基于 XGBoost 的机载激光雷达与高光谱影像结合的特征选择算法

张爱武[1,2*]，董喆[1,2]，康孝岩[1,2]

[1]首都师范大学三维信息获取与应用教育部重点实验室，北京 100048；

[2]首都师范大学空间信息技术教育部工程研究中心，北京 100048

摘要 为了解决地物分类的机载激光雷达(LiDAR)与高光谱特征构造中存在的特征维数过高的问题，提出了一种基于 XGBoost 与皮尔逊相关系数相结合的特征选择算法——XGB-PCCS，同时设计了 XGBoost 与序列后向选择相结合的特征选择算法——XGB-SBS 与之对比。采用真实数据验证所设计的两种算法，结果表明：两种算法均可在保证分类结果准确率的基础上有效地减小特征集维数；XGB-SBS 算法保留的特征维度为 33，得到的总体分类精度为 95.63%，Kappa 系数为 0.943；XGB-PCCS 算法保留的特征维度为 25，总体分类精度为 95.55%，Kappa 系数为 0.942。XGB-PCCS 算法的人为干预程度较低，运行时间较短，保留的特征集更精简。此外，对比了两种算法得到的特征子集，并总结了 LiDAR 点云与高光谱影像多模态特征构造中重要程度较高的 24 种特征。

关键词 遥感；特征选择；XGBoost 算法；皮尔逊相关系数；机载激光雷达；高光谱图像

中图分类号 P237 **文献标识码** A **doi**：10.3788/CJL201946.0404003

Feature Selection Algorithms of Airborne LiDAR Combined with Hyperspectral Images Based on XGBoost

Zhang Aiwu[1,2*], Dong Zhe[1,2], Kang Xiaoyan[1,2]

[1]*Key Laboratory of 3D Information Acquisition and Application, Ministry of Education, Capital Normal University, Beijing* 100048, *China*;

[2]*Engineering Research Center of Space Information Technology, Ministry of Education, Capital Normal University, Beijing* 100048, *China*

Abstract In order to solve the problem of high feature dimension in the feature construction of airborne light detection and ranging (LiDAR) and hyperspectral images for the classification of ground objects, we propose a feature selection algorithm based on extreme gradient boosting (XGBoost) combined with Pearson correlation coefficients (PCCS), named XGB-PCCS. Meanwhile, another feature selection algorithm based on XGBoost combined with sequential backward selection (SBS), named XGB-SBS, is designed to compare with XGB-PCCS. The real data is used to verify the two algorithms designed above. The results show that both algorithms can effectively reduce the dimension of feature sets on the basis that the accuracy of classification results is ensured. As for the XGB-SBS algorithm, the retained feature dimension is 33, the overall classification accuracy is 95.63%, and the Kappa coefficient is 0.943. In contrast, as for the XGB-PCCS algorithm, the retained feature dimension is 25, the overall classification accuracy is 95.55%, and the Kappa coefficient is 0.942. The XGB-PCCS algorithm has low degree of human intervention and short running time, and the retained feature set is compact. In addition, the feature subsets obtained by the two algorithms are compared, and 24 kinds of features with high importance in the multi-modal feature construction of LiDAR point cloud and hyperspectral images are summarized.

Key words remote sensing; feature selection; XGBoost algorithm; Pearson correlation coefficient; light detection and ranging (LiDAR); hyperspectral images

OCIS codes 280.3640; 100.2960; 100.4145

收稿日期：2018-10-23；**修回日期**：2018-12-26；**录用日期**：2018-12-29

基金项目：国家自然科学基金(41571369)、国家重点研发计划(2016YFB0502500)、北京市自然科学基金(4162034)、青海省科技计划(2016-NK-138)、科技创新服务能力建设-基本科研业务费(科研类)项目(025185305000/143)

* **E-mail**：zhangaw98@163.com

1 引　言

高光谱遥感又称成像光谱遥感，可同时获取目标区域的一维光谱信息和二维几何空间信息[1]，具有图谱合一的特点与优势，能够精确描述地物的光谱和纹理特性[2]。机载激光雷达（LiDAR）是一种主动式的对地观测技术，能够实时获取地形表面的三维空间信息[2-4]。上述两种类型数据提取出的特征在水平和垂直位置上的精度、信息都有其各自的优缺点[5]。传统的基于单一类型数据特征的分类研究往往只能表现出研究对象的部分特性，具有相当的局限性[6]。高光谱数据可以提供更多水平空间上的信息，如更丰富的光谱和纹理信息，但在垂直方向上则无法提供更多信息。而 LiDAR 数据对于描述三维空间信息更有优势，但激光点的密度对精确分类有一定程度的影响。将上述两种数据的特征进行联合，不同种类的特征信息之间可以相互补充，因此表达出的语义信息会更加丰富。同时，上述两类特征之间所具有的相关性也在一定程度上有助于提升地物分类结果的准确率。

但从另一方面看，多种不同数据的特征联合必然会导致特征集维数大幅增加，而如何有效地从高维特征数据集中提取或选择出最佳的特征子集用于后续的地物分类，得到最佳的分类结果，也是众多研究领域的重点。特征选择是一个不断搜索并择优的过程，是从一组特征集中挑选出一部分最有效的特征以降低特征空间维数，并在一定程度上提高分类精度的过程[7]。在实际应用中，通常采用启发式搜索算法在运算效率与特征子集的质量之间寻找一个较好的平衡点，即近似最优解[8]。合适的特征选择方法不仅可以简化模型，缩短样本的训练时间，还可以通过降低过拟合增强学习模型的泛化性能，有助于提高分类结果的准确率。

XGBoost 是 Chen 等[9]于 2015 年提出的一种基于梯度 Boosting 的集成学习算法，该算法可自动利用中央处理器（CPU）多线程进行并行计算，是近年来新兴的高效算法。XGBoost 不但可以用于分类，还可以通过统计得到特征变量的重要性及排序。此外，虽然该算法便于对各个特征的重要性进行分析，但却不利于进行准确的特征选择。

序列后向选择（SBS）算法[8]是指从特征全集开始，每次从特征集中剔除一个特征，使得剔除该特征后的评价函数值达到最优。该算法在特征数量不是非常大的情况下具有较高的分类准确率[10]，但其却无法考虑特征之间的相关性，对于相关性较高的特征不能一并剔除，从而导致存在冗余的排序步骤。而且，该算法属于贪心算法，容易陷入局部最优值。

皮尔逊相关系数（PCC）可用于计算两个连续型变量之间的相关性，是反映两变量相关关系方向和密切程度的指标。由于特征计算值在一定的区间范围内可以认为是连续的[11]，因此本文选取皮尔逊相关系数来衡量各个特征之间的相关程度，从而对特征选择步骤起到一定程度的简化作用。

本文首先针对完全匹配的 LiDAR 点云和高光谱影像构造并提取了 46 种特征，然后设计了一种基于 XGBoost 特征重要性排序并结合皮尔逊相关系数的特征选择（XGB-PCCS）算法，该算法综合考虑了特征向量相关关系的方向及密切程度，从而对特征进行选择；此外，设计了 XGBoost 与序列后向选择相结合的多模态特征选择（XGB-SBS）算法与之对比；最后采用真实数据对本文设计的两种算法进行了验证和分析，并从选择时长、选择出的特征子集构造以及后续分类精度三方面对上述两种算法进行对比和评估。

2 特征提取

2.1 LiDAR 数据特征的提取

LiDAR 数据特征可分为两部分。一部分是来自 LiDAR 点云的自身特征，称作直接特征，如高程、回波及强度信息等，通常可以从点云数据中直接读取（如 *.las 格式），提取过程相对简单[3]；另一部分是 LiDAR 数据的衍生特征，需要对点云数据进行局部统计，从而提取出相应的几何特征。

2.1.1 直接特征

直接特征包括高程 F_{H}、回波强度 F_{I}、回波次数 F_{NR}、回波编号 F_{RN}[11-12]。

2.1.2 衍生特征

点云的几何特征可以通过对某一个点的三维邻域内的所有点进行统计分析得到。本课题组采用球体邻域对 LiDAR 点云进行统计计算，分别提取得到高程相关、投影相关、面相关及协方差相关 4 种衍生特征[11]。其中：高程相关特征包括邻域高程差 F_{Z}、邻域高程均值 F_{Hm}、邻域高程方差 F_{Hvar}；投影相关特征包括 xy 平面最小外接矩面积 $F_{\mathrm{S}xy}$、xz 平面最小外接矩面积 $F_{\mathrm{S}xz}$、yz 平面最小外接矩面积 $F_{\mathrm{S}yz}$；面相关特征包括竖直角 F_{VA}、邻域竖直角均值 F_{VAm}、法向散射系数 F_{Ns}、点至拟合平面的距离 F_{D}、平面拟合残差 F_{Ds}、表面系数 F_{S}；协方差矩阵相关

特征包括特征值($F_{\lambda1}$、$F_{\lambda2}$、$F_{\lambda3}$)以及各向异性系数F_{Ac}、平面指数F_{Pc}、线性指数F_{Lc}、球面指数F_{Sc}、信息熵F_{Ec}、结构张量总变化指数F_{Oc}。

上述各特征的提取方法可参考文献[11]。

2.2　高光谱影像特征的提取

2.2.1　光谱特征

1) 直接光谱特征

主成分分析(PCA)是一种较为常用且有效的降维方法。PCA变换是一种线性变换,通过对光谱数据进行矩阵变换尽量保留原有数据的信息,变换后得到的主成分分量之间彼此不相关,且随着主成分编号的递增,各分量所包含的信息量递减[13-14]。

本课题组对实验数据集中的高光谱影像进行PCA变换后得到的前10个波段的信息量为99.97%,故仅选取前10个波段作为提取到的直接光谱特征进行后续特征的选择。

2) 间接光谱特征

单波段遥感影像局部邻域内像素值的统计信息,如均值和方差,也可以作为光谱特征加入到特征向量中参与分类研究[3]。为避免重复提取相似特征,本课题组只对PCA变换后获取到的首个波段进行邻域均值和方差的计算,邻域窗口的大小选择为3×3,分别记为F_{PCA_m}和F_{PCA_var}。

由于本课题组选用的数据集中的植被类型较为单一,且植被总面积仅占整个数据集的12%,故未提取高光谱影像的植被指数特征。

2.2.2　纹理特征

纹理是灰度分布在空间位置上反复出现而形成的,因而在图像空间中相隔某一距离的两个像素之间会存在一定的灰度关系,即为图像中灰度的空间相关特性[15]。灰度共生矩阵(GLCM)就是一种常用的纹理分析方法,是图像中两个像素点灰度级联合分布的统计形式[16],反映了图像灰度关于方向、领域、变化幅度的综合信息,能够较好地反映纹理灰度级之间的相关性规律。

本课题组选用大小为3×3的窗口,沿0°、45°、90°及135°这4个方向角以1 pixel位移距离进行统计,仅实现灰度共生矩阵角二阶矩F_{GLCM_ASM}、信息熵F_{GLCM_Ent}、惯性矩F_{GLCM_Hom}、相异性F_{GLCM_Dis}、对比度F_{GLCM_Con}、相关性F_{GLCM_Cor},以及归一化灰度矢量(GLDV)角二阶矩F_{GLDV_ASM}、对比度F_{GLDV_Con}、均值F_{GLDV_Mean}共9种纹理特征的提取[15-19]。

3　特征选择

特征选择是一个不断搜索并择优的过程,是从一组特征集中挑选出一部分最有效的特征以降低特征空间维数,并在一定程度上提高分类精度的过程[7]。在实际应用中,通常采用启发式搜索算法,在运算效率与特征子集的质量之间寻找一个较好的平衡点,即为近似最优解[8]。

特征选择的目的有三:便于构建用时更短、消耗更低的预测分类模型;能够使模型具有更好的理解性和解译性;提高分类预测的准确性。

下面依次介绍XGBoost算法、XGB-SBS算法和XGB-PCCS算法。

3.1　XGBoost算法的原理及其特征重要性量度

Boosting算法是数据挖掘领域中比较流行且有效的集成学习算法,通过将各个弱分类器加权叠加形成强分类器来有效降低误差,得到精确度更高、更准确的分类结果[20]。Gradient Boosting算法是在Boosting算法的基础上改进而来的,其思想是不断地降低残差,使先前模型中的残差在梯度方向上进一步降低,从而得到新的模型[20]。XGBoost则是在Gradient Boosting算法的基础上进行改进后的算法,是由Chen等[9]于2015年提出的,全称为eXtreme Gradient Boosting。XGBoost算法对损失函数进行的是二阶泰勒展开,在目标函数中加入了树模型复杂度作为正则项,在训练过程中借鉴随机森林的思想,即每次迭代过程中对样本进行抽样,采用部分样本的部分特征去训练的方法,并充分利用了多核CPU并行计算的优势,大幅提高了模型的运算速度和预测分类精度[21-22]。

此外,XGBoost算法可以统计出每个特征变量的重要性,并据此得到特征的重要性排序。排序结果便于对特征的重要性进行分析,但并不利于准确地进行特征选择。当前较为常见的做法是在重要性排序之后人为选定前k个特征,重新训练模型并预测分类。该方法虽然简单有效,但人为干预性较强,且选定的特征数量及特征种类不一定为最佳的特征子集。因此,本课题组提出了一种基于XGBoost特征重要性排序联合皮尔逊相关系数计算的特征选择方法,以达到确定最佳特征子集的目的。

XGBoost算法中的特征重要性计算标准如下:weight指该特征用来切割树结点的次数;gain指该特征被用来切割树结点时所产生的平均增益;cover指该特征在树结构内被应用的平均覆盖率。本实验中所应用到的重要性排序均选用weight作为重要性得分的计算标准。

3.2　XGB-SBS 算法

序列后向选择算法是指从特征全集开始，每次从特征集中剔除一个特征，使剔除该特征后的评价函数值达到最优。该算法属于贪心算法，容易陷入局部最优值。

本课题组设计的 XGB-SBS 算法首先利用 XGBoost 算法中特征变量的重要性量度对特征进行重要性排序，采用序列后向搜索算法[8]依次从当前特征集中剔除该轮迭代中重要性得分最低（即排序最靠后）的特征，剩余的保留特征重新进行新一轮的预测和排序，记录每轮迭代过程中的预测分类准确率，并将其作为评价函数值，用于确定预测分类准确率最高的保留特征子集，即为最终特征选择的结果。具体算法流程如图 1 所示。

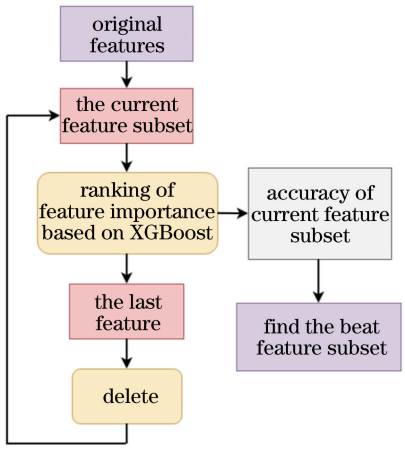

图 1　XGB-SBS 算法流程图

Fig. 1　Flow chart of XGB-SBS algorithm

为保证模型训练的可靠性和稳定性，每轮迭代中对训练集进行 10 次随机抽取。每次抽取出训练集的 70%用于训练模型及特征集重要性排序，并保证每次抽取出的样本类别比例与整个训练集中样本类别的比例相同，训练集中剩余的 30%用于求取当前次随机抽取的分类精度。将分类精度最高轮次的重要性排序中最末位的特征作为下一轮待剔除的特征，并将 10 次抽取得到的 10 个分类精度的平均值作为当前轮次保留特征集下的预测分类精度，用以找出预测分类精度最高的轮次，从而确定 XGB-SBS 算法下的最佳特征子集。

该算法的核心是根据重要性排序将全部特征逐一剔除，可通过迭代更为准确地得到各个特征的重要性排序，同时可以准确地得到测试分类精度最大的保留特征集。但该算法的缺陷在于无法考虑特征之间的相关性，对于相关性较高的特征不能一并剔除，从而导致存在冗余的排序步骤，因而当数据量较大时，该方法的运行选择时间可能会比较长。

3.3　XGB-PCCS 算法

两个连续型变量之间的相关性通常可用皮尔逊相关系数进行计算，它是反映两变量相关关系的方向和密切程度的指标。由于特征计算值在一定区间范围内可以认为是连续的[11]，故本课题组选用皮尔逊相关系数来衡量各个特征之间的相关程度。故可将与每轮重要性排序最末位特征的相关系数绝对值较大且排位靠后的特征视为同等不重要特征，在下一轮训练和预测中一并剔除，从而对特征选择步骤起到一定程度的简化。

基于上述 XGB-SBS 算法提出如下改进：在每轮确定即将剔除的特征之后，计算该剔除特征与当前保留特征集中其他特征的皮尔逊相关系数，将所有相关系数的绝对值的均值作为阈值，绝对值大于阈值且在该轮重要性排序中位于后 50%的特征，在下一轮中也一并剔除。XGB-PCCS 算法流程如图 2 所示。

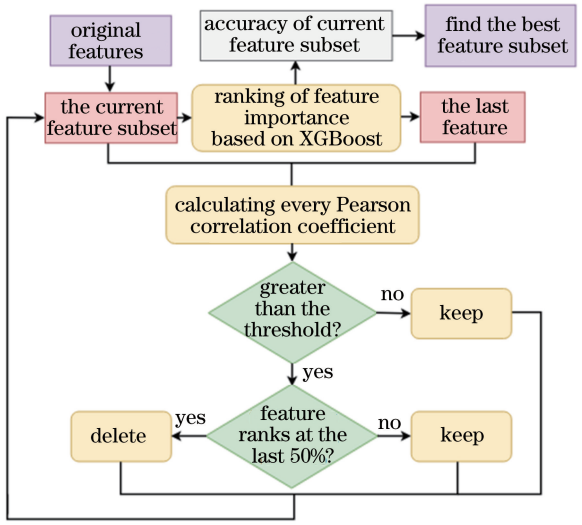

图 2　XGB-PCCS 算法流程图

Fig. 2　Flow chart of XGB-PCCS algorithm

改进后的算法通过计算与每次重要性排序最末位特征的皮尔逊相关系数，将与其相关程度较高的特征一并剔除。改进后的算法与常用的人为选定重要性排序前 k 个特征的选择方法相比更为准确，人为干预程度更小，同时也减少了运行分类器建立模型及进行重要性量度和排序的次数，在数据量较大的情况下可明显缩短特征选择的时间。

4　实验结果及分析

4.1　研究区域及实验数据集

本研究选用的数据集 grss_dfc_2018 含机载

LiDAR 数据、高光谱影像以及地表真实分类，如图 3 所示，均来自 2018 IEEE GRSS Data Fusion Contest（http：// www. grss-ieee. org/community/technical-committees/data-fusion/），该数据由休士顿大学高光谱图像分析实验室和国家机载激光测绘中心（NCALM）提供。数据获取的时间为 2017 年 2 月 16 日，16：31—18：18（GMT），地点位于休士顿大学校园及其邻近地区。

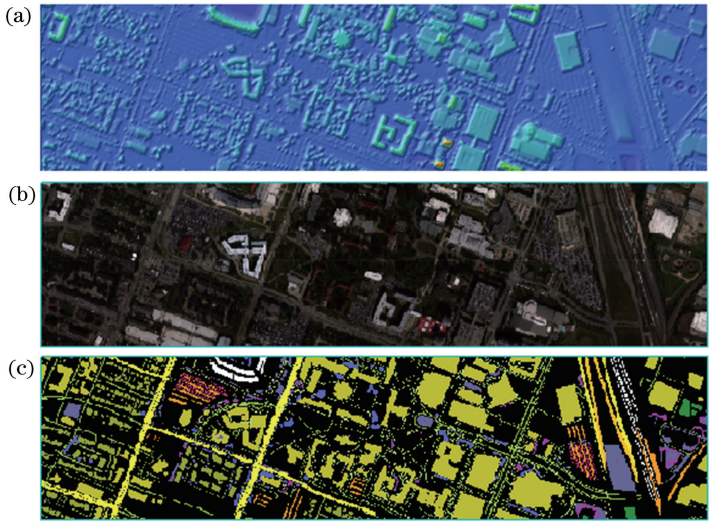

图 3　研究区域图。(a) LiDAR 衍生 DSM；(b)高光谱影像；(c)地表真实分类

Fig. 3　Map of research area. (a) LiDAR derived digital surface model; (b) hyperspectral image; (c) surface truth classification

机载 LiDAR 传感器的飞行高度为 500 m（相对地面高度），扫描条带宽度为 445 m，重叠率为 50%，行距为 225 m，点云数据的平均密度为 32 point/m^2。搭载的 OpTeTi TIN MW 运行了 3 个不同波长的激光通道，本课题组仅选用了通道 1 的数据，即基于 1550 nm 的激光得到的点云数据。

高光谱搭载传感器飞行高度为 2100 m（相对地面高度），条带宽度为 1500 m，平均飞行速度为 65 m/s。该高光谱数据包含 48 个波段，光谱范围为38～1050 nm，空间分辨率为 1 m。

首先对本课题组选用的研究数据集进行特征提取，其中 LiDAR 数据特征有 25 种，高光谱影像特征有 21 种，共计 46 种特征。另外，该数据集共包含 20 个类别，从数据集提供的各类样本中随机抽取 30%作为训练集，剩余的 70%用于精度评价，样本分类及各类别样本点的数量详见表 1。

4.2　XGB-SBS 算法选择的结果

XGB-SBS 算法选择出的特征数量与训练集分类精度之间的关系如图 4 所示。实验结果表明，特征集中重要程度较低的特征逐一被剔除，随特征集维度逐渐减小，前期验证精度的变化不太明显，但当特征被剔除到一定数量后，验证精度明显大幅下降。在此过程中，验证精度最高处所应用的特征集即为该算法下的最佳特征子集。

本次实验最佳的保留特征数为33，训练集的预

表 1　训练集与测试集中的样本类别及样本数量

Table 1　Class name and number of samples in training and test sets

No.	Class name	Number of sample points in training set	Number of sample points in test set
1	Healthy grass	2940	6859
2	Stressed grass	9751	22751
3	Artificial turf	205	479
4	Evergreen trees	4078	9517
5	Deciduous trees	1506	3515
6	Bare earth	1355	3161
7	Water	80	186
8	Residential buildings	11932	27840
9	Non-residential buildings	67125	156627
10	Roads	13760	32106
11	Sidewalks	10209	23820
12	Crosswalks	455	1063
13	Major thoroughfares	13904	32444
14	Highways	2959	6906
15	Railways	2081	4856
16	Paved parking lots	3450	8050
17	Unpaved parking lots	44	102
18	Cars	1964	4583
19	Trains	1611	3758
20	Stadium seats	2047	4777
	Total	151456	353400

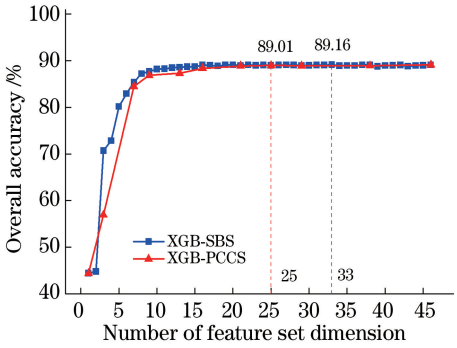

图 4　不同选择算法保留的总体分类精度与特征集维数的关系

Fig. 4　Relationship between overall classification accuracy and number of feature set dimension retained by different selection algorithms

测精度为 89.16%，保留特征见表 2，每次排序仅剔除一个特征，剔除次序及剔除特征见表 3。

表 2　不同特征选择算法下的最佳特征

Table 2　Optimal features at different feature selection algorithms

Algorithm	Number	Optimal features
XGB-SBS	33	F_{H}, F_{I}, F_{NR}, F_{Z}, F_{Hm}, F_{Hvar}, $F_{S\,xy}$, F_{Syz}, F_{VAm}, F_{Ns}, $F_{\lambda 1}$, $F_{\lambda 2}$, $F_{\lambda 3}$, F_{Ac}, F_{Pc}, F_{Sc}, F_{Ec}, F_{Oc}, $F_{PCA\,1\text{-}10}$, F_{PCA_m}, F_{PCA_var}, F_{GLCM_Con}, F_{GLCM_Dis}, F_{GLCM_Cor}
XGB-PCCS	25	F_{H}, F_{I}, F_{Z}, F_{Hm}, F_{Hvar}, F_{VA}, F_{VAm}, F_{Ns}, F_{Ac}, F_{Pc}, F_{Ec}, $F_{PCA\,1\text{-}10}$, F_{PCA_m}, F_{PCA_var}, F_{GLCM_Con}, F_{GLCM_Cor}

4.3　XGB-PCCS 算法选择的结果

XGB-PCCS 算法选择下的特征数量与训练集分类精度之间的关系如图 4 所示。实验结果表明，通过计算与每次重要性排序最末位特征的皮尔逊相关系数，逐次将相关程度较高且当前轮次重要性排名靠后的特征一并剔除，可得到一个训练集预测精度的峰值，此时所保留的特征即为该算法得到的最佳保留特征子集。

表 3　基于 XGB-SBS 算法求最佳特征子集时的特征剔除次序

Table 3　Order of feature deletion for finding optimal feature subset based on XGB-SBS algorithm

Order	Deleted features
1-5	F_{GLDV_Con}, F_{GLDV_Mean}, F_{GLDV_ASM}, F_{GLCM_ASM}, F_{GLCM_Ent}
6-10	F_{RN}, $F_{S\,xz}$, F_{Lc}, F_{D}, F_{GLCM_Hom}
11-13	F_{VA}, F_{S}, F_{Ds}

本次实验中最佳的保留特征数为 25，训练集预测精度为 89.01%，保留特征见表 2，每次排序可剔除多个特征，剔除次序及剔除特征见表 4。

首轮特征重要性排序中归一化灰度矢量对比度特征 F_{GLDV_Con} 排于最末位。该特征与其余 45 种特征之间的皮尔逊相关系数的绝对值 P 的关系如图 5 所示，相关系数总体差异显著，大部分特征的绝对值在 0～0.2 内，但同时也存在绝对值较大的特征。本轮求得皮尔逊相关系数绝对值的平均值约为0.1507，其中共有 11 种特征系数大于均值，可视为与待剔除的 F_{GLDV_Con} 特征相关程度较高，但由于 F_{H}、F_{Hm}、F_{Sc}、F_{PCA_var}、F_{GLCM_Con} 这 5 种特征在本轮重要性排序中的排名靠前，视为对分类结果影响较大的特征，故本轮仅剔除 F_{GLDV_Con} 及 F_{GLCM_Hom}、F_{GLCM_Dis}、F_{GLCM_Ent}、F_{GLCM_ASM}、F_{GLDV_Mean}、F_{GLDV_ASM} 共 7 种特征，剩余的 39 种特种进入下一轮训练和筛选。

表 4　基于 XGB-PCCS 算法求最佳特征子集时的特征剔除次序

Table 4　Order of feature deletion for finding optimal feature subset based on XGB-PCCS algorithm

Order	Deleted features	
	According to feature importance	According to Pearson correlation coefficient
1	F_{GLDV_Con}	F_{GLCM_Hom}, F_{GLCM_Dis}, F_{GLCM_Ent}, F_{GLCM_ASM}, F_{GLDV_Mean}, F_{GLDV_ASM}
2	$F_{S\,xz}$	$F_{\lambda 1}$, $F_{\lambda 2}$, $F_{\lambda 3}$, $F_{S\,yz}$, $F_{S\,xy}$, F_{RN}, F_{Oc}, F_{NR}
3	F_{Ds}	F_{S}, F_{Sc}, F_{Lc}, F_{D}

从最佳特征子集的结果来看，F_{PCA_var} 和 F_{GLCM_Con} 均是筛选过后的保留特征，可以体现出皮尔逊相关系数与特征重要性相结合作为特征筛选依据的重要性。比较表 3 和表 4 可知，XGB-PCCS 算法明显缩减了迭代排序轮次，理论上当数据量较大时可节省特征选择的运行时间。

4.4　分类精度评价

选用 XGBoost 分类器分别对原始 46 种特征、XGB-SBS 选择的 33 种特征子集、XGB-PCCS 选择的 25 种特征子集应用相同的参数进行分类预测。选用的评判预测分类精度的指标为总体分类精度 OA 和 Kappa 系数。本次实验使用的是轻便型笔记本计算机（型号：华硕 S56C；CPU：Intel Core i5 3317U，主频 1.7 GHz；内存 4 GB）

两种特征选择算法的运行选择时间及分类预测精度见表 5。

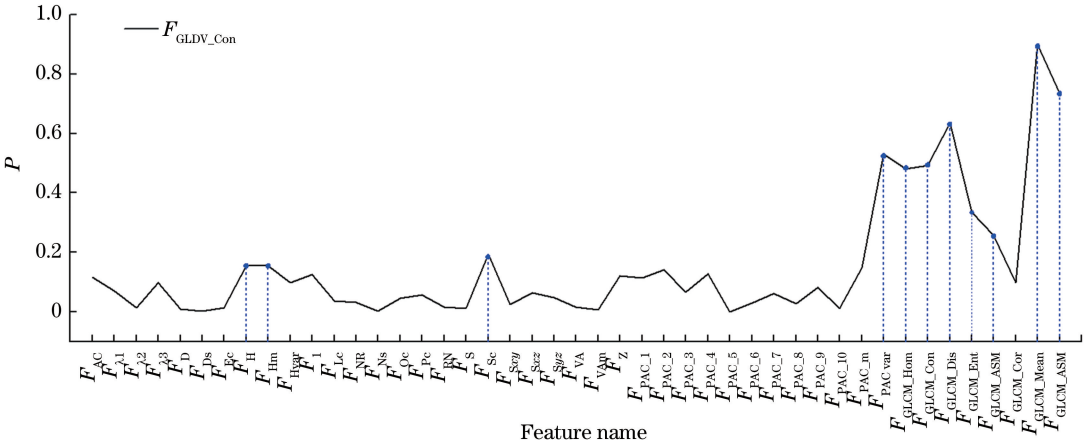

图 5　F_{GLDV_Con} 与其余 45 种特征的皮尔逊相关系数的绝对值

Fig. 5　Absolute values of Pearson correlation coefficients between F_{GLDV_Con} and other 45 kinds of features

表 5　两种特征选择算法的运行选择时间及分类预测精度

Table 5 Running time consumption and predicted classification accuracy of two feature selection algorithms

Result	XGBoost	XGB-SBS	XGB-PCCS
Time consumption /s		17818	5255
Dimension of feature subset	46	33	25
OA /%	95.53	95.63	95.55
Kappa	0.942	0.943	0.942

由表 5 可以看出：XGB-SBS 算法进行特征选择之后得到的分类结果相比未经过特征选择的总体分类精度提高了 0.1%，Kappa 系数提升了 0.001；XGB-PCCS 算法进行特征选择之后得到的分类结果相比未经过特征选择的总体分类精度提高了 0.02%，Kappa 系数则保持不变。虽然两种特征选择方法相较于不经过特征选择直接进行分类所得到的分类精度提升不大，但应用的特征数量皆有减少，可以认为是在削减特征维数的基础上保证了分类结果的准确率。

另一方面，XGB-PCCS 算法通过计算与每次重要性排序最末位特征的皮尔逊相关系数并将与其相关程度较高的特征一并剔除的方法，减少了运行分类器建立模型及进行重要性量度和排序的次数，在数据量较大的情况下可明显缩短特征选择的时间。

4.5　面向地物分类的多模态特征构造

表 2 分别列出了 XGB-SBS 与 XGB-PCCS 两种算法选择出的最佳特征子集的构造。通过对比可以看出，高程 F_{H}、回波强度 F_{I}、邻域高差 F_{Z}、邻域高程均值 F_{Hm}、邻域高程方差 F_{Hvar}、邻域竖直角均值 F_{VAm}、法向散射系数 F_{Ns}、各向异性系数 F_{Ac}、平面指数 F_{Pc}、信息熵 F_{Ec}，PCA 变换得到的前 10 个主成分 $F_{PCA1\text{-}10}$，PCA 变换后获取到的首个波段的邻域均值 F_{PCA_m} 和邻域方差 F_{PCA_var}，灰度共生矩阵对比度 F_{GLCM_Con} 和相关性 F_{GLCM_Cor} 这 24 个特征是两个特征子集都包含的，故认为对于本研究所用数据集而言，这 24 个特征是对地物分类预测起决定作用的特征。而回波次数 F_{NR}，xy 与 yz 平面最小外接矩面积 F_{Sxy}、F_{Syz}，竖直角 F_{VA}，协方差特征值 $F_{\lambda1}$、$F_{\lambda2}$、$F_{\lambda3}$，球面指数 F_{Sc}，结构张量总变化指数 F_{Oc} 及灰度共生矩阵相异性 F_{GLCM_Dis} 这 10 个特征相较于本课题组提取的剩余特征而言也是相对比较重要的。

5　结　　论

针对完全匹配的 LiDAR 点云及高光谱影像两种不同类型的数据共构造并提取了 46 种特征，然后设计了 XGBoost 与序列后向选择相结合的多模态特征选择算法 XGB-SBS，并在该算法的基础上结合皮尔逊相关系数设计了一种新的特征选择算法 XGB-PCCS。

XGB-SBS 算法选择出的特征子集由原始的 46 个下降至 33 个，得到的分类结果总体分类精度为 95.63%，Kappa 系数为 0.943。该算法可以通过迭代得到更准确的特征重要性排序，同时可以确定原始 46 个特征中可以使测试分类精度达到最大的保留特征集。但该算法的缺点在于没有考虑特征之间的相关性，对于相关性较高的特征不能一并剔除，从而导致存在冗余的排序步骤，当数据量较大时，该算法的运行时间可能会比较长。

XGB-PCCS 算法在综合考虑了特征向量相关关系的方向及密切程度后对特征做出选择。选择出的特征子集的大小由原始的 46 个缩减至 25 个，得

到的总体分类精度为95.55%，Kappa系数为0.942。该算法的人为干预程度较低，运行时间较短，可以更加有效地缩减特征数量，但筛选出的特征子集对于后续分类精度的提升意义不大。

虽然本研究采用的两种特征选择算法选取的特征子集对分类准确率的提升都不显著，但保留的特征数量都有一定程度的缩减，都达到了在保证分类结果准确率的基础上有效削减特征维数的目的。两种算法相比较而言，XGB-PCCS算法保留的特征集相对更为精简，运行时间更短。

在后续的研究中可以考虑特征尺度对面向地物分类的多模态特征构造的影响，以及特征维数大幅增加后两种算法的选择结果对分类精度的影响是否更明显。

参　考　文　献

[1] Zhang B. Advancement of hyperspectral image processing and information extraction[J]. Journal of Remote Sensing, 2016, 20(5): 1062-1090.
张兵. 高光谱图像处理与信息提取前沿[J]. 遥感学报, 2016, 20(5): 1062-1090.

[2] Zhu J T, Huang R. Feature selection and classification of hyperspectral data and LiDAR data based on Adaboost[J]. Remote Sensing Information, 2014, 29(6): 68-72.
朱江涛, 黄睿. 基于Adaboost的高光谱与LiDAR数据特征选择与分类[J]. 遥感信息, 2014, 29(6): 68-72.

[3] Dong B G. Research on classification technologies of land cover by fusing airborne LiDAR point clouds and remote sensing imagery [D]. Zhengzhou: PLA Information Engineering University, 2013.
董保根. 机载LiDAR点云与遥感影像融合的地物分类技术研究[D]. 郑州：解放军信息工程大学, 2013.

[4] Pan S Y, Guan H Y. Object classification using airborne multispectralLiDAR data [J]. Acta Geodaetica et Cartographica Sinica, 2018, 47(2): 198-207.
潘锁艳, 管海燕. 机载多光谱LiDAR数据的地物分类方法[J]. 测绘学报, 2018, 47(2): 198-207.

[5] Cheng X J, Cheng X L, Hu M J, *et al*. Buildings detection and contour extraction by fusion of aerial images and LIDAR point cloud[J]. Chinese Journal of Lasers, 2016, 43(5): 0514002.
程效军, 程小龙, 胡敏捷, 等. 融合航空影像和LIDAR点云的建筑物探测及轮廓提取[J]. 中国激光, 2016, 43(5): 0514002.

[6] Zhang Q C, Tong G F, Li Y, *et al*. River detection in remote sensing images based on multi-feature fusion and soft voting[J]. Acta Optica Sinica, 2018, 38(6): 0628002.
张庆春, 佟国峰, 李勇, 等. 基于多特征融合和软投票的遥感图像河流检测[J]. 光学学报, 2018, 38(6): 0628002.

[7] Yao X, Wang X D, Zhang Y X, *et al*. Summary of feature selection algorithms [J]. Control and Decision, 2012, 27(2): 161-166, 192.
姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述[J]. 控制与决策, 2012, 27(2): 161-166, 192.

[8] Yao D J, Yang J, Zhan X J. Feature selection algorithm based on random forest[J]. Journal of Jilin University (Engineering and Technology Edition), 2014, 44(1): 137-141.
姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版), 2014, 44(1): 137-141.

[9] CHEN T, GUESTRIN C. Xgboost: a scalable tree boosting system[C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016 : 785-794.

[10] Liu M, Li Z R, Zhang H T, *et al*. Feature selection algorithm application in near-infrared spectroscopy classification based on binary search combined with random forest pruning[J]. Laser & Optoelectronics Progress, 2017, 54(10): 103001.
刘明, 李忠任, 张海涛, 等. 基于二分搜索结合修剪随机森林的特征选择算法在近红外光谱分类中的应用[J]. 激光与光电子学进展, 2017, 54(10): 103001.

[11] Zhang A W, Xiao T, Duan Y H. A method of adaptive feature selection for airborne LiDAR point cloud classification [J]. Laser & Optoelectronics Progress, 2016, 53(8): 082802.
张爱武, 肖涛, 段乙好. 一种机载LiDAR点云分类的自适应特征选择方法[J]. 激光与光电子学进展, 2016, 53(8): 082802.

[12] Liang X W. Research on feature selection and precise classification technique from LIDAR data [D]. Taiyuan: North University of China, 2015.
梁小伟. 机载LIDAR数据特征选择与精确分类技术研究[D]. 太原：中北大学, 2015.

[13] Ye Z, He M Y. PCA and windowed wavelet transform for hyperspectral decision fusion classification[J]. Journal of Image and Graphics, 2015, 20(1): 132-139.
叶珍, 何明一. PCA与移动窗小波变换的高光谱决策融合分类[J]. 中国图象图形学报, 2015, 20(1): 132-139.

[14] Zang Z, Lin H, Yang M H. Comparative study on

descending dimension classification of hyperspectral data between ICA algorithm and PCA algorithm[J]. Journal of Central South University of Forestry & Technology, 2011, 31(11): 18-22.
臧卓, 林辉, 杨敏华. ICA 与 PCA 在高光谱数据降维分类中的对比研究[J]. 中南林业科技大学学报, 2011, 31(11): 18-22.

[15] Li J, Yang Y Q, Shen W, *et al*. Research on fabric texture based on gray level co-occurrence matrix[J]. Advanced Textile Technology, 2013, 21(3): 12-16.
李静, 杨玉倩, 沈伟, 等. 基于灰度共生矩阵的织物纹理研究[J]. 现代纺织技术, 2013, 21(3): 12-16.

[16] Tian Y Q, Guo P, Lu H Q. Texture feature extraction of multiband remote sensing image based on gray level co-occurrence matrix[J]. Computer Science, 2004, 31(12): 162-163, 195.
田艳琴, 郭平, 卢汉清. 基于灰度共生矩阵的多波段遥感图像纹理特征的提取[J]. 计算机科学, 2004, 31(12): 162-163, 195.

[17] Ren G Z, Jiang T. Study on GLCM-based texture extraction methods[J]. Computer Applications and Software, 2014, 31(11): 190-192, 325.
任国贞, 江涛. 基于灰度共生矩阵的纹理提取方法研究[J]. 计算机应用与软件, 2014, 31(11): 190-192, 325.

[18] Bigdeli B, Samadzadegan F, Reinartz P. Fusion of hyperspectral and LIDAR data using decision template-based fuzzy multiple classifier system[J]. International Journal of Applied Earth Observation and Geoinformation, 2015, 38: 309-320.

[19] Bigdeli B, Pahlavani P. High resolution multisensor fusion of SAR, optical and LiDAR data based on crisp vs. fuzzy and feature vs. decision ensemble systems[J]. International Journal of Applied Earth Observation and Geoinformation, 2016, 52: 126-136.

[20] Jiang J W, Liu W G. Application of XGBoost algorithm in manufacturing quality prediction[J]. Intelligent Computer and Applications, 2017, 7(6): 58-60.
蒋晋文, 刘伟光. XGBoost 算法在制造业质量预测中的应用[J]. 智能计算机与应用, 2017, 7(6): 58-60.

[21] Yang C. Road network extraction from remote sensing images based on XGBoost[J]. Microcomputer & Its Applications, 2017, 36(24): 28-31.
杨灿. 基于 XGBoost 的遥感图像中道路网络的提取[J]. 微型机与应用, 2017, 36(24): 28-31.

[22] Li Y Z, Wang Z Y, Zhou Y L, *et al*. The improvement and application of Xgboost method based on the Bayesian optimization[J]. Journal of Guangdong University of Technology, 2018, 35(1): 23-28.
李叶紫, 王振友, 周怡璐, 等. 基于贝叶斯最优化的 Xgboost 算法的改进及应用[J]. 广东工业大学学报, 2018, 35(1): 23-28.