法是影响分类精度的两大因素。因此本研究讨论了3种具有排序或约束功能的特征选择方法和3种分类算法。

2.4.1 特征选择方法

本研究选择随机森林、互信息和L1正则化方法对原始数据进行降维处理。其中互信息法是一种过滤式特征选择算法，随机森林和L1正则化法均属于嵌入式算法，但二者所基于的原理不同（随机森林根据每个特征变化后对结果的影响程度

判断其重要性；而L1正则化方法则通过向成本函数中添加L1范数，将系数矩阵稀疏化，从而达到降维目的），因此，3种不同原理的算法被选用比较特征选择的效果。

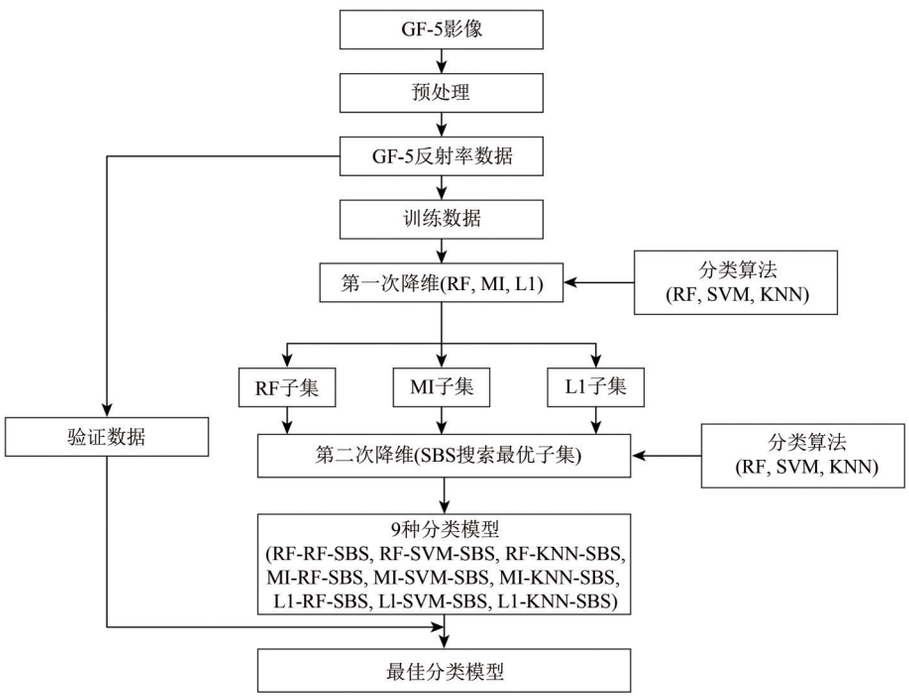


图2 本研究技术路线图

Fig. 2 Flow chart of this study

（1）随机森林。随机森林是一种基于集成思想的机器学习算法，集成了多棵决策树的分类结果，具有精度高、稳定性强、不受高维空间影响的特点，被广泛应用于分类问题（Htitiou等，2019；Yoo等，2019）。在随机森林算法中，每棵决策树的袋外数据OOB（Out Of Bag）可以计算得到一个误差率，用来评价预测因子的重要性，具体步骤如下（Liu和Sun，2019；Wang等，2019）：首先，选择相应的袋外数据，计算每棵决策树的袋外数据误差err_{OOB1}；其次，随机对袋外数据所有的样本特征X加入噪声进行干扰，再次计算袋外数据误差，记为err_{OOB2}；最后，若随机森林中共有N棵决策树，则特征X的重要性值为：

$$X_{\mathrm{importance}}=\left(\sum_{i=1}^{N}err_{\mathrm{OOB2}}-err_{\mathrm{OOB1}}\right)/N \tag{1}$$

（2）互信息法。互信息法属于一种过滤算法，此类算法独立于任何模型，仅依赖于数据的统计学特征来评估变量的重要性，主要包括两步（Li等，2019；Radley等，2020）：首先，根据特征评价标准对重要性进行排序，然后选择合适的阈值或者变量个数，剔除低阶特征。互信息是一种信息度量，代表了一个变量中包含另一个变量的信息量，用于判断变量之间的相关性。假设两个随机变量X和Y的联合分布为$p(x,y)$，边缘分布分别为$p(x)$和$p(y)$，则互信息$I(\boldsymbol{X},\boldsymbol{Y})$是联合分布$p(x,y)$与边缘分布$p(x)$、$p(y)$的相对熵：

$$I(\boldsymbol{X},\boldsymbol{Y})=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{2}$$

（3）L1正则化法。高维输入不仅增加了模型的复杂度和不稳定程度，也增加了过拟合的风险。而正则化作为一种回归形式，通过约束模型的权重来产生稀疏权值矩阵，降低多项式的阶数，最终模型中系数不为0的变量则为重要变量。对于给定数据集$\boldsymbol{D}=\left\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\right\}$，基于L1正则化的嵌入式特征选择模型可以表示为（Park和Hastie，2007；Zhou等，2017）：

$$\min_{\boldsymbol{w}}\sum_{i=1}^{m}(y_i-\boldsymbol{w}^{\mathrm{T}}x_i)^2+\lambda\|\boldsymbol{w}\|_1 \tag{3}$$

式中，$\boldsymbol{w}$为线性拟合的系数向量，λ为正则化参数，$\|\boldsymbol{w}\|_1$为系数向量的一阶范数。

本研究使用线性SVM作为基础模型，通过添加L1惩罚项进行特征选择，其中惩罚系数C的搜

索范围为［0.01，0.1］，步长为0.01。

（4）序列后向选择法。序列后向选择法是一种贪心算法，属于包装式特征选择法。该算法首先需要确定分类算法，然后从总特征（个数为n）开始，遍历所有$n-1$个特征组合，选择分类精度最高的组合进行下一步遍历，直到将变量个数缩减到最小（即n=1）。最后对比每次遍历得到的结果，找到分类精度最高的组合，即得到最佳特征子集。当面对高维数据时，序列后向选择算法的运算量大，因此有必要先用滤波法或嵌入法来寻找一个低维度的特征集合，然后利用序列后向选择算法得到最优特征子集。

2.4.2 分类方法

本研究选择了3种常用的机器学习分类算法，包括随机森林分类、支持向量机分类和K近邻分类算法。其中随机森林分类算法和支持向量机分类算法分别为两种嵌入式特征选择算法的基础模型，而K近邻分类算法则独立于3种降维方法。本研究选择上述3种分类算法的目的是为了检验基于同种算法进行降维和分类是否为最优策略。

（1）随机森林分类算法。随机森林分类器是由多棵决策树分类器的投票结果所决定，基本流程为（董超和赵庚星，2020；刘杰 等，2020）：1）在训练样本中，通过boosttrap方式有放回的抽取N个训练样本集；2）使用N个样本集构建N棵决策树模型，在每棵决策树中，从M个特征向量中随机选择m（m<M）个特征用于内部节点划分，节点分裂标准采用基尼指数（Gini）；3）对N棵决策树的分类结果进行集成，采用投票的方式确定最终分类结果。在随机森林算法中，决策树的个数N以及节点分裂时输入的特征个数m影响着模型的分类精度和运算速度。本研究使用网络搜索交叉验证进行参数寻优，最终N和m取值分别为300和0.75。

（2）支持向量机分类算法。支持向量机是由Vapnik提出的一种基于核的机器学习模型，由于其严格的数学理论支持、良好的泛化以及最优数值求解能力，在遥感分类领域具有广泛的应用（Maldonado 和 López，2018；魏友华 等，2020）。支持向量机直接从训练数据中确定决策函数，用于寻找一个超平面对样本进行分割。在非线性情况下，支持向量机通过核函数将样本映射到高维空间，构建一个最优分类超平面，最终转化为一个凸二次规划问题进行求解。支持向量机可以在高维小样本情况下凸显其卓越的泛化能力，得到高精度的分类结果（Sukawattanavijit 等，2017；Yan和Jia，2018）。支持向量机的核函数选用径向基核函数，其中的参数gamma和C使用网络搜索交叉验证进行参数寻优。

（3）K近邻分类算法。K近邻是一种经典的机器学习分类方法。该算法简单高效，分类性能显著，常用于数据挖掘和统计学领域（Zhang，2020）。与其他监督分类算法不同，K近邻法不需要样本进行训练，通过将已知类别的样本作为参照，根据未知样本与所有已知样本之间的距离对其进行归类。具体流程如下：首先，对数据进行归一化处理，计算未知样本与所有已知样本之间的距离或相似度，找到与未知样本最近的K个已知样本；其次，根据已知样本所属的类别判断未知样本的归属，若K个已知样本均属于类别A，则未知样本也将被归为A类；若K个已知样本不属于同一类别，则根据少数服从多数的原则，将未知样本归为占比例最高的类别。该算法关键参数K采用网络搜索交叉验证进行参数寻优，搜索空间设置为［2，20］。

2.5 精度验证方法

本研究使用总体分类精度和召回率进行分类精度验证与评价。总体分类精度表示所有正确预测样本占总样本的比重，召回率代表某一类别里，所有样本中被正确预测的比重。其中总体分类精度作为降维的评价标准，总体分类精度和召回率作为农田分类模型的评价标准。

3 结果与分析

3.1 不同模型随维度降低的总体分类精度变化

在3种特征选择算法中，随机森林和互信息法可以得到所有变量的重要性值，并以此作为基准对变量进行排序。所以，对于随机森林和互信息法，本研究采用维度逐步递减的方式，以10维为步长，分析不同分类算法处理下总体分类精度的变化趋势，并确定特征子集（图3和图4）。而基于L1正则化的特征选择方法通过定义惩罚系数C的

大小得到筛选结果，变量的个数并不固定。所以本研究根据惩罚系数C的搜索范围与步长分析其总体分类精度变化趋势，从而确定特征子集（图5）。

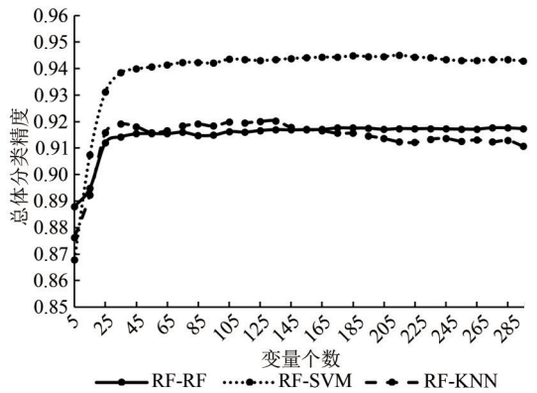


图3 随机森林排序下的不同模型的总体分类精度变化趋势

Fig. 3 Trend of overall accuracy of different models under random forest sorting

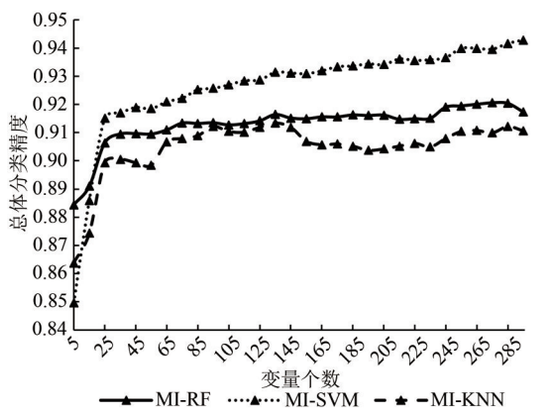


图4 互信息排序下的不同模型的总体分类精度变化趋势

Fig. 4 Trend of overall accuracy of different models under multi-information sorting

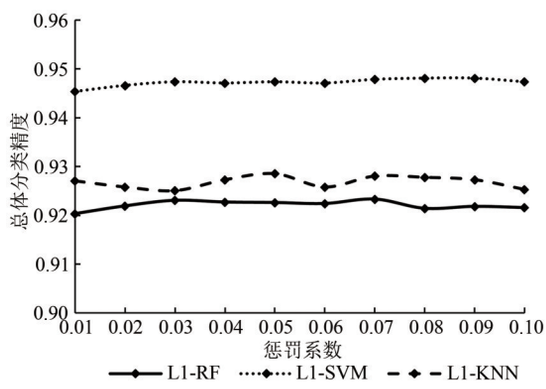


图5 L1正则化约束条件下的不同模型的总体分类精度随惩罚系数的变化趋势

Fig. 5 Trend of overall accuracy of different models under L1 regularization sorting

在随机森林特征重要性排序（图3）中，变量超过15维的情形下，RF-SVM模型得到的分类精度始终保持最高，而RF-RF和RF-KNN表现出较为复杂的现象：（1）当输入变量的维度大于155维时，RF-RF的总体分类精度大于RF-KNN；（2）而当维度低于155维但高于15维时，则出现相反的趋势。当输入变量个数仅为5维时，RF-RF在3种模型中表现出最高的精度，而RF-SVM则表现出较低的精度。根据上述精度变化可以得出，RF的泛化能力较强，在较少的输入变量条件下仍可以保持较高的分类精度，而SVM在处理高维数据时更有优势。当变量个数大于35维时，KNN的分类精度随维度的降低出现升高，则说明该算法对冗余数据的鲁棒性较差。

根据图3表现的整体趋势可以看出，维度在295—25的区间内，3种分类方法的总体分类精度变化较小；当变量个数小于25维时，总体分类精度出现大幅度的降低。所以，本研究将随机森林排序中重要性排名为前25的变量作为特征子集。

在互信息排序（图4）中，变量维度从295降低到5的过程中，MI-RF、MI-KNN和MI-SVM模型的总分类精度分别降低了3.29%，4.68%和9.32%。当变量维度大于25时，MI-SVM的精度均高于MI-RF和MI-KNN，而随着维度持续降低，MI-SVM的总体分类精度出现大幅度下降。当输入变量个数为5维时，MI-SVM得到的总体分类精度低于MI-RF和MI-KNN。相比之下，MI-RF的总体分类精度变化最小，并且在输入变量个数较少的情况下仍然保持较高的精度 。虽然MI-KNN在维度降低过程中表现出的变化趋势并不稳定（先降低后升高再降低），但在维度降至25维之前，总体分类精度仅降低1%，说明其变化幅度较小。因此，本研究将互信息排序下重要性值最高的前25个变量作为MI-RF、MI-SVM和MI-KNN的特征子集。

在正则化约束条件（图5）中，惩罚系数的变化对总体分类精度的影响很小，每种模型的精度最大差值分别为0.30%（L1-RF），0.35%（L1-KNN）和0.27%（L1-SVM）。当惩罚系数取值为0.01时所对应的变量个数为27，此时模型的总体分类精度分别为92.03%（L1-RF），94.54%（L1-

SVM）和92.70%（L1-KNN）。与降维前得到的结果相比，精度分别提高了0.30%，0.25%和1.64%。因此，本研究将惩罚系数取值为0.01时得到的特征变量作为特征子集。

3.2 不同模型的特征子集及其自相关性

表2为降维前以及第一次降维处理后不同模型所对应的总体分类精度。从中可以看出，使用L1正则化方法处理后的模型（L1-RF、L1-SVM和L1-KNN），分类精度均有提高，而MI在降维后的模型精度均出现下降。RF仅在使用KNN分类模型支持下精度有所提高（提高0.52%）。分类精度提高的模型分别为RF-KNN、L1-RF、L1-SVM和L1-KNN，其中L1-KNN模型精度提升最大，为1.64%，而L1-SVM模型的精度提升最小，为0.25%。尽管如此，L1-SVM模型总分类精度仍然是所有模型中最高的，达到94.54%。

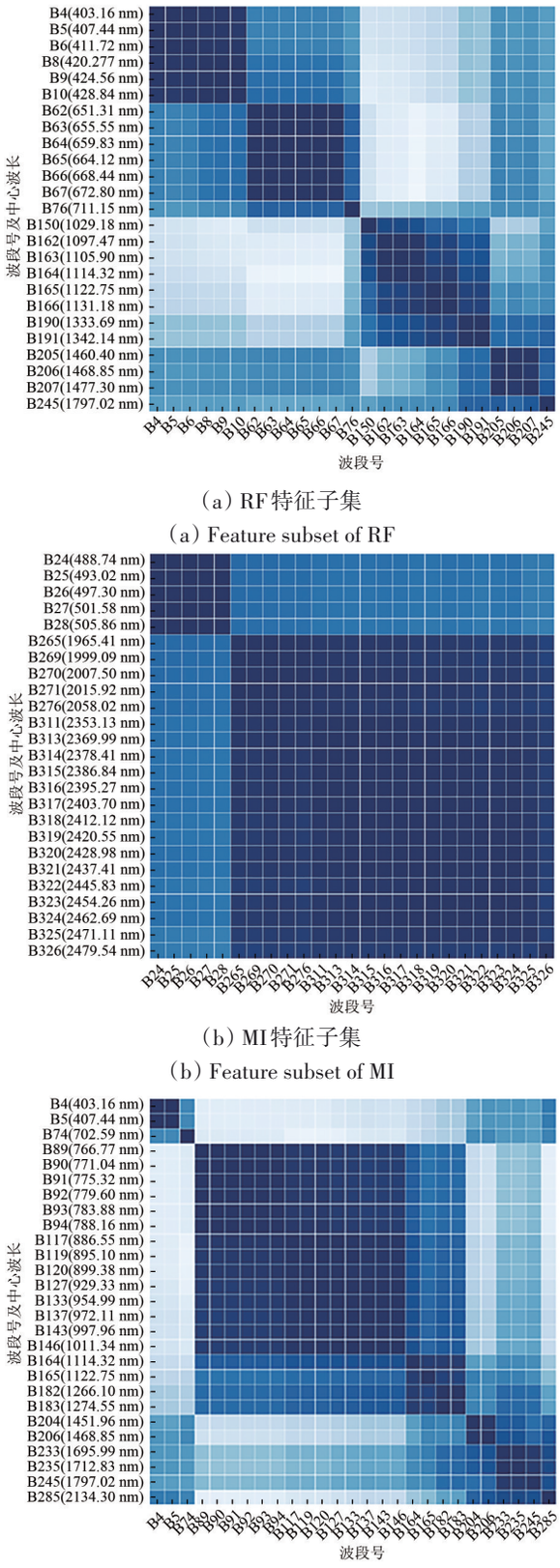
表2 降维前及第1次降维后各模型的总体分类精度比较

Table 2 Comparison of overall accuracy of each model before and after the first feature selection process

模型	降维前总体分类精度/%	第1次降维后总体分类精度/%	变化值/%
RF-RF	91.73	91.20	-0.53
RF-SVM	94.29	93.12	-1.17
RF-KNN	91.06	91.58	0.52
MI-RF	91.73	90.64	-1.09
MI-SVM	94.29	91.51	-2.78
MI-KNN	91.06	89.94	-1.12
L1-RF	91.73	92.03	0.30
L1-SVM	**94.29**	**94.54**	**0.25**
L1-KNN	91.06	92.70	1.64

注：表中的变化值为降维后的总体分类精度减去降维前的总体分类精度得到的结果。表中加粗部分为最佳模型及其精度指标值。

为了分析不同特征选择方法造成的精度变化差异，本研究对3个特征子集的自相关性进行了分析（图6）。从图6可以看出，MI特征子集中包含的大部分变量集中在蓝波段和短波红外区间的连续波段，其极高的自相关性对分类精度产生了影响。因此，过滤法并不适合高维且自相关性很强的数据集。相比之下，RF和L1特征子集中的变量不仅具有较强的代表性，而且变量波谱分布范围较为广泛。



（a）RF特征子集

（a）Feature subset of RF

（b）MI特征子集

（b）Feature subset of MI

（c）L1特征子集

（c）Feature subset of L1

图6 RF、MI和L1特征子集的自相关性

Fig. 6 Autocorrelation of RF, MI and L1 feature subsets

RF和L1两种方法获得的特征子集在分类过程中仍然具有明显的分类效果差异。使用RF特征子集时，仅RF-KNN模型的分类精度得到了提高，而L1特征子集更具代表性，3种模型均有不同程度的精度提升。这是由于L1正则化在模型中添加惩罚项的方式防止出现过拟合现象，过滤了影响模型整体精度的特征。RF通过比较每个特征变化后对结果的影响程度得到其重要性值及排序。但使用重要性值较大的变量所组成的集合进行分类，并未从整体的角度考虑到变量之间的互补或互斥特性，因此得到的分类精度低于L1正则化。

根据图6波段分布可以发现，RF和L1子集中分别有11和10个波段分布在短波红外。不同的是，RF子集中剩余的波段大部分集中在蓝波段和红波段，红边波段和近红外波段均仅有一个变量。而L1子集中，分别有7个和8个变量分布在红边和近红外波段，蓝波段中有2个变量，而红波段无分布。为讨论两种子集的分类精度差异，本研究对基于L1和RF子集模型的分类结果进行了分析（表3）。在使用RF、SVM和KNN分类算法的情况下，L1子集对农田的召回率较RF子集分别提高了1.63%，3.45%和1.73%。该结果表明，与RF子集中的蓝波段和红波段相比，L1子集中的红边和近红外波段更有利于提高模型对农田的识别能力。

表3 基于RF和L1特征子集的分类模型对不同地物类型的召回率

Table 3 Recall rate of classification model based on RF and L1 feature subset for different feature types

模型	召回率/%				
	农田	森林	水体	不透水面	裸土
RF-RF	90.75	90.48	97.88	89.17	91.22
L1-RF	92.38	89.58	95.75	90.75	93.69
RF-SVM	90.45	92.38	97.66	89.96	97.53
L1-SVM	**93.90**	**95.69**	**96.82**	**90.26**	**97.24**
RF-KNN	90.55	92.59	94.06	90.65	91.52
L1-KNN	92.28	94.19	94.06	89.67	94.08

注：表中加粗部分为最佳模型及其精度指标值。

3.3 基于后向序列选择的降维结果及分析

在后向序列选择方法搜索最佳特征子集后，各模型的总体分类精度均有不同程度的提升（表4）。其中精度提升最大的为MI-KNN-SBS模型（提高1.64%），精度提升最小的为L1-SVM-SBS模型（提高0.1%）。尽管L1-SVM-SBS模型的精度提升最小，其总体分类精度（94.64%）仍然优于其他模型。表4也展示了使用后向序列选择处理后，各模型的农田召回率。与第一次降维后得到的农田召回率相比，混合特征选择算法提高了所有模型的农田识别精度。除了L1-RF-SBS模型外，其他所有模型的农田召回率均提高了1%以上，其中RF-SVM-SBS模型的农田召回率提高最明显（增加3.86%）。这说明与单一特征选择方法相比，混合式特征选择方法不仅提升了总体分类精度，还剔除了影响农田识别度的波段，有效提高了农田识别精度。在所有模型中，L1-SVM-SBS模型的农田召回率最高（95.83%），其次是RF-SVM-SBS模型（94.31%）。本研究提出的方法在常规特征选择处理后，又通过使用后向序列选择法再次降维，得到最佳子集。该方法既减小了模型的输入维度，又一定程度上提高了分类精度。与单一特征选择方法相比，该方法从理论上用最优方式解决降维问题，以遍历的形式，简单、迅速的给出特定模型下的最优子集。

表4中3种基于L1子集的模型（L1-RF-SBS，L1-SVM-SBS，L1-KNN-SBS）也具有较高维度的输入，这说明了L1子集中有更多的变量对模型精度提升具有贡献性。此外，基于SVM分类算法的模型在使用SBS方法降维后得到的精度提升均较低，同时与其他方法相比，最优子集仍保持较高的维度。这说明SVM分类器更适用于高维空间，对具有较高自相关性的特征子集仍然保持较好的分类精度。本研究最终选择L1-SVM-SBS模型对GF-5高光谱数据进行分类，分类结果和验证数据的混淆矩阵分别如图7和表5所示。农田的召回率为95.83%，说明该模型适用于农田识别。验证数据中分别有5.02%和7.78%的裸土像元被误分为农田和不透水面，这是由于混合像元的影响，导致裸土的识别精度相对较低。

表4 第2次降维后各模型的总分类精度以及所对应的输入维度

Table 4 Overall accuracy of each model and the corresponding input variable dimension after the second feature selection process

模型	第2次降维后总体分精度/%	与第1次降维结果相比总体分类精度提升百分比	输入维度	农田召回率/%	与第1次降维结果相比农田召回率提升百分比
RF-RF-SBS	91.66	0.46	13	92.37	1.62
RF-SVM-SBS	93.32	0.20	18	94.31	3.86
RF-KNN-SBS	92.01	0.43	14	91.97	1.42
MI-RF-SBS	91.29	0.65	7	91.46	2.78
MI-SVM-SBS	91.56	0.05	14	88.72	1.14
MI-KNN-SBS	90.28	1.64	3	86.18	2.11
L1-RF-SBS	92.85	0.82	26	93.19	0.81
L1-SVM-SBS	**94.64**	**0.10**	**23**	**95.83**	**1.93**
L1-KNN-SBS	92.97	0.27	17	93.49	1.21

注：L1-SVM-SBS模型的输入波段号为B4，B5，B74，B90，B91，B92，B93，B94，B120，B127，B133，B137，B143，B146，B164，B165，B182，B183，B204，B206，B233，B235，B285。表中加粗部分为最佳模型及其精度指标值。

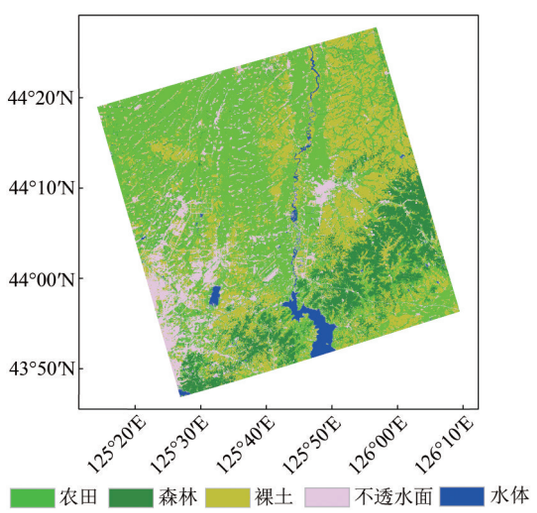


图7 L1-SVM-SBS模型分类结果图

Fig. 7 Classification map using L1-SVM-SBS model

表5 基于L1-SVM-SBS模型的分类结果混淆矩阵

Table 5 Confusion matrix of classification results based on L1-SVM-SBS model

类型	农田	森林	水体	裸土	不透水面	总计
农田	943	31	0	6	4	984
森林	12	979	0	0	7	998
水体	1	0	451	19	0	471
裸土	51	4	3	879	79	1016
不透水面	5	0	0	18	991	1014
总计	1012	1014	454	922	1081	4483

4 结 论

高光谱数据在地物识别方面具有独特的优势，但其丰富的波谱信息同时也为数据处理带来了挑战。本研究提出了一种混合式特征选择方法用于高光谱数据农田分类，主要研究结论如下：

（1）使用高光谱数据进行农田分类时，L1正则化方法能够得到最优的特征选择结果，从而得到较高的分类精度。RF和L1正则化得到的特征子集具有较强的代表性，而基于MI方法得到的特征子集具有较高的自相关性，严重影响了分类精度。使用同种分类算法情况下，L1子集比RF子集得到的精度更高，说明L1子集中的红边、近红外波段比RF子集中的蓝、红波段更适用于农田分类。

（2）3种分类算法在高分五号高光谱数据降维和农田分类中表现出明显的差异。SVM在高维空间中更容易实现较高的分类精度，对冗余数据有较强的抗噪性。而RF在低维空间中表现出更强的泛化能力，可以在有限的输入特征中学习到更强的数据规律。

（3）本研究提出的混合式特征选择方法有效的将嵌入式（或过滤式）和包装法的优点相结合，即借助了机器学习算法强大的泛化能力，又结合了包装法的搜索能力，通过使用较少但更具代表性的波段，实现更精确的分类。在后续的研究中，将会通过加入纹理、植被指数等特征进一步丰富原始数据特征空间，发挥该算法的优势，从而选出更有价值的信息。

参考文献（References）

Alvarez-Meza A M, Lee J A, Verleysen M and Castellanos-Dominguez G. 2017. Kernel-based dimensionality reduction using Renyi's α-entropy measures of similarity. Neurocomputing, 222: 36-46 [DOI: 10.1016/j.neucom.2016.10.004]

Aneece I and Thenkabail P. 2018. Accuracies achieved in classifying five leading world crop types and their growth stages using optimal earth observing-1 Hyperion hyperspectral narrowbands on Google earth engine. Remote Sensing, 10: 2027 [DOI: 10.3390/rs10122027]

Bolón-Canedo V and Alonso-Betanzos A. 2019. Ensembles for feature selection: a review and future trends. Information Fusion, 52: 1-12 [DOI: 10.1016/j.inffus.2018.11.008]

Ding X H, Zhang S Q, Li H P, Wu P, Dale P, Liu L J and Cheng S. 2020. A restrictive polymorphic ant colony algorithm for the optimal band selection of hyperspectral remote sensing images. International Journal of Remote Sensing, 41(3): 1093-1117 [DOI: 10.1080/01431161.2019.1655810]

Dong C and Zhao G X. 2020. Influence of time series data quality on land cover classification accuracy. Remote Sensing Technology and Application, 35(3): 558-566. (董超, 赵庚星. 2020. 时序数据集构建质量对土地覆盖分类精度的影响研究. 遥感技术与应用, 35(3): 558-566) [DOI: 10.11873/j.issn.1004-0323.2020.3.0558]

Dong X F, Gan F P, Li N, Yan B K, Zhang L, Zhao J Q, Yu J C, Liu R Y and Ma Y N. 2020. Fine mineral identification of GF-5 hyperspectral image. Journal of Remote Sensing, 24(4), 454-464 (董新丰, 甘甫平, 李娜, 闫柏琨, 张磊, 赵佳琪, 于峻川, 刘镕源, 马燕妮. 2020. 高分五号高光谱影像矿物精细识别. 遥感学报, 24(4): 454-464) [DOI: 10.11834/jrs.20209194]

Fan D D, Li Q Z, Wang H Y, Zhang Y, Du X and Shen Y. 2019. Improvement in recognition accuracy of minority crops by resampling of imbalanced training datasets of remote sensing. Journal of Remote Sensing, 23(4): 730-742 (樊东东, 李强子, 王红岩, 张源, 杜鑫, 沈宇. 2019. 通过训练样本采样处理改善小宗作物遥感识别精度. 遥感学报, 23(4): 730-742) [DOI: 10.11834/jrs.20197478]

Ghorbanian A, Kakooei M, Amani M, Mahdavi S, Mohammadzadeh A and Hasanlou M. 2020. Improved land cover map of Iran using sentinel imagery within Google earth engine and a novel automatic workflow for land cover classification using migrated training samples. ISPRS Journal of Photogrammetry and Remote Sensing, 167: 276-288 [DOI: 10.1016/j.isprsjprs.2020.07.013]

González J, Ortega J, Damas M, Martín-Smith P and Gan J Q. 2019. A new multi-objective wrapper method for feature selection-Accuracy and stability analysis for BCI. Neurocomputing, 333: 407-418 [DOI: 10.1016/j.neucom.2019.01.017]

Han Z and Song W. 2019. Spatiotemporal variations in cropland abandonment in the Guizhou-Guangxi karst mountain area, China. Journal of Cleaner Production, 238: 117888 [DOI: 10.1016/j.jclepro.2019.117888]

Hao P Y, Chen Z X, Tang H J, Li D D and Li H. 2019. New workflow of plastic-mulched farmland mapping using multi-temporal sentinel-2 data. Remote Sensing, 11(11): 1353 [DOI: 10.3390/rs11111353]

Htitiou A, Boudhar A, Lebrini Y, Hadria R, Lionboui H, Elmansouri L, Tychon B and Benabdelouahab T. 2019. The performance of random forest classification based on Phenological metrics derived from sentinel-2 and landsat 8 to map crop cover in an irrigated semi-arid region. Remote Sensing in Earth Systems Sciences, 2(4): 208-224 [DOI: 10.1007/s41976-019-00023-9]

Jia K and Li Q Z. 2013. Review of features selection in crop classification using remote sensing data. Resources Science, 35(12): 2507-2516 (贾坤, 李强子. 2013. 农作物遥感分类特征变量选择研究现状与展望. 资源科学, 35(12): 2507-2516)

Kussul N, Lavreniuk M, Skakun S and Shelestov A. 2017. Deep learning classification of land cover and crop types using remote sensing data. IEEE Geoscience and Remote Sensing Letters, 14(5): 778-782 [DOI: 10.1109/LGRS.2017.2681128]

Li M, Chen H, Shi X, Liu S, Zhang M and Lu S F. 2019. A multi-information fusion "triple variables with iteration" inertia weight PSO algorithm and its application. Applied Soft Computing, 84: 105677 [DOI: 10.1016/j.asoc.2019.105677]

Li Q Z. 2018. Prospect of grain production and supply service mode of China in internet plus era. China Agricultural Informatics, 30(1): 93-99 (李强子. 2018. 互联网+时代中国粮食的生产与供给服务模式思考. 中国农业信息, 30(1): 93-99) [DOI: 10.12105/j.issn.1672-0423.20180109]

Li Y, Li T and Liu H. 2017. Recent advances in feature selection and its applications. Knowledge and Information Systems, 53(3): 551-577 [DOI: 10.1007/s10115-017-1059-8]

Liu D and Sun K. 2019. Random forest solar power forecast based on classification optimization. Energy, 187: 115940 [DOI: 10.1016/j.energy.2019.115940]

Liu J, Liu J K, An J J and Zhang C. 2020. Precise crop classification based on multi-features from time-series Landsat8 OLI images and random forest algorithm. Agricultural Research in the Arid Areas, 38(3): 281-288, 298 (刘杰, 刘吉凯, 安晶晶, 章超. 2020. 基于时序Landsat8 OLI多特征与随机森林算法的作物精细分类研究. 干旱地区农业研究, 38(3): 281-288, 298) [DOI: 10.7606/j.issn.1000-7601.2020.03.37]

Liu X P, Li X, Tan Z Z and Chen Y M. 2011. Zoning farmland protection under spatial constraints by integrating remote sensing, GIS and artificial immune systems. International Journal of Geographical Information Science, 25(11): 1829-1848 [DOI: 10.1080/13658816.2011.557380]

Liu X S, Gong Z W and Wu J. 2018. Land use information extraction using multiple features derived from hyperspectral images. Journal of Nanjing Forestry University (Natural Science Edition), 42(4): 141-147 (刘晓双, 龚直文, 吴见. 2018. 基于多特征的高光谱遥感土地利用信息提取. 南京林业大学学报(自然科学版). 42(4): 141-147) [DOI: 10.3969/j.issn.1000-2006.201705029]

Liu Y N, Xun X D, Hu X N, Liu S F, Cao K Q, Chai M Y, Liao Q J, Zuo Z Q, Hao Z Y, Duan W B, Zhou W Y N, Zhang J and Zhang Y. 2020. Development of visible and short-wave infrared hyperspectral imager onboard GF-5 satellite. Journal of Remote Sensing, 24(4): 333-344 (刘银年, 孙德新, 胡晓宁, 刘书锋, 曹开钦, 柴孟阳, 廖清君, 左志强, 郝振贻, 段微波, 周魏乙诺, 张静, 张营. 2020. 高分五号可见短波红外高光谱相机设计与研制. 遥感学报, 24(4): 333-344) [DOI: 10.11834/jrs.20209196]

Maldonado S and López J. 2018. Dealing with high-dimensional class-

imbalanced datasets: embedded feature selection for SVM classification. Applied Soft Computing, 67: 94-105 [DOI: 10.1016/j.asoc.2018.02.051]

Park M Y and Hastie T. 2007. L_1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(4): 659-677 [DOI: 10.1111/j.1467-9868.2007.00607.x]

Radley S, Sybi C J and Premkumar K. 2020. Multi information amount movement aware- routing in FANET: flying ad-hoc networks. Mobile Networks and Applications, 25(2): 596-608 [DOI: 10.1007/s11036-019-01395-4]

Sánchez-Maroño N, Alonso-Betanzos A and Tombilla-Sanromán M. 2007. Filter methods for feature selection – a comparative study// Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning. Birmingham, UK, December: Springer: 178-187 [DOI: 10.1007/978-3-540-77226-2_19]

Sukawattanavijit C, Chen J and Zhang H S. 2017. GA-SVM algorithm for improving land-cover classification using SAR and optical remote sensing data. IEEE Geoscience and Remote Sensing Letters, 14(3): 284-288 [DOI: 10.1109/LGRS.2016.2628406]

Sylvester E V A, Bentzen P, Bradbury I R, Clément M, Pearce J, Horne J and Beiko R G. 2018. Applications of random forest feature selection for fine-scale genetic population assignment. Evolutionary Applications, 11(2): 153-165 [DOI: 10.1111/eva.12524]

Wang X X, Gao X W, Zhang Y Z, Fei X Y, Chen Z, Wang J, Zhang Y Y and Zhao H M. 2019. Land-cover classification of coastal wetlands using the RF algorithm for Worldview-2 and Landsat 8 images. Remote Sensing, 11(16): 1927 [DOI: 10.3390/rs11161927]

Wei Y H, Wang Y, He X M, Guo K and Chang R C. 2020. Method of terrain classification based on GF-5 satellite remote sensing images. Modern Electronics Technique, 43(18): 85-88 (魏友华, 王瑶, 何雪梅, 郭科, 常睿春. 2020. 基于“高分五号”遥感图像的地物分类方法. 现代电子技术, 43(18): 85-88) [DOI: 10.16652/j.issn.1004-373x.2020.18.022]

Xu L, Ming D P, Zhou W, Bao H Q, Chen Y Y and Ling X. 2019. Farmland extraction from high spatial resolution remote sensing images based on stratified scale pre-estimation. Remote Sensing, 11(2): 108 [DOI: 10.3390/rs11020108]

Yan X A and Jia M P. 2018. A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing. Neurocomputing, 313: 47-64 [DOI: doi.org/10.1016/j.neucom.2018.05.002]

Yin H, Prishchepov A V, Kuemmerle T, Bleyhl B, Buchner J and Radeloff V C. 2018. Mapping agricultural land abandonment from spatial and temporal segmentation of Landsat time series. Remote Sensing of Environment, 210: 12-24 [DOI: 10.1016/j. rse. 2018. 02.050]

Yoo C, Han D, Im J and Bechtel B. 2019. Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images. ISPRS Journal of Photogrammetry and Remote Sensing, 157: 155-170 [DOI: 10.1016/j.isprsjprs.2019.09.009]

Yu Q Y, Xiang M T, Wu W B and Tang H J. 2019. Changes in global cropland area and cereal production: an inter-country comparison. Agriculture, Ecosystems and Environment, 269: 140-147 [DOI: 10.1016/j.agee.2018.09.031]

Yuan J W, Wu C, Du B, Zhang L P and Wang S G. 2020. Analysis of landscape pattern on urban land use based on GF-5 hyperspectral data. Journal of Remote Sensing, 24(4): 465-478 (袁静文, 武辰, 杜博, 张良培, 王树根. 2020. 高分五号高光谱遥感影像的城市土地利用景观格局分析. 遥感学报, 24(4): 465-478) [DOI: 10.11834/jrs.20209252]

Zhang S C. 2020. Cost-sensitive KNN classification. Neurocomputing, 391: 234-242 [DOI: 10.1016/j.neucom.2018.11.101]

Zhou X L, Su G Q, Wang L J, Nie S D and Ge X M. 2017. The inversion of 2D NMR relaxometry data using L1 regularization. Journal of Magnetic Resonance, 275: 46-54 [DOI: 10.1016/j.jmr.2016.12.003]

Hybrid feature selection for cropland identification using GF−5 satellite image

CHEN Zhulin[1,2], JIA Kun[1,2], LI Qiangzi[3], XIAO Chenchao[4], WEI Dandan[4], ZHAO Xiang[1,2], WEI Xiangqin[3], YAO Yunjun[1,2], LI Juan[3]

1.State Key Laboratory of Remote Sensing, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China;
2.Beijing Engineering Research Center for Global Land Remote Sensing Products, Beijing Normal University, Beijing 100875, China;
3.Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China;
4.Land Satellite Remote Sensing Application Center, Ministry of Natural Resource of the People's Republic of China, Beijing 100048, China

Abstract: Accurate farmland area identification is the basis of crop yield estimation and an important indicator in food security assessment. As an important data source for farmland identification, remote sensing data can provide dynamic and fast observation results

for classification. GF-5, which is the only hyperspectral satellite in the China High-resolution Earth Observation System, has great research and application potential in farmland identification. However, the dimensionality curse caused by the redundant bands in hyperspectral data seriously affects the calculation speed and classification accuracy of models. To solve this problem, this research proposes a hybrid feature selection algorithm for farmland identification. First, on the basis of the feature importance provided by the feature selection algorithm, the feature dimension is gradually reduced from 295 to 5 with a step length of 10. The overall accuracy of the classification results corresponding to each feature dimension is recorded. Second, the turning point (a dimension number whose corresponding overall accuracy hardly decreases when the input variable number is smaller than it) is determined based on the overall accuracy, and the corresponding variables are adopted as the feature subset. Lastly, the Sequential Backward Selection (SBS) method is used to search for the best subset. Three feature selection algorithms (i. e., Random Forest (RF), Multi-Information (MI), and L1 regularization (L1)) and three classification algorithms (RF, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN)) are examined. Results indicate that the autocorrelations of the three subsets differ significantly. Most of the bands selected by the MI method are continuous and concentrated in the blue and shortwave infrared range. Therefore, the extremely high autocorrelation that exists in this subset has a negative effect on classification accuracy. By contrast, the correlation between bands in the RF and L1 feature subsets is relatively weak. However, the two feature sets still result in different classification accuracy. According to the variable distribution, many red-edge and near-infrared bands are contained in the L1 feature subset. These bands demonstrate better ability to distinguish farmland, forest, and soil than the blue and red bands selected by the RF algorithm. The classification algorithms also have different capacities. In the high-dimensional space, the SVM algorithm exhibits high robustness to noise, resulting in high accuracy. However, when the dimension decreases to a critical value, the accuracy of SVM decreases sharply. By contrast, although RF is not as robust as SVM in the high-dimensional space, it has excellent generalization ability in the low-dimensional space. Compared with the subsets obtained after the first dimensionality reduction process, the optimal feature subsets obtained by SBS searching improve the classification accuracy of each model. The L1-SVM-SBS model with a 23-dimensional input achieves the highest overall classification accuracy (94.64%) and cropland recall rate (95.83%). This study provides a new method of farmland identification using hyperspectral data. By selecting numerous representative and informative bands, this method not only improves farmland classification accuracy, but can also be used as a reference for other classification problems involving hyperspectral remote sensing.

Key words: cropland identification, GF-5, feature selection, hyperspectral remote sensing, L1 regularization, sequential backward selection

Supported by National Key Research and Development Program of China (No. 2019YFE0127300, 2016YFB0501404); National Natural Science Foundation of China (No. 42171318)