第42卷，第10期
2022年10月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42，No. 10，pp3283-3290
October，2022

高光谱图像分类的 ReliefF-RFE 特征选择算法构建与应用

项颂阳[1,3]，许章华[1,2,4,5,6*]，张艺伟[1,2]，张 琦[1,3]，
周 鑫[1,2]，俞 辉[1,3]，李 彬[1,2]，李一帆[1,2]

1. 福州大学地理与生态环境研究中心，福建 福州 350108
2. 福州大学环境与安全工程学院，福建 福州 350108
3. 福州大学数字中国研究院(福建)，福建 福州 350108
4. 福建省资源环境监测与可持续经营利用重点实验室，福建 三明 365004
5. 空间数据挖掘与信息共享教育部重点实验室，福建 福州 350108
6. 福州大学信息与通信工程博士后科研流动站，福建 福州 350108

摘 要 高光谱图像具有波段连续、维数高、数据量大、相邻波段相关性强的特点，可为地物分类提供更为丰富的细节信息。但是，数据中存在大量冗余信息与噪声，在图像分类中如直接利用其所有波段特征而不进行有效分析与选择，将会导致较低的计算效率和较高的计算复杂度，分类精度亦可能随着波段维数增加而出现先增后减的“休斯(Hughes)现象”。为快速地从高达数十个甚至数百个波段的高光谱图像中提取出具有较好识别能力的特征子集，从而避免“维度灾难”，将过滤式 ReliefF 算法和封装式特征递归消除算法(RFE)相结合，构建了 ReliefF-RFE 特征选择算法，可用于高光谱图像分类的特征选择。该算法根据权重阈值，利用 ReliefF 算法快速剔除大量无关特征，缩小并优化特征子集的范围；利用 RFE 算法进一步搜索最优特征子集，将缩小范围后的特征子集中与分类器关联性小、冗余的特征进行递归筛选，进而得到分类性能最佳的特征子集。采用 Indian pines 数据集、Salinas-A 数据集与 KSC 数据集等 3 个标准数据集作为实验数据，将 ReliefF-RFE 算法的应用效果与 ReliefF 和 RFE 算法进行对比。结果显示，在 3 个数据集中，应用 ReliefF-RFE算法的高光谱图像分类平均总体精度(OA)为 92.94%、F-measure 为 92.81%，Kappa 系数为 91.94%；ReliefF-RFE 算法的平均特征维数是 ReliefF 算法的 37%，而平均运算时间则是 RFE 算法的 75%。由此表明，ReliefF-RFE 算法能够在保证分类精度的同时，克服过滤式 ReliefF 算法无法有效减小特征之间冗余以及封装式 RFE 算法时间复杂度较高的缺陷，具有更为均衡的综合性能，适用于高光谱图像分类的特征选择。

关键词 高光谱图像；特征选择；ReliefF 算法；RFE 算法；ReliefF-RFE 算法

中图分类号：TP79 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)10-3283-08

引 言

随着计算机科学和光谱成像学的发展，高光谱图像被广泛应用于精准农业、环境监测和公共安全等诸多领域。高光谱图像波段连续且数据量大，光谱和空间特征丰富，为地物解译与图像分类提供了更多可用的信息。但是，直接使用高维度特征进行地物分类并不合适。高维数据中含有大量的冗余信息和噪声，当参与分类的特征波段数不断增加，分类精度和特征数并不成正比，而是出现先上升再下降的“休斯(Hughes)现象”[1]。在保证地物分类精度的情况下，如何从原始波段中析取相关性小、维数低、信息量大且冗余度小的高光谱数据，解决维度灾难问题是十分必要的[2]。特征提取和特征选择是高光谱图像的两种主要降维方法。较之于特征

收稿日期：2021-10-10，**修订日期**：2022-01-16

基金项目：国家自然科学基金面上项目(42071300)，福建省自然科学基金面上项目(2020J01504)，福建省资源环境监测与可持续经营利用重点实验室开放基金项目(ZD202102)，3S 技术与资源优化利用福建省高校重点实验室开放课题(fafugeo201901)，晋江市福大科教园区发展中心科研项目(2019-JJFDKY-17)，中国博士后面上基金项目(2018M630728)资助

作者简介：项颂阳，1997 年生，福州大学地理与生态环境研究中心硕士研究生 e-mail：xsydedanjuan@163.com

*通讯作者 e-mail：fafuxzh@163.com

提取，基于特征选择的降维方法既能不改变既有的特征空间和特征值，保留原始地物信息，又能够剔除大部分与实际地物不相关或冗余的特征，从而达到大量减少特征个数，提高模型精度，减少运算时间的目的[3]。

为解决高光谱图像分类中存在的多维度、强相关和多冗余等问题，学界围绕特征选择开展了相关研究。特征选择按照和学习器的不同结合方式，主要囊括了过滤式(Filter)、封装式(Wrapper)、嵌入式(Embedded)和集成式(Ensemble)四种方法[4]，其中，过滤式和封装式已在高光谱图像中得到较为有效的应用。例如，Ren 等[5]利用改进的分区过滤式ReliefF算法，实现高光谱图像的显著降维；Ye 等[6]采用跨域ReliefF(corss-domain ReliefF，CDRF)算法以特征权重的跨域更新规则来考虑源域和目标域的跨场景特征选择问题，有效提高了跨场景高光谱数据集的分类精度；张东彦等[7]利用ReliefF 算法筛选对大豆识别最有效的特征，发现基于优选特征提取精度明显高于原始波段反射率，虽然略低于全部 19 个特征的分类效果，但是数据量降低了 63.16%；刘代超等[8]研究了封装式特征递归消除算法(recursive feature elimination，RFE)与随机森林分类模型相结合方法对林地与非林地的识别潜力，发现特征选择可以有效减少数据输入维数，并取得最高分类精度。然而，单一的过滤式或封装式算法并不能兼顾分类精度和分类效率。过滤式方法在进行特征选择时，具有结构简单、训练速度快、独立于具体训练模型、易于设计和理解等优点，但因其并不考虑特征之间的相关性，故不能有效减小特征之间的冗余，其得到的特征子集也不是最优的；而基于特征子集搜索的封装式方法考虑了特征维之间的相互作用，有效去除特征之间的冗余，能选出最优特征子集，但因搜索过程与分类器紧密相关，所以特征子集搜索时间复杂度较高[9]。

将过滤式算法和封装式算法结合能够筛选出与地物类别密切相关且相互之间冗余度最小的特征子集。Marwa 等[10]提出了 Filter 与 Wrapper 的高维数据特征构造的多目标混合滤波器包装进化算法，结合两者算法能够优势互补，在消除冗余特征方面效果显著；Yuan 等[11]将最大化特征空间与类编码空间之间的相关信息法与递归特征消除相结合，能够有效地集成高维蛋白质数据；Lin 等[12]提出了一种交互增益递归特征消除(IG-RFE)方法，该方法基于对称不确定性和归一化交互增益相结合，迭代去除不重要的特征，能够在生物数据分析中综合特征个体的识别能力和特征之间的相互作用，可以更好地评估特征的重要性。

针对高光谱图像分类中过滤式算法无法有效去除冗余特征和封装式算法时间复杂度较高的问题，提出一种结合过滤式和封装式的 ReliefF-RFE 特征选择算法，该算法可综合过滤式算法训练速度快、结构简单以及封装式算法能有效去除特征之间冗余的优点。尝试运用该算法对 Indian pines，Salinas-A 与 KSC 等 3 个标准数据集进行特征选择与分类，分析其分类精度、特征维数及运算时间，并将其与单一的 ReliefF 算法和 RFE 算法进行比较，验证 ReliefF-RFE 算法在高光谱图像分类特征选择中的综合性能。

1 实验部分

1.1 ReliefF 算法原理

Relief 算法由 Kira 和 Rendell 于 1992 年提出，主要用于解决二值分类的特征降维问题。针对 Relief 算法无法处理多分类的问题，Kononenko 于 1994 年对 Relief 算法进行改进，提出了 ReliefF 算法。ReliefF 算法是一种经典的多变量过滤式特征选择方法，其核心思想是根据特征与类别标签之间的相关性计算特征的权值[13-14]。该算法中特征和类别标签的相关性是基于特征对近距离样本的区分能力度量的，具体计算过程如下：首先通过随机抽样选择一个样本 x_i，然后分别计算和 R_i 相同类别中 k 个最近邻样本的距离 $\sum_{j=1}^{k}\mathrm{diff}(f_l, R_i, H_j)$，和 R_i 不同类别中 k 个最近邻样本的距离 $\sum_{j=1}^{k}\mathrm{diff}(f_l, R_i, M_j(C))$，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，通过计算类间距与类内距不断迭代 m 次更新其权值，并根据最终权值进行特征选择，其权值计算公式为

$$W^{i+1}(f_l)=W^{i}(f_l)-\frac{\sum_{j=1}^{k}\mathrm{diff}(f_l, R_i, H_j)}{m\times k}+\sum_{C\neq \mathrm{label}(R_i)}\frac{P(C)}{1-P(\mathrm{label}(R_i))}\times\frac{\sum_{j=1}^{k}\mathrm{diff}(f_l, R_i, M_j(C))}{m\times k} \tag{1}$$

式(1)中：$W^i(f_l)$为第 i 个样本中第 l 个特征 f 的权值；$H_j(j=1, 2, \cdots, k)$为与 R_i 同类的 k 个最近邻样本中的第 j 个样本；$P(C)$为在训练样本中属于类别 C 的样本所占比值；$P(\mathrm{label}(R_i))$为与同类的样本占总样本的比值，其中 $\mathrm{label}(R_i)$为 R_i 的标签；$M_j(C)(j=1, 2, \cdots, k)$为与 R_i 不同类的 k 个最近邻样本中的第 j 个样本。距离计算公式为

$$D_f(i, j)=\sqrt{\sum_f (i_f-j_f)^2/\sigma_f} \tag{2}$$

式(2)中：$\mathrm{diff}(f, R_1, R_2)$为样本 R_1 与 R_2 在第 f 个特征上的归一化距离，R_{1f}和 R_{2f}分别为样本 R_1 和 R_2 的第 f 个特征，$\max(f)$和 $\min(f)$分别为所有样本中对应特征 f 的最大值和最小值。

1.2 递归特征消除算法原理

递归特征消除法是一种贪婪的优化算法，是封装式算法的代表。该算法的主要思想是反复创建模型，逐步剔除最不重要特征。该方法是一个循环过程，每个过程都包含以下 3 个步骤：(1)用当前数据集训练分类器，获得与分类器特征相关的信息即每个特征的权重；(2)根据事先制定的规则，计算所有特征的排序准则分数；(3)在当前数据集中移除对应于最小排序准则分数的特征。该循环过程一直执行到特征集合中剩余最后一个变量时结束，执行的结果为获得一列按照特征重要性排序的特征序号列表，这个迭代循环过程实际上是一个序列后向选择的过程，它在整个循环过程中先是去

除了与判别不相关的特征，保留了对判别相对重要的优化特征子集，因而可以达到优化特征子集选择，提高判别精度的目的。然而，封装式由于每次与特征子集的评价都要进行分类器的训练和测试，计算复杂度很高。因此，封装式特征选择的计算开销通常比过滤式特征选择要大得多。

1.3　ReliefF-RFE 算法构建与原理

过滤式算法运算高效，但是其并不考虑特征之间的相关性。然而，与类相关性强的单个特征组合在一起并不一定有利于分类[15]，是因为相关特征中包含一些与其他相关特征具有相同分类信息的特征，即“冗余特征”。这类特征不仅会增加分类器的计算复杂度，而且可能导致分类器性能急剧下降。因此，需要寻找一种比较高效的特征搜索算法，从而获得最佳的分类特征子集。封装式算法的特征选择与具体分类器紧密相关，因直接使用分类器的识别率来评价特征性能，并将选择所得特征直接用于构造最终的分类模型，所以封装模型相对于过滤模型具有更好的分类识别性能。但是，由于每一次选择都涉及到分类器的建模计算，所以封装式算法的运算时间要比过滤式算法慢很多。综合考虑过滤式算法和封装式算法的优缺点，在过滤式 ReliefF 算法基础上，选取考虑到特征之间相关性的封装式 RFE 算法，构建 ReliefF-RFE 算法。首先利用 ReliefF 算法对不同的特征赋予不同的权重，通过权重阈值有效剔除无关特征并保留有助于分类的特征，减少后续特征子集搜索范围，其次，将 ReliefF 算法筛选后留下的特征，利用 RFE 算法进一步筛选出与特征间相关性最大且相互之间冗余性最小的最佳分类特征子集。

设输入训练集 x，最近邻样本个数 k，步长 step，输出 x_k 为训练集 x 与目标类别有最大相关性且相互之间具有最小冗余性的特征子集，那么，ReliefF-RFE 算法具体可分为以下 8 个主要步骤：

(1)从训练集 x 中进行随机抽样选择一个样本 R_i；

(2)计算与样本 R_i 相同类别的 k 个猜中近邻的距离 $\sum_{j=1}^{k}\mathrm{diff}(f_l, R_i, H_j)$ 和其他不同类的 k 个猜错近邻的距离 $\sum_{j=1}^{k}\mathrm{diff}(f_l, R_i, M_j(C))$；

(3)根据与 R_i 不同类样本所占比值，最终计算出特征的权重 $W^i(f_l)$；

(4)筛选出权重系数总和达 95% 的特征子集，通过权重阈值剔除无关特征，得到 ReliefF 算法筛选后的特征子集 x_j；

(5)将特征子集 x_j 作为 RFE 算法的输入，同时设置最优特征子集 x_k 为空，max=0；

(6)在特征子集 x_j 构建随机森林分类器，并通过随机森

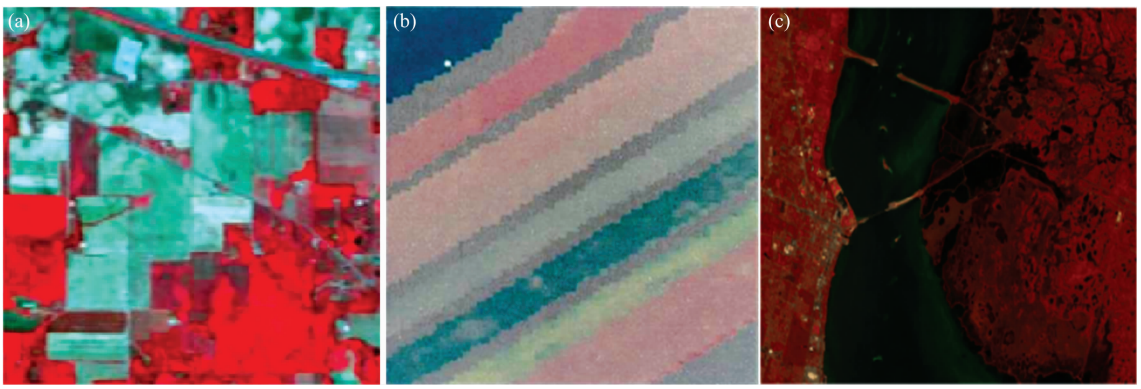

图 1　各标准数据集的高光谱图像

(a)：Indian pines；(b)：Salinas-A；(c)：KSC

Fig. 1　Hyperspectral images of each standard dataset

(a)：Indian pines；(b)：Salinas-A；(c)：KSC

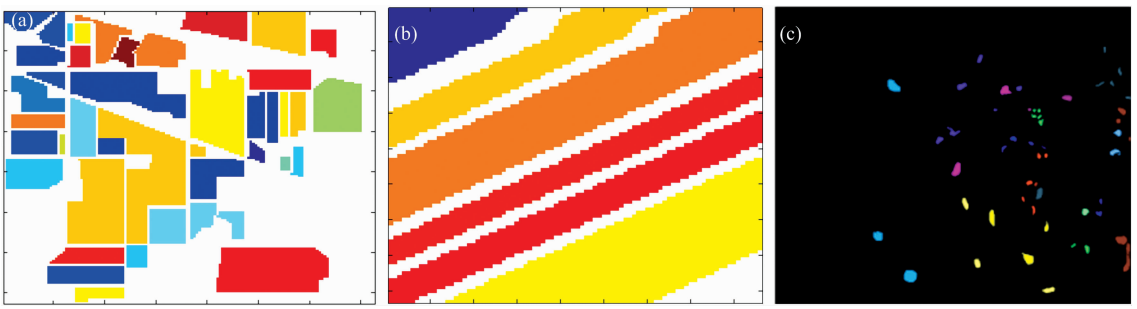

图 2　各标准数据集的真实地物参考图

(a)：Indian pines；(b)：Salinas-A；(c)：KSC

Fig. 2　References of real features for each standard dataset

(a)：Indian pines；(b)：Salinas-A；(c)：KSC

林的 feature_importances_属性获得每个特征的重要性；

(7)将特征子集 x_j 中每个特征的重要性进行排序，根据步长 step 从当前的这组特征集合中修剪最不重要的特征，计算修剪后集合的 Accuracy 值。若 Accuracy≥max，则 max=Accuracy，且取当前集合为最优特征子集，否则 max 和最优特征子集不变；

(8)在修剪的集合上递归地重复该过程，逐次进行迭代，直到找到 Accuracy 值最高的一组特征子集 x_k。

1.4 数据与方法

为验证算法的适用性，采用高光谱图像分类领域所公认的 Indian pines，Salinas-A 及 KSC 等 3 个标准数据集，数据集中包含了许多地物类别(图 1 和图 2)。Indian pines 数据集与 Salinas-A 数据集的地物类别分别为 16 类和 6 类，地物的分布比较规则，每类地物的分布整体性好；而 KSC 数据集的地物类别共 13 类，分布较为分散。各数据集描述如表 1 所示。

表 1 高光谱数据集描述

Table 1 Hyperspectral dataset description

数据集	Indian pines	Salinas-A	KSC
特征维数	200	204	176
样本数	10 249	5 348	5 211
影像尺寸	145×145	86×83	512×614
分辨率/m	20	3.7	18
地物类别	16	6	13

(1)Indian pines 数据集是使用 AVIRIS 成像光谱仪获取的美国印第安纳州西北部地区的试验区域图像，地表覆盖类型混合了林地、农田、道路、房屋建筑等。标记样本分布不均衡，部分类别样本较少。各种农作物基本都处于生长初期，对地表的林冠覆盖程度只有 5%，裸地和作物残渣对植被像元分类影响明显。以上原因导致数据集类间相似度非常高，分类难度大大增加。Indian pines 图像大小 145×145 像素，波长范围 0.4～2.5 μm，220 个波段，空间分辨率 20 m，去除坏波段和水体吸收后剩余 200 个波段。

(2)Salinas-A 数据集也是用 AVIRIS 成像光谱仪拍摄的，同 Indian pines 图像相似，是对美国 Salinas 山谷进行成像。它区别于 Indian pines 的是，Salinas-A 数据集能够达到 3.7 m 的空间分辨率。图像原本包含的波段有 224 个，对没有水反射的 108～112，154～167 以及第 224 波段都需要剔除，剩余波段图像有 204 个。

(3)KSC 数据集由佛罗里达肯尼迪航天中心的 AVIRIS 传感器收集，包含 224 个波段。该数据集由 512×614 个像素组成，其中每个像素具有 8 m 的空间分辨率。同时，光谱覆盖范围为 0.4～2.5 μm，去除 48 个噪声比较高的波段后，共使用了 176 个波段。

为验证本算法的可靠性，将 ReliefF-RFE 算法及单一的 ReliefF、RFE 算法分别应用于前述 3 个高光谱数据集，分析并比较 3 种算法的分类效果，分类器统一采用随机森林分类器。在进行实验比较之前，先设置不同算法的参数。ReliefF 算法通过随机抽样选择样本计算某个特征的权值，为了有效处理权重偏差，避免最终运算结果的权重不同，对某一特征的不同权值，通过 10 次运算求取平均值作为该特征的权值，设定最邻近样本数 k 为 10，对大多数分类任务最为可靠有效；随机森林的最佳参数通过网格搜索交叉验证工具(GridSearchCV)进行参数寻优。随机选取 70%的样本作为训练样本，30%样本作为验证样本，统计十折交叉验证下最佳特征子集的总体精度(overall accuracy，OA)、F-measure、Kappa 系数和筛选的最小特征维数以及特征选择算法的运算时间，各指标的值越大，特征维数和运算时间越少，说明对应特征选择算法的性能越好，具体公式如式(3)—式(7)

$$\mathrm{OA}=\frac{\mathrm{TN}+\mathrm{TP}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}\times 100\% \tag{3}$$

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\times 100\% \tag{4}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times 100\% \tag{5}$$

$$\text{F-measure}=\frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}\times 100\% \tag{6}$$

$$\mathrm{Kappa}=\frac{p_0-p_c}{1-p_c}=\frac{N\sum_{i=1}^{R}x_{ii}-\sum_{i=1}^{R}(x_{i+}x_{+i})}{N^2-\sum_{i=1}^{R}(x_{i+}x_{+i})} \tag{7}$$

式中：Precision 表示精确度；Recall 表示召回率；TP(true positive)表示被模型预测为正的正样本；TN(true negative)表示被模型预测为负的负样本；FP(false positive)表示被模型预测为正的负样本；FN(false negative)表示被模型预测为负的正样本；R 为类别数；x_{ii} 为第 i 行第 i 列上的数目；x_{i+} 为第 i 行的总数目；x_{+i} 为第 i 列的总数目；N 为数据集总数目。

2 结果与讨论

2.1 Indian pines 数据集算法应用分析

分别将 3 种特征选择算法与随机森林分类器相结合，对 Indian pines 数据集进行分类，由此获得该数据集的最佳特征子集。利用验证样本计算混淆矩阵评价分类精度，统计 3 种特征选择算法的 OA，F-measure 和 Kappa 系数以及特征维数和运算时间(表 2)。结果表明，ReliefF 算法的分类精度最低且筛选出的特征维数最高，其原因可能与 ReliefF 算法筛选出来的特征仍存在冗余，增加了模型复杂度和过拟合的风险，不利于提高模型可解释性，使用的特征维数越多，并不

表 2 Indian pines 数据集最佳特征子集下的各评价指标

Table 2 Classification results based on the best feature subset of the Indian pines dataset

算法	OA /%	F-measure /%	Kappa 系数	特征维数	t /h
ReliefF	85.53	85.15	0.834 0	157	58.56
RFE	86.41	86.10	0.844 2	36	107.33
ReliefF-RFE	86.28	86.00	0.842 2	45	91.58

意味分类预测效果越好；RFE 算法取得了较高的分类精度，同时大幅降低了建模所需的特征数量，增强了模型的泛化能力，但是单独采用 RFE 算法进行特征选择所需要的运算时间较长，当数据集规模足够大时，搜索最佳特征子集的运算时间将不能够被接受；较之于 ReliefF 算法，采用过滤式和封装式相结合的 ReliefF-RFE 算法，其 OA 提高了 0.75%，F-measure 提高了 0.85%，Kappa 系数提高了 0.008 2；在特征维数方面，ReliefF-RFE 算法能够去除 ReliefF 算法的冗余特征，其特征维数是 ReliefF 算法的 29%；在运算时间方面，ReliefF-RFE 算法是 RFE 算法的 85%，说明 ReliefF-RFE 算法降低了封装式算法的运算时间，提升了模型的整体运算效率。

从 3 种特征选择算法的分类结果图可知(图 3)，部分地物存在一定的错分误分现象，但总体上，各算法的分类效果与真实地物参考图基本接近。

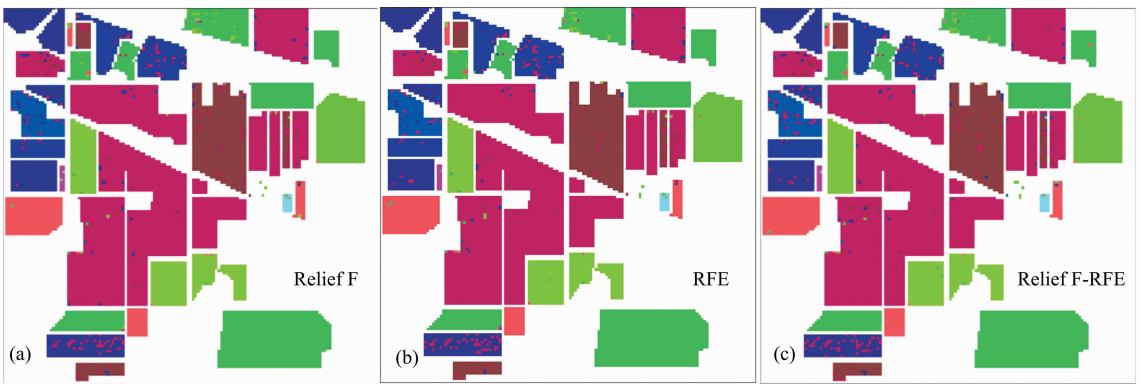

图 3　Indian pines 数据集高光谱图像 3 种特征选择算法的分类结果

Fig. 3　Classification results of three feature selection algorithms for hyperspectral images of Indian pines dataset

2.2　Salinas-A 数据集算法应用分析

利用随机森林分类器结合 3 种特征选择算法得到 Salinas-A 数据集的最佳特征子集，可以发现，3 种算法都取得了较好的预测效果(表 3)。ReliefF 算法的特征维数最高，RFE 算法以较长的运算时间为代价，用最少的特征数实现了和 ReliefF 算法相同的预测效果，这也说明 RFE 算法能够进一步剔除冗余特征，减少数据集处理的复杂度。在分类精度方面，ReliefF-RFE 算法首先通过 ReliefF 算法初步筛选掉一批无相关特征，再利用 RFE 算法进一步去除冗余特征，用较少规模的特征子集取得比单一 ReliefF 算法和 RFE 算法更好的分类精度，其 OA 提高了 0.07%，F-measure 提高了 0.06%，Kappa 系数提高了 0.000 8；在特征维数方面，ReliefF-RFE 算法是 ReliefF 算法的 39%，有效降低了特征维数；在运算时间方面，ReliefF-RFE 算法的运算时间是 RFE 算法的 76%。

表 3　Salinas-A 数据集最佳特征子集下的各评价指标

Table 3　Classification results based on the best feature subset of the Salinas-A dataset

算法	OA /%	F-measure /%	Kappa 系数	特征维数	t /h
ReliefF	99.31	99.32	0.991 4	180	6.34
RFE	99.31	99.32	0.991 4	35	11.35
ReliefF-RFE	99.38	99.38	0.992 2	71	8.63

从 3 种特征选择算法的分类细节上来看(图 4)，各地物的边缘区域会存在错分情况，ReliefF-RFE 算法在总体上的错分情况较少。

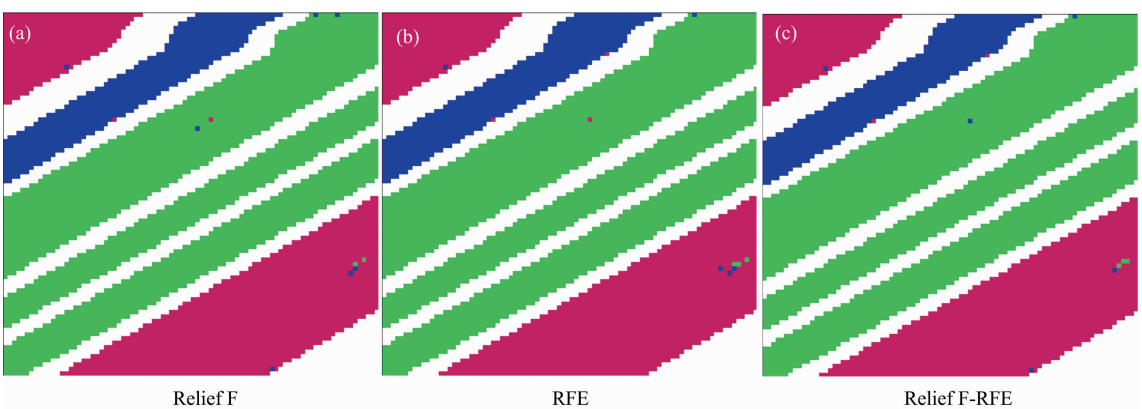

图 4　Salinas-A 数据集高光谱图像 3 种特征选择算法的分类结果

Fig. 4　Classification results of three feature selection algorithms for hyperspectral images of Salinas-A dataset

2.3 KSC 数据集算法应用分析

参照 Indian pines 与 Salinas-A 数据集的步骤，利用 KSC 数据集，对 3 种算法进行分析。结果发现，ReliefF 算法筛选的特征维数依然较大，冗余特征的存在在一定程度上影响了模型性能；RFE 算法以特征递归消除的方式搜索最优特征子集，具有更好的预测效果；ReliefF-RFE 算法的降维效果最好，分别是 ReliefF 算法和 RFE 算法特征维数的 44% 和 89%，并且在运算时间上具有显著的优势，是 RFE 算法运算时间的 63%，在损失小部分分类精度情况下，大幅降低了特征选择算法的时间复杂度。

在 KSC 数据集中，相较于前两者算法，ReliefF-RFE 算法的分类精度最低(图 5)，其原因可能在于：KSC 数据集地物类别精细、数量高达 13 类，如湿地地物类别多达 6 种，然而其光谱曲线的可分性有限，同种地物类别会有不同的群体状况表现，增加了分类难度。

表 4 KSC 数据集最佳特征子集下的各评价指标

Table 4 Classification results with the best feature subset of the KSC dataset

算法	OA /%	F-measure /%	Kappa 系数	特征维数	t /h
ReliefF	93.22	93.14	0.924 0	145	8.22
RFE	93.41	93.31	0.936 2	72	14.98
ReliefF-RFE	93.16	93.06	0.923 8	64	9.58

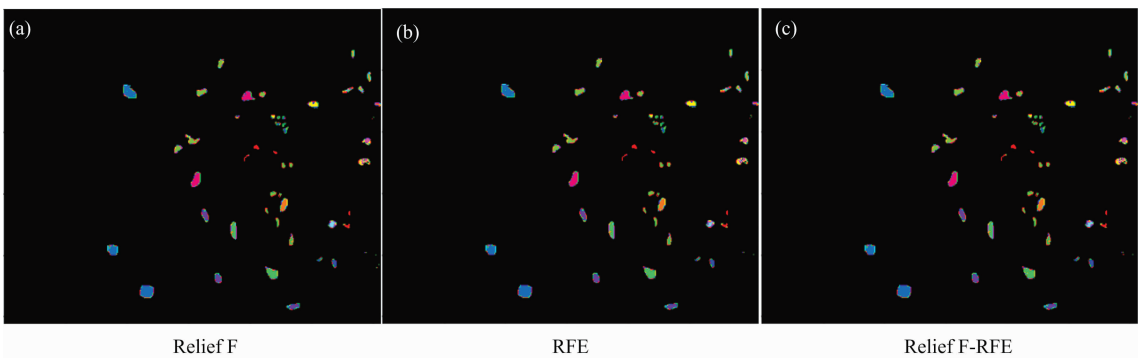

图 5 KSC 数据集高光谱图像 3 种特征选择算法的分类结果

Fig. 5 Classification results of three feature selection algorithms for hyperspectral images of KSC dataset

2.4 3 种算法特征选择的综合比较分析

对比 ReliefF-RFE 算法与 ReliefF，RFE 算法对 3 个高光谱数据集分类应用结果可以发现：其一，在分类精度方面[图 6(a)]，从总体分类精度可以发现，RFE 算法最佳，其平均 OA 为 93.04%、F-measure 为 92.91%，Kappa 系数为 92.06%；其次为 ReliefF-RFE 算法，其平均 OA 为 92.94%、F-measure 为 92.81%，Kappa 系数为 91.94%；ReliefF 算法最低，其平均 OA 为 92.69%、F-measure 为 92.54%，Kappa 系数为 91.65%；从各数据集中可以发现，ReliefF-RFE 算法在 Indian pines 数据集上比 ReliefF 算法拥有更好的预测性能，其 OA 提高了 0.75%，F-measure 提高了 0.85%，Kappa 系数提高了 0.008 2，在 Salinas-A 数据集上，其 OA 分别较其他两种算法提高 0.07%，F-measure 提高 0.06%，Kappa 系数提高 0.000 8。其二，在特征维数和运算时间角度方面

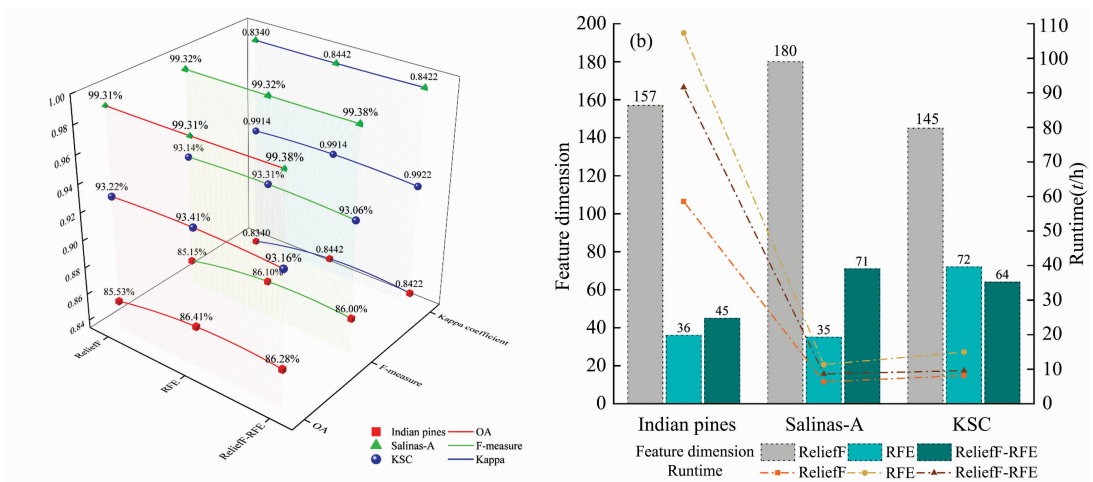

图 6 3 种特征选择算法的综合对比

(a)：分类精度；(b)：特征维数和运算时间

Fig. 6 Comprehensive comparison of three feature selection algorithms

(a)：Classification accuracy；(b)：Feature dimension and runtime

[图 6(b)]，ReliefF 算法在 3 个数据集上的特征维数均最高；ReliefF-RFE 算法在 3 个高光谱数据集中筛选的特征维数远低于 ReliefF 算法，平均特征维数是 ReliefF 算法的 37%，由此可以证明，高维数据集中含有与地物类别无关的特征及互相关性较高的冗余特征，结合封装式算法的特征选择能够去除过滤式算法中的冗余特征；RFE 算法在 3 个数据集上拥有最高的时间复杂度，ReliefF-RFE 算法在 3 个高光谱数据集中的运算时间均远低于单一的 RFE 算法，平均运算时间是 RFE 算法的 75%，这是由于 RFE 算法每次都要结合分类器进行迭代，造成耗时过多，ReliefF-RFE 算法则在 ReliefF 算法筛选结果的基础上降低了数据集的特征数，减小了特征子集的搜索范围，从而大幅降低了时间复杂度。总体而言，ReliefF-RFE 算法能够克服过滤式 ReliefF 算法无法有效减小特征之间冗余以及封装式 RFE 算法时间复杂度较高的缺陷，在分类精度、特征维数、运算时间共 3 个方面拥有均衡的综合性能。

3　结　论

针对高光谱图像波段连续且波段间相关性强的特点，在波段信息重复率较大的情况下，如果直接将原始波段信息用于地物分类，可能存在较多与地物类别不相关的冗余特征，由此引起“休斯现象”，影响模型的分类性能。在保证分类精度的情况下，本文提出一种将过滤式 ReliefF 算法和封装式 RFE 算法结合的 ReliefF-RFE 算法，并采用 Indian pines，Salinas-A 与 KSC 等 3 个标准高光谱数据集进行试验分析，结果表明：

(1)在 3 个高光谱数据集中，ReliefF-RFE 算法的平均总体精度(OA)为 92.94%、F-measure 为 92.81%，Kappa 系数为 91.94%，优于 ReliefF 算法但劣于 RFE 算法，总体精度良好。

(2)较之于 ReliefF 算法，ReliefF-RFE 算法继承了 RFE 算法的降维优势，能够大幅降低特征维数，其平均特征维数是 ReliefF 算法的 37%。

(3)较之于 RFE 算法，ReliefF-RFE 算法继承了 ReliefF 算法运算效率高的特点，运算时间是 RFE 算法平均运算时间的 75%。

综合来看，本文所提出的 ReliefF-RFE 算法能够发挥 ReliefF 和 RFE 的优点并规避两种算法的对应不足，将其应用于高光谱图像分类的特征选择，能够在保证分类精度的情况下，克服过滤式 ReliefF 算法无法有效减小特征之间冗余以及封装式 RFE 算法时间复杂度较高的缺陷，具有更为均衡的综合性能。ReliefF-RFE 算法能够在特征选择时间要求较高的情况下，用最小的代价，处理大样本的高光谱数据集。

References

[1] DU Pei-jun, XIA Jun-shi, XUE Chao-hui, et al(杜培军, 夏俊士, 薛朝辉, 等). National Remote Sensing Bulletin(遥感学报), 2016, 20(2): 236.

[2] REN Xiao-dong, LEI Wu-hu, GU Yu, et al(任晓东, 雷武虎, 谷　雨, 等). Computer Science(计算机科学), 2015, 42(S2): 162.

[3] ZHANG Bing(张　兵). National Remote Sensing Bulletin(遥感学报), 2016, 20(5): 1062.

[4] LI Zhi-qin, DU Jian-qiang, NIE Bin, et al(李郅琴, 杜建强, 聂　斌, 等). Computer Engineering and Applications(计算机工程与应用), 2019, 55(24): 10.

[5] Ren J S, Wang R X, Liu G, et al. Remote Sensing, 2020, 12(7): 1104.

[6] Ye M C, Xu C X, Chen H, et al. International Journal of Wavelets, Multiresolution and Information Processing, 2019, 17(5): 17.

[7] ZHANG Dong-yan, YANG Yu-ying, HUANG Lin-sheng, et al(张东彦, 杨玉莹, 黄林生, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2021, 37(9): 110.

[8] LIU Dai-chao, LI Xiao-song, LI Xiang-chen, et al(刘代超, 李晓松, 李向晨, 等). Bulletin of Surveying and Mapping(测绘通报), 2020, (8): 5.

[9] XU Ling-ling, CHI Dong-xiang(徐玲玲, 迟冬祥). Computer Engineering and Applications(计算机工程与应用), 2020, 56(24): 12.

[10] Marwa H, Slim B, Chih C H, et al. Memetic Computing, 2019, 11(2): 193.

[11] Yuan M S, Yang Z J, Huang G Z, et al. Pattern Recognition Letters, 2017, 92: 17.

[12] Lin X H, Li C, Ren W J, et al. Computational Biology and Chemistry, 2019, 83: 107149.

[13] DING Si-fan, WANG Feng, WEI Wei(丁思凡, 王　锋, 魏　巍). Computer Science(计算机科学), 2021, 48(4): 91.

[14] ZHANG Xiao-nei, ZHAI Wen-peng, HOU Hui-rang, et al(张小内, 翟文鹏, 侯惠让, 等). Journal of Electronics & Information Technology(电子与信息学报), 2021, 43(7): 2032.

[15] WANG Xiang, HU Xue-gang(王　翔, 胡学钢). Journal of Computer Applications(计算机应用), 2017, 37(9): 2433.

Construction and Application of ReliefF-RFE Feature Selection Algorithm for Hyperspectral Image Classification

XIANG Song-yang[1,3], XU Zhang-hua[1,2,4,5,6*], ZHANG Yi-wei[1,2], ZHANG Qi[1,3], ZHOU Xin[1,2], YU Hui[1,3], LI Bin[1,2], LI Yi-fan[1,2]

1. Research Center of Geography and Ecological Environment, Fuzhou University, Fuzhou 350108, China
2. College of Environmental and Safety Engineering, Fuzhou University, Fuzhou 350108, China
3. The Academy of Digital China, Fuzhou University, Fuzhou 350108, China
4. Fujian Provincial Key Laboratory of Resources and Environment Monitoring & Sustainable Management and Utilization, Sanming University, Sanming 365004, China
5. Key Laboratory of Spatial Data Mining & Information Sharing, Ministry of Education, Fuzhou 350108, China
6. Postdoctoral Research Station of Information and Communication Engineering, Fuzhou University, Fuzhou 350108, China

Abstract Hyperspectral images are characterized by continuous bands, high dimensionality, large data volume and strong correlation between adjacent bands, which can provide richer detailed information for feature classification. However, there is a lot of redundant information and noise in data, and the direct use of all band features without effective analysis and selection in image classification will lead to low computational efficiency and high computational complexity, and the "Hughes phenomenon" that the classification accuracy may increase and then decrease with the increase of band dimension. In order to quickly extract a subset of features with good recognition ability from hyperspectral images with tens or even hundreds of bands to avoid the "dimensional disaster". This paper combines the filtered ReliefF algorithm and the wrapped recursive feature elimination algorithm (Recursive feature elimination, RFE) to build the ReliefF-RFE feature selection algorithm, which can be used for feature selection in hyperspectral image classification. The algorithm uses the ReliefF algorithm to quickly eliminate many irrelevant features based on weight thresholds to narrow and optimize the range of feature subsets. The RFE algorithm is used to further search for the optimal feature subsets, and the recursive elimination of the less relevant features and redundant to the classifier in the narrowed feature subsets is performed to obtain the feature subsets with the best classification performance. In this paper, three standard datasets, including the Indian pines dataset, Salinas-A dataset and KSC dataset, are used as experimental data to compare the application effect of the ReliefF-RFE algorithm with ReliefF and RFE algorithms. The results show that the hyperspectral image classification by applying the ReliefF-RFE algorithm has an average overall accuracy (OA) of 92.94%, F-measure of 92.81%, and Kappa coefficient of 91.94%; in the three datasets, the average feature dimension of ReliefF-RFE algorithm is 37% of that of ReliefF algorithm, while the average operation time is 75% of that of the RFE algorithm. It shows that the ReliefF-RFE algorithm can ensure the classification accuracy while overcoming the defects of the filtered ReliefF algorithm, which cannot effectively reduce the redundancy among features and the wrapped RFE algorithm, which has high time complexity and has a more balanced comprehensive performance, which is suitable for feature selection in hyperspectral image classification.

Keywords Hyperspectral image; Feature selection; ReliefF algorithm; RFE algorithm; ReliefF-RFE algorithm

(Received Oct. 10, 2021; accepted Jan. 16, 2022)

* Corresponding author