

顾及样本优化选择的机器学习云检测研究

张辉¹ 周仿荣² 徐真¹ 文刚² 马御棠² 韩旭^{3,*} 吴磊³

(1 云南电网有限责任公司, 昆明 650011)

(2 南方电网公司云南电网电力科学研究院电力遥感技术联合实验室, 昆明 650217)

(3 苏州深蓝空间遥感技术有限公司, 苏州 215505)

摘要 针对云层日变化、云类型、云相态、云光学厚度等特征差异带来的光谱差异, 导致传统阈值算法对云识别精度不高的问题, 文章提出了一种顾及样本优化选择, 耦合物理阈值方法和机器学习的云检测算法模型, 利用“葵花8号”卫星(Himawari-8)数据进行日间云检测。通过样本优化选择, 使样本中尽可能包括不同情形下的云特征, 为机器学习模型提供良好的样本基础, 增加模型泛化能力; 同时输入特征除了考虑反照率、亮温、亮温差以及天顶角等因素外, 还加入了基于反照率和亮温差的物理阈值方法云识别结果; 最后基于极限随机树模型进行云检测。结果表明: 模型云检测交叉验证精度为96.41%, 总漏检率和总虚检率分别为2.08%和0.91%; 通过云-气溶胶激光雷达与红外探路者卫星观测(CALIPSO)产品数据进行对比分析, 结果显示云检测总体精度为97.1%。

关键词 样本优化 极限随机树 机器学习 云检测 航天遥感

中图分类号: P237

文献标志码: A

文章编号: 1009-8518(2024)01-0161-13

DOI: 10.3969/j.issn.1009-8518.2024.01.014

Study on Machine Learning Cloud Detection Considering Optimal Selection of Samples

ZHANG Hui¹ ZHOU Fangrong² XU Zhen¹ WEN Gang²

MA Yutang² HAN Xu^{3,*} WU Lei³

(1 Yunnan Power Grid Co., Ltd., Kunming 650011, China)

(2 Grid Joint Laboratory of Power Remote Sensing Technology, Electric Power Research Institute, Yunnan Power Grid Company Ltd., China Southern Power, Kunming 650217, China)

(3 Suzhou Deep Blue Space Remote Sensing Technology Co., Ltd., Suzhou 215505, China)

Abstract Aiming at the problem that the traditional threshold algorithm have low accuracy of cloud detection due to spectral differences caused by characteristic differences such as cloud diurnal variation, cloud type, cloud phase state, and cloud optical thickness, This paper proposes a cloud detection algorithm model that takes into account optimal selection of samples, coupled with the physical threshold method and machine learning, and uses the data of Himawari-8 for daytime cloud detection. Through sample optimization selection, the samples include cloud features in different situations as much as possible, providing a good sample basis for

收稿日期: 2023-03-27

基金项目: 云南省重大科技专项(202202AD080010)

引用格式: 张辉, 周仿荣, 徐真, 等. 顾及样本优化选择的机器学习云检测研究[J]. 航天返回与遥感, 2024, 45(1): 161-173.

ZHANG Hui, ZHOU Fangrong, XU Zhen, et al. Study on Machine Learning Cloud Detection Considering Optimal Selection of Samples[J]. Spacecraft Recovery & Remote Sensing, 2024, 45(1): 161-173. (in Chinese)

the machine learning model and increasing the model generalization ability. At the same time, in addition to considering factors such as albedo, brightness temperature, brightness temperature difference, and zenith angle, the input features also add cloud recognition results based on the physical threshold method based on albedo and brightness temperature difference. And cloud detection is carried out based on the Extremely randomized trees (ET) model. The results show that cloud detection cross-validation accuracy of the model is 96.41%, with the total omission error of 2.08% and total commission error of 0.91%, respectively. The results are compared with the product data based on CALIPSO with an overall detection accuracy of 97.1%.

Keywords sample optimization; extremely randomized trees; machine learning; cloud detection; space remote sensing

0 引言

全球表面云覆盖面积约占地球表面积 69%^[1], 对全球辐射平衡和气候变化具有深刻影响。同时, 云覆盖对太阳辐射进行遮挡, 使得卫星传感器难以获取地表信息, 对卫星遥感定量反演工作带来不确定性。因此, 云/云影的有效获取是研究全球辐射平衡、气候变化以及遥感定量反演的重要前提。

卫星遥感观测是研究云检测、云微物理特性等一系列工作的重要手段之一。云在卫星接收的光谱中表现为较高的反照率和较低的辐射亮温, 因此, 对于云的探测, 传统方法多利用有云和晴空下地物在光谱上的差异设置阈值进行云的检测。早期多为单一静态阈值^[2-3], 后来逐步发展为动态自适应阈值^[4]、波段组合阈值^[5-6]、时序阈值^[7-8]等。光谱阈值法虽然计算速度快, 效率高, 并在部分地区取得了不错的结果, 但其对于卫星传感器光谱通道敏感, 且在特定时间和地域获取的阈值应用到其他时间和地域又会产生偏差^[9], 并且阈值的确认也需要做大量的实验, 有效的阈值选择难以把握。

云检测本质上属于分类问题, 机器学习技术因为其较强的信息挖掘能力也被广泛应用于云检测研究中^[10]。机器学习一般分为监督学习和非监督学习, 而云检测研究中监督学习算法更为流行, 例如贝叶斯算法^[11]、支持向量机 (Support Vector Machine, SVM)^[12-13]、随机森林 (Random Forest, RF)^[14-15]和人工神经网络 (Artificial Neural Network, ANN)^[16]等。利用机器学习进行云检测研究通常以反照率、亮温以及通道组合作为输入特征, 以目视解译标记或激光雷达观测结果作为云样本^[17]。利用机器学习进行云检测的输入一般以能表征云时空变化和微物理特征为原则, 因此输入特征应尽可能全面表征云的特性。对于云样本标记, 利用目视解译虽然能够精确获取云和晴空像元, 但通常目视解译获取的样本数据有限; 利用激光雷达, 例如正交偏振云-气溶胶偏振雷达 (CALIOP)^[15]观测结果同样可以较为精确获取云和晴空像元, 但激光雷达卫星空间覆盖有限, 且与目标卫星存在过境时间差异导致时空匹配不一致问题。然而机器学习本质上属于数据驱动的统计模型, 其精度和鲁棒性在很大程度上取决于样本的数量、品质和是否具有代表性等因素^[18]。因此云样本的准确性和代表性是利用机器学习进行云检测的重要影响因素之一^[19]。

日本气象厅发射的新一代地球静止气象卫星 Himawari-8 搭载的高级葵花成像仪 (Advanced Himawari Imager, AHI) 具有 16 个光谱波段, 能够实现对整个圆盘区域每十分钟的观测^[20], 被广泛应用于气象监测、山火监测等方向。Himawari-8 卫星具有更高的光谱分辨率和时间分辨率, 为研究云的光谱特征和时空变化特征提供了良好的基础。其官方团队开发了云检测阈值算法^[21], 并得到云掩膜产品; 同时该团队还开发了云类型 (Cloud Type, CTYPE)、云光学厚度 (Cloud Optical Depth, COD) 和云相态 (Cloud Phase, CLOP) 等产品, 为表征云微物理特性提供了有效参考。

本文针对云日变化、云类型、云相态、云光学厚度等特征差异带来的光谱差异, 导致传统阈值算法无法对云进行有效识别以及一般机器学习云检测对样本和输入特征考虑较少的问题, 以具有高时间分辨率的 Himawari-8 数据为基础, 构建顾及不同天气类型和时刻、云类型、云光学厚度、云相态等要素条件下的云样本, 同时输入特征除了包括反照率、亮温、亮温差以及天顶角等, 还针对机器学习未考虑云物理机理的问题, 引入基于反照率和亮温差的物理阈值方法识别结果作为输入特征。然后在变量重要性度量、变量反向选择和参数调优的基础上选择极限随机树算法进行云检测, 并且与常用的随机森林云检测算法进行对比分析; 为定量评估本文构建的云检测模型的准确度, 通过利用十折交叉验证方法以及云-气溶胶激光雷达与红外探路者卫星 (CALIPSO) 官方云检测产品两方面进行精度评定。

1 研究区域与数据

1.1 研究区域概况

本文研究区域主要分布在云南地区, 云南地处我国西南边陲, 位于东经 $97^{\circ}31' \sim 106^{\circ}11'$, 北纬 $21^{\circ}8' \sim 29^{\circ}15'$, 属于低纬度和高海拔地区, 地势呈西北高、东南低, 为山地高原地形, 气温总体呈北低南高的分布^[22]。受到地域和气候的影响, 云南地区的云层具有明显复杂多变的特点。

1.2 数据

本文使用的数据来自 Himawari-8 卫星和 CALIPSO 卫星。Himawari-8 卫星可实现对全圆盘区域 10 min 每次的高频次观测, 其上搭载的 AHI 传感器光谱通道覆盖从可见光到红外范围的 16 个波段, 其波长范围从 $0.47 \sim 13.3 \mu\text{m}$, 具体波段属性见表 1。除了卫星原始反照率和亮温数据外, 本文使用 Himawari-8 官方产品数据用于样本优化选择, 主要使用参数包括 CTYPE、COD 和 CLOP。

CALIPSO 卫星为主动式激光雷达卫星, 具有正交偏振能力, 可以提供全球云和气溶胶观测数据, 并用于云和气溶胶在调节地球气候中的作用以及两者的相互作用。携带的正交偏振云-气溶胶偏振雷达采用了偏振技术, 是世界上首个应用型的星载云和气溶胶激光雷达, 具有三个通道 (1 064 nm、532 nm 垂直及平行通道), 能够较为准确地识别出云以及反演云的微物理特性。本文主要使用的云检测结果来自 CALIPSO 卫星官方云产品 (2 级 VFM 产品), 数据时间为 2019 年 3 月—2022 年 2 月。

表 1 AHI 波段属性

类别	波段序号	中心波长/ μm	空间分辨率/km
可见光	01	0.47	1.0
	02	0.51	1.0
	03	0.64	0.5
近红外	04	0.86	1.0
	05	1.60	2.0
	06	2.30	2.0
红外	07	3.90	2.0
	08	6.20	2.0
	09	6.90	2.0
	10	7.30	2.0
	11	8.60	2.0
	12	9.60	2.0
	13	10.40	2.0
	14	11.20	2.0
	15	12.40	2.0
	16	13.30	2.0

2 研究方法

本文提出的云检测模型主要包括样本优化选择、多源特征构建、机器学习算法、模型参数调优、精度评定五方面内容, 主要流程如图 1 所示。机器学习模型本质上属于数据驱动模型, 其精度和泛化能力在很大程度上取决于特征的选择以及样本的数等。因此本文重点对多源特征数据集的构建以及云样本品质和是否具有代表性进行探究。

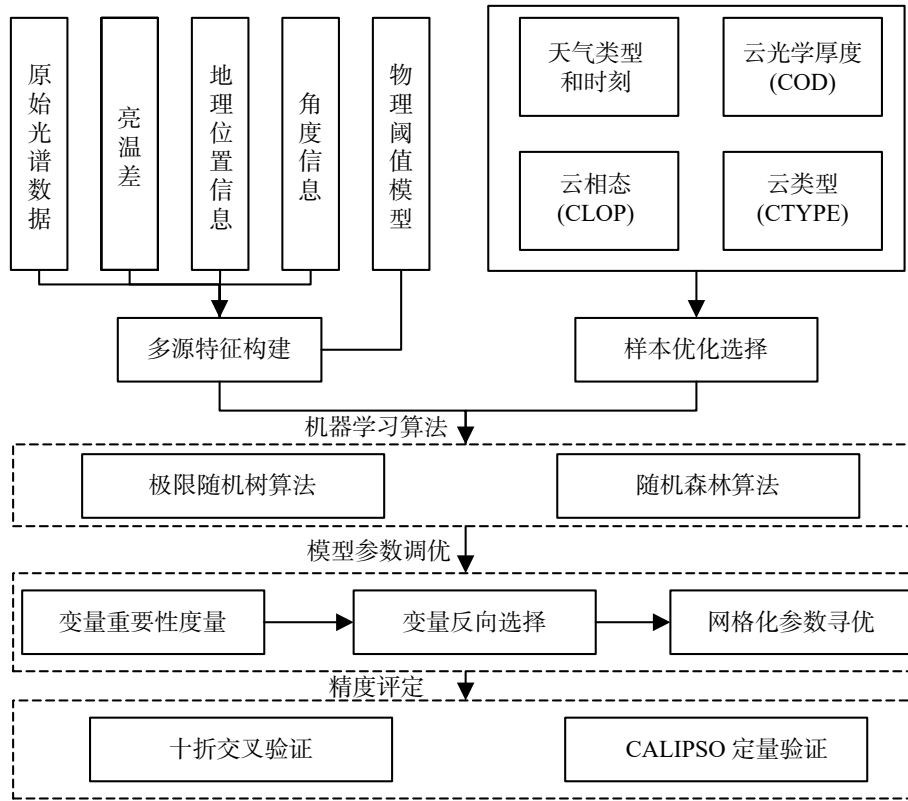


图 1 云检测流程
Fig.1 Flowchart of cloud detection

2.1 样本优化选择

随着机器学习和深度学习技术在卫星遥感领域的不断深入发展和应用，以样本为基础的数据驱动模型逐渐成为遥感信息提取的一种新的研究方向，对样本的规模、品质、多样性等提出了更高要求。本文对云样本品质以及是否具有代表性进行探究，通过优化选择过程，使样本尽可能覆盖不同情形下的云和晴空。

在样本优化选择过程中充分考虑时间维度、天气类型、云相态（CLOP）、云光学厚度（COD）和云类型（CTYPE）。云是大气中的水蒸气遇冷液化成的小水滴或凝华成的小冰晶所混合组成的漂浮在空中的可见聚合物。根据云的定义，云一般按相态可以分为水云、冰云与混合云等，而云相态会直接影响云对辐射的吸收、散射和透射。从图 2（a）和图 2（b）可以看出，云南地区水云相比于冰云和混合云分布更广且离散，如果样本选择过程中不考虑云相态会导致水云样本明显高于冰云和混合云。

COD 是云微物理特性中的重要参数，其表征云的消光能力，一般云量少且云层薄时对应的 COD 值为 2~3 左右。根据云南地区 COD 值小于 2（如图 3（a））和 COD 值小于 3 的结果图（如图 3（b））对比发现，COD 值小于 2 的像元覆盖与目视解译过程中认为的薄云像元更为接近。因此认为 COD 值小于 2 的值为薄云，为样本选择中薄云判断提供依据。

国际卫星云气候学计划（International Satellite Cloud Climatology Project, ISCCP）根据云顶高度和 COD，将云分成 9 类，即卷云、卷层云、深对流、高积云、高层云、雨层云、积云、层积云和层云。根据高度划分，前三类为高云，中间三类为中云，后三类为低云。不同云类型的微物理特性差异导致卫星传感器观测到的反照率和亮温存在差异。从图 4（a）可以看出，在可见光和近红外波段，各类型云在不同波段变化趋势一致，但反照率具有明显差异；不同类型云在同一波段上差异显著，例如积云与雨层云在 3 波段（albedo03）反照率相差 0.26，主要受到不同云类型所处高度、光学厚度、相态、云粒子有效半

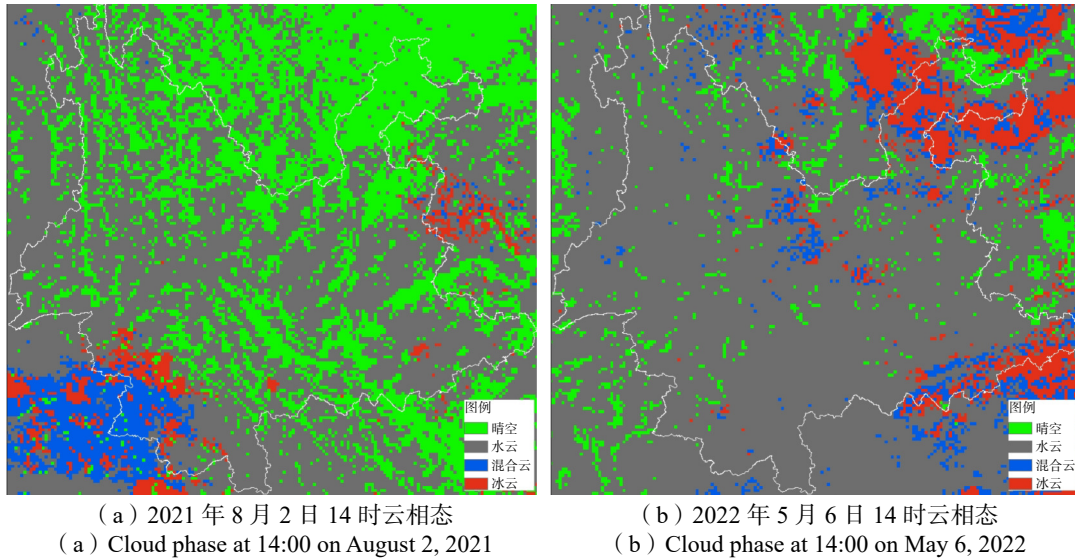


图 2 云南地区云相态空间分布
Fig.2 Spatial distribution of cloud phase in Yunnan

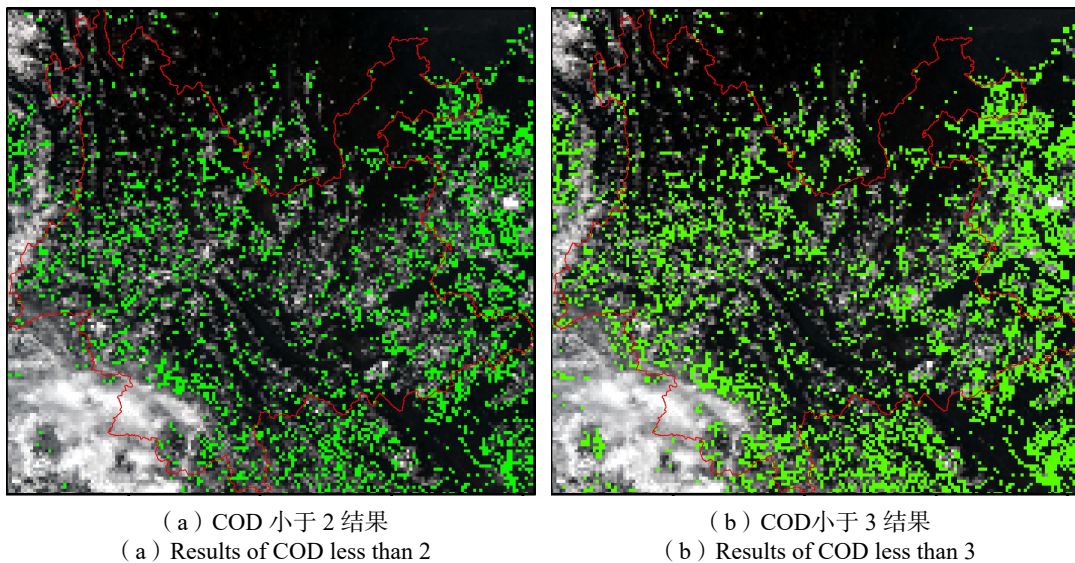


图 3 云南地区 COD 空间分布
Fig.3 Spatial distribution of COD in Yunnan

径等因素影响；根据图 4 (b)，在热红外波段，中低云在 7 波段 (tbb07) 和 10 波段 (tbb10) 亮温差异较小，但在 11 波段 (tbb11)、14 波段 (tbb14)、15 波段 (tbb15) 具有明显差异，主要由于波长较长时，不同类型云对电磁波的吸收和反射特性差异更明显。高云平均亮温为 254.5 K，显著低于中低云平均亮温 276.5 K，主要由于高云相态主要以冰云和混合云为主且高度较高；根据图 4 (c)，各个云类型在 BTD07 (14 和 7 波段亮温差) 和 BTD10 (14 和 10 波段亮温差) 变化趋势基本一致，但不同类型云亮温差差异明显；根据图 4 (d)，各个云类型在 BTD11 (14 和 11 波段亮温差) 和 BTD15 (14 和 15 波段亮温差) 同样具有明显差异，主要由于不同类型云的构成和微物理特性差异导致。

基于以上分析发现，不同云类型的反照率、亮温和亮温差同样存在明显差异，所以样本对云类型的考虑是必要的，样本中标记各个时间段云类型数据，在一定程度上能为以数据为驱动的机器学习模型带来更高的精度和鲁棒性。

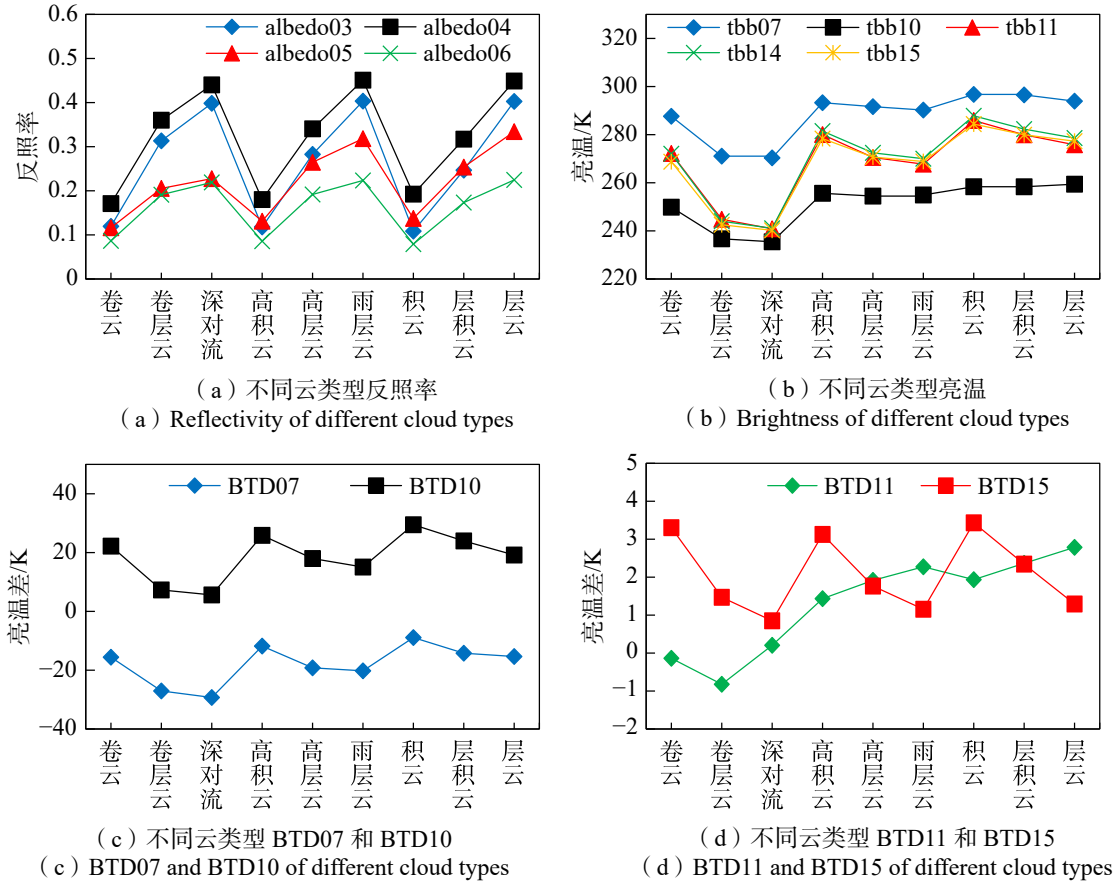


图4 不同云类型反照率、亮温和亮温差差异对比

Fig.4 Comparison of reflectivity, brightness temperature and brightness temperature difference of different cloud types

综上所述，为了提高云识别有效性，云样本需要考虑加入 COD 小于 2 的像元用以标识薄云情况，加入云相态和云类型用以标识不同云相态和类型对辐射的吸收、散射和透射程度。

样本优化选择具体步骤如下：

步骤 1：首先考虑时间维度和天气类型，从 2020 年 4 月—2022 年 5 月范围内选取 Himawari-8 数据，在时间上覆盖 4 个季节、12 个月以及白天中不同时刻，天气类型包括晴天、阴天、多云、雨、雾、下雪后（避免地表积雪对识别带来干扰）等。然后选取 Himawari-8 云产品参数 QA（Quality Assurance）为高置信度且确定为云的像元作为云，并标记为 1；QA 为高置信度且确定为晴空的像元作为晴空，并标记为 0，最终生成样本 L_1 。

步骤 2：基于样本 L_1 ，依据 COD、CLOP 和 CTYPE 进行判断，获取样本 L_2 。样本 L_2 生成过程如下：

$$S_{cod} = \begin{cases} D_1 & d < 2 \\ D_2 & d \geq 2 \end{cases} \quad (1)$$

$$S_{clop} = \begin{cases} P_1 & p=1 \\ P_2 & p=2 \\ P_3 & p=3 \end{cases} \quad (2)$$

$$S_{type} = T_k \quad (k=1,2,3, \dots, 9) \quad (3)$$

式中 S_{cod} 、 S_{clop} 和 S_{type} 分别代表云光学厚度、云相态和云类型样本筛选条件下的样本集合； d 为 COD 值； p 和 k 分别为云相态和云类型掩码； D_1 和 D_2 分别代表 COD 小于 2 和大于等于 2 的数据集； P_1 、 P_2 、 P_3 分别代表水云、混合云和冰云数据集； T_k 为 9 种云类型样本数据集。由于薄云数据相对较小，所

以 D_1 中数据量为 D_2 的 $1/3$, 避免数据量过小。云类型和云相态数据量均以中位数 M 为标准, 小于中位数的数据取全部数据集, 大于中位数数据量随机取 M 个数据。通过对筛选后的三类样本数据取并集, 形成云样本 L_2

$$L_2 = S_{\text{cod}} \cup S_{\text{clou}} \cup S_{\text{type}} \tag{4}$$

步骤 3: 通过对 L_2 随机选取样本点进行目视确认, 删除云和晴空指示不明的像元, 形成最终优化选择后的样本。

通过以上步骤获取的云样本包括不同时间、天气类型、云相态、云类型以及薄云情况下的数据, 增加了云样本代表性。

2.2 多源特征构建

输入特征作为机器学习和深度学习等统计模型中重要因素之一, 输入特征的优选组合是提升模型准确性和鲁棒性的重要措施。输入特征选取原则为是否能够在不同情形区分云和晴空, 与其他研究不同的是本文除了考虑反照率、亮温、亮温差以及天顶角等因素外, 还加入了基于反照率和亮温差异构建的物理阈值方法云检测结果 (Mask), 该模型是对云南地区可见光波段反照率、热红外波段亮温以及中红外与热红外亮温差的多组云测试结果组合, 具体见式 (5) ~ (8):

$$R_3 > 0.2 \tag{5}$$

$$B_7 < -9 \cap T_{14} < 300 \tag{6}$$

$$R_3 > 0.1 \cap B_7 < -5 \cap T_{14} < 290 \tag{7}$$

$$T_{14} < 240 \tag{8}$$

满足式 (5) ~ (8) 任一条件均为云像元, 式中各变量释义见表 2, 各个阈值条件通过对云南地区测试获取。本文输入特征具体见表 2。

表 2 输入特征表
Tab.2 Table of input features

输入特征	中心波长/ μm	解释
albedo03(R_3)	0.64	3 波段反照率
albedo04(R_4)	0.86	4 波段反照率
albedo05(R_5)	1.60	5 波段反照率
albedo06(R_6)	2.30	6 波段反照率
tbb07(T_7)	3.90	7 波段亮温
tbb10(T_{10})	7.30	10 波段亮温
tbb11(T_{11})	8.60	11 波段亮温
tbb14(T_{14})	11.20	14 波段亮温
tbb15(T_{15})	12.40	15 波段亮温
BTD07(B_7)		14 波段与 7 波段亮温差
BTD10(B_{10})		14 波段与 10 波段亮温差
BTD11(B_{11})		14 波段与 11 波段亮温差
BTD15(B_{15})		14 波段与 15 波段亮温差
Lon		经度
Lat		纬度
SAZ		卫星天顶角
SOZ		太阳天顶角
Mask		物理阈值方法云检测结果

2.3 机器学习算法

随机森林由 Leo Breiman^[23] 受到 Amit 和 Geman^[24] 早期工作的启发在 2001 年提出。随机森林由 Bootstrap 样本训练的决策树集合组成，并根据随机选择的预测器子集中的最佳子集划分树中的每个节点。其可以用于分类响应变量（称为分类），也可以用于连续响应（称为回归）。与人工神经网络、支持向量机等方法相比，RF 具有更好的学习性能，对噪声的鲁棒性也更强，同时减少过拟合现象的发生。

极限随机树（Extremely Randomized Trees, ET）方法由 Pierre Geurts 等人于 2006 年提出^[25]。ET 是 RF 在计算效率方面和高度随机化的扩展。ET 根据经典的自上向下过程构建一组未修剪的决策树，类似于 RF 方法。但是，该方法与 RF 有两点主要的区别：

- 1) RF 应用的是引导聚集算法（Bootstrap Aggregating, Bagging），但 ET 不采用自助抽样法（Bootstrap）来选择采样集作为每个决策树的训练集，而是每棵决策树应用的是全部原始训练集；
- 2) RF 在一个随机子集内获取最佳的分裂属性，主要是基于基尼重要度或者均方差的原则，这与传统的决策树保持一致，而 ET 随机选择一个特征值划分决策树，增强了基分类器节点分裂的随机性。

从数据学习的维度理解，ET 进一步增强了样本空间的随机性。

2.4 模型参数调优

通过对所选变量的重要性度量，量化排名所有输入变量在模型的重要性，并在此基础上进行变量反向选择和网格化寻优确定最终参数。

2.4.1 变量重要性度量

变量重要性度量主要采用基于平均不纯度减少的方法（MDI），通过计算每棵树内杂质减少累积的平均值和标准差来实现对特征的重要性度量。如图 5 所示，物理阈值方法云检测结果在模型中的重要性最高，说明物理阈值方法提取的云和晴空结果在一定程度上具有较高可信度，在模型中也表现出较为重要作用。其次为第三波段反照率，而第三波段也经常做云检测特征。卫星天顶角、经度和纬度的重要性分列后三位，可能原因为 Himawari-8 是地球静止轨道卫星，卫星天顶角在云南区域变化不大，因此对云层和晴空的表现特征不明显；经度和纬度分别在一定程度上表征地理位置，但是在云南范围内云和晴空的覆盖情况一般与地理位置相关性不大，因此导致经度和纬度在模型中的重要性不高。

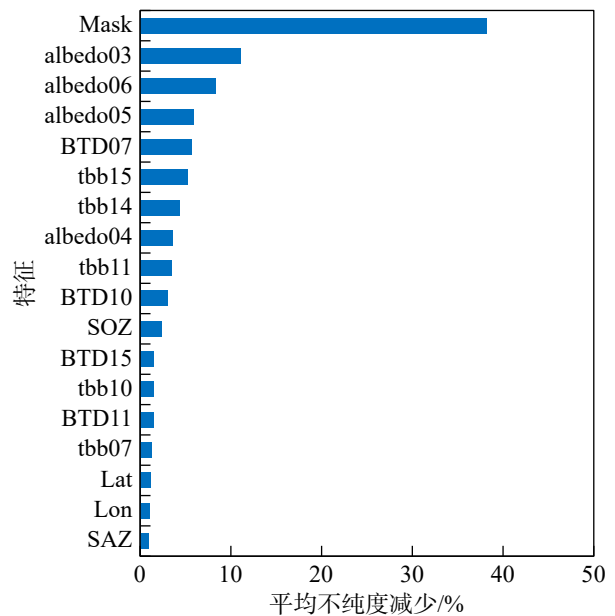


图 5 变量重要性度量

Fig.5 Feature importance

2.4.2 变量反向选择

为了减少模型运行成本和计算量，并且避免数据冗余和相关性，考虑对变量进行反向选择，获取模型最优变量。其基本思想为，通过对变量重要性度量中重要性最差的变量进行剔除，根据模型识别精度进行定量判定，若模型精度不发生明显变化则移除该变量。变量重要性度量结果显示卫星天顶角在变量重要性中居末位，因此在变量反向选择的过程中优先删除卫星天顶角。在变量反向选择过程中（表 3）发现，保留全部变量时，云检测精度（云被正确分类的概率）为 96.41%，总分类精度（云和晴空都被正确分类的概率）为 97.01%，而总漏检率和总虚检率分别为 2.08% 和 0.91%；在分别删除变量重要性后三位的卫星天顶角、经度和纬度后，模型精度均有小幅度下降，因此考虑不删除变量，选择全部变量作为输入数据。

表 3 变量反向选择过程中精度变化

Tab.3 Accuracy changes during variable reverse selection

移除变量	云检测精度/%	总分类精度/%	总漏检率/%	总虚检率/%
无	96.41	97.01	2.08	0.91
SAZ	96.40	97.00	2.08	0.92
Lon	96.40	96.99	2.09	0.92
Lat	96.36	96.96	2.11	0.93

2.4.3 网格化寻优

网格化寻优基本过程为遍历搜索，即在所有候选的参数选择中，通过循环遍历，尝试每一种可能性，表现最好的参数就是最终的结果。网格化寻优过程包括网格搜索和交叉验证。网格搜索，搜索的是参数，即在指定的范围内，按步长依次调整参数，利用调整的参数进行模型训练，从范围内所有参数中找到在验证集上精度最高的参数，本质为模型训练验证并进行比较的过程。

本文对模型最大迭代次数 (n_estimators) 进行网格化寻优，设置最小值为 50，最大值为 1 200，步长为 50，对参数数组进行遍历搜索，并获取对应每个数值对应的得分。以 ET 模型为例，结果 (图 6) 显示，在最大迭代次数在 400~1 200 区间内，对应得分相对接近，在最大迭代次数为 700 时得分最高，并且在大于 700 后基本处于稳定状态，因此 ET 模型选择最大迭代次数为 700 进行模型建立。

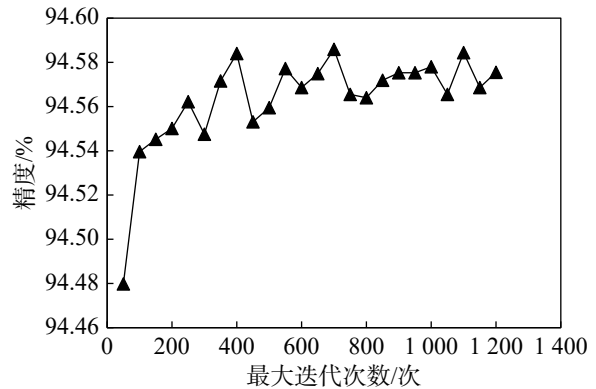


图 6 最大迭代次数网格寻优结果
Fig.6 Grid search results of n_estimators

2.5 精度评定

云检测属于二分类问题，因此一般基于混淆矩阵进行精度评定。TP (True Positive) 表示预测为云且实际也为云的数量；TN (True Negative) 表示预测为晴空且实际也为晴空的数量；FP (False Positive) 表示预测为云但实际为晴空的数量；FN (False Negative) 表示预测为晴空但实际为云的数量。

在分类指标定义后，利用以下四个指标对模型精度进行评定：

$$CP = \frac{TP}{P} \tag{9}$$

$$TP = \frac{TP+TN}{P+N} \tag{10}$$

$$MP = \frac{FN}{P+N} \tag{11}$$

$$FP = \frac{FP}{P+N} \tag{12}$$

式中 P 为样本中为云的样本数量；N 为样本中为晴空的样本数量；CP 为云检测精度，表征云被正确分类的概率；TP 为总分类精度，表征云和晴空都被正确分类的概率；MP 为总漏检率，表征实际为云，而预测为晴空的概率；FP 为总虚检率，表征实际为晴空，预测为云的概率。

3 结果与分析

本文采用十折交叉验证方法对模型进行精度验证，其方法主要将样本数据集分成 10 份，将其中 9 份作为训练数据，1 份作为测试数据，交叉验证重复 10 次，平均 10 次的结果最终得到总体精度。这个方法

的优势在于保证所有样本数据都可以参与验证。

RF 和 ET 云检测精度对比结果 (表 4) 显示, ET 云检测精度和总分类精度均高于 RF。因此选择在验证精度上表现较好的 ET 进行云检测。

表 4 精度指标结果
Tab.4 Accuracy index results

名称	ET 精度/%	RF 精度/%
CP	96.41	96.36
TP	97.01	96.95
MP	2.08	2.11
FP	0.91	0.94

选取 2021 年 8 月 2 日 14 时数据对云检测结果进行验证和分析, 根据图 7 (a) 真彩色图像显示, 该时刻具有云层集中、晴空集中以及云与晴空交叉分布特征, 而根据图 7 (b) 云检测结果 (灰色为云, 浅蓝色为晴空), 整体看, 云与晴空分布与真彩色图像匹配度较好; 在图像左下角云层集中区, 真彩色图像显示存在小范围偏暗区域, 目视识别为薄云, 而云检测能够将这部分识别为云, 且在一定程度上符合云层分布的空间连续性。对于图像右上角晴空集中区, 云检测结果能将晴空像元检测出来, 对于区域内离散分布的碎云也能够与晴空进行分离; 对于图像中部云和晴空交叉分布区域, 云检测结果中的云像元与真彩色图像中的偏亮像元能够较好地进行匹配; 通过视觉分析, 云检测结果对于云边缘和薄云的识别也与真彩色影像具有较好一致性。

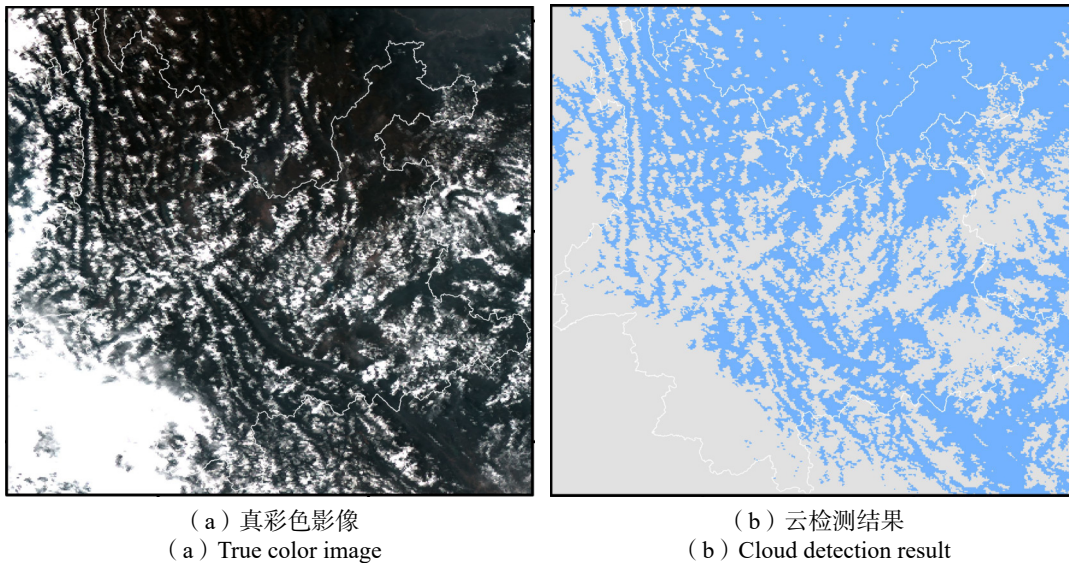
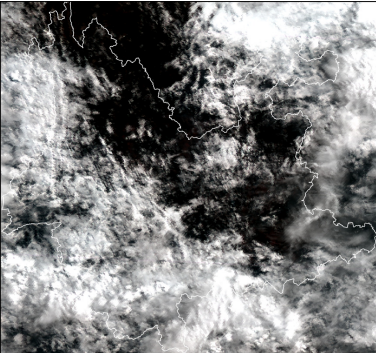
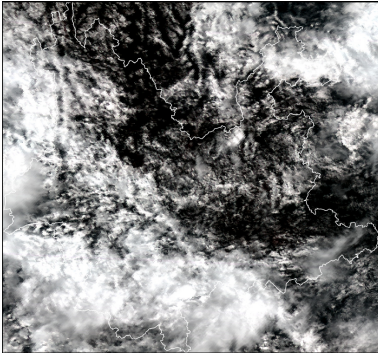
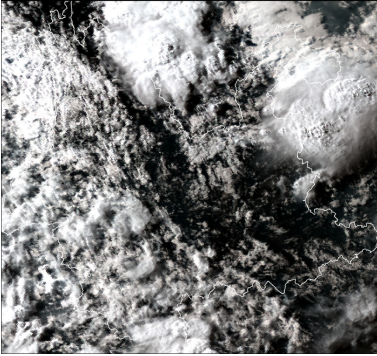
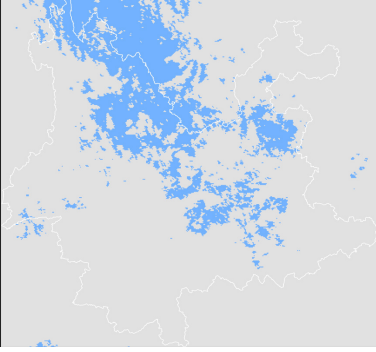
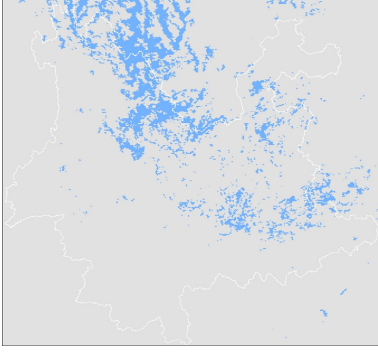
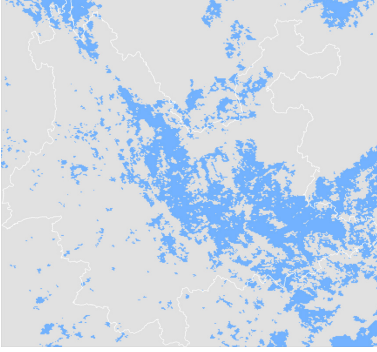


图 7 2021 年 8 月 2 日 14 时真彩色影像与云检测结果对比
Fig.7 Comparison of true color image with cloud detection result at 14:00 on August 2, 2021

为了验证模型在一天早中晚的云检测效果, 选取 2022 年 6 月 2 日早 8 时、中午 12 时和晚 18 时进行对比分析。如表 5 所示, 云南地区云像元占比显著高于晴空像元, 左下和右上区域云层相对集中, 而左上和右下区域云层相对分散。在早 8 时, 晴空像元主要分布在影像左上区域, 中部区域云和晴空交叉分布, 而云检测结果与目视识别判断结果趋势基本一致, 且对区域内相对云和晴空边界分离较好; 根据真彩色影像, 中午 12 时比 8 时晴空像元相对减少, 左上区域晴空被零散的云覆盖, 呈现出波纹状, 而云检测结果与这一趋势相对应, 因此模型可以捕捉一天中不同时刻云层的变化特征。根据真彩色影像, 晚 18 时相比于 8 时和 12 时, 右下区域晴空像元增多, 同时对于区域内较薄的云层也可以做出有效检测。因此, 针对一天中不同时刻的云变化显著的情况下, 模型也可以实现对云和晴空的有效分离。

表 5 不同时刻真彩色影像与云检测对比

Tab.5 Comparison of true color images with cloud detection at different moments in time

	早 8 时	中午 12 时	晚 18 时
真彩色影像			
云检测结果			

为了进一步验证模型的精度和鲁棒性，本文选取在样本集覆盖时间外的 CALIPSO 卫星官方云产品对模型云检测结果进行验证。CALIPSO 数据覆盖四个季节，每个季节不同月份随机抽选 7~10 天数据，确保每个季度的云检测结果都可以得到验证。基于 CALIPSO 过境时刻数据（部分时刻接近，时间误差不超过 5 min）进行云检测，并对所有验证数据以及各个季度数据进行精度评定，其中所有验证数据量为 24 286 个。如图 8 所示，全部数据验证云检测精度为 97.1%，其中夏季云检测精度最高，为 98.77%，秋季云检测精度最低，为 95.38%，说明本文在顾及样本优化选择后构建的云南地区云检测机器学习模型具有较好的精度和鲁棒性，能够对云和晴空实现较好的分离。

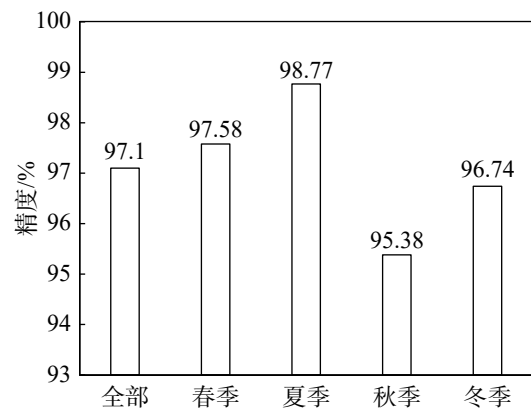


图 8 基于 CALIPSO 的云检测季节验证精度
Fig.8 Season verification accuracy of cloud detection based on CALIPSO

4 结束语

本文在样本优化选择和引入物理阈值方法为输入特征的基础上，构建基于机器学习的云检测模型，对云南地区进行云检测，其中重点考虑样本数据品质和代表性对机器学习模型的重要性，以及改进机器学习模型未考虑云检测的物理机理的情况。与以往研究主要有两点不同：一是在样本优化选择过程中，

考虑时间维度、天气类型、CLOP、COD和CTYPE因素,以此让样本集中包含不同情形下的云特征,增加样本代表性;二是在多源特征构建过程中,引入基于反照率和亮温差异构建的物理阈值方法,使机器学习模型在一定程度上考虑到云检测物理机理过程。

由于卫星空间分辨率和混合像元的影响,导致一个像元可能存在有云和无云同时覆盖情况,对于这种情况,本方法存在漏检和误检情况,难以准确判断,这种情形下需要结合高空间分辨率卫星影像进行研究。此外,本文主要是基于Himawari-8卫星数据进行相关研究,在未来工作中,将基于其他卫星数据利用本方法开展实验,以验证方法的有效性。

参考文献 (References)

- [1] 唐雅慧,周毓荃,蔡森,等.基于CloudSat与CALIPSO联合观测研究全球云分布特征[J].大气科学学报,2020,43(5):917-931.
TANG Yahui, ZHOU Yuquan, CAI Miao, et al. Global Distribution of Clouds Based on CloudSat and CALIPSO Combined Observations[J]. Transactions of Atmospheric Sciences, 2020, 43(5): 917-931. (in Chinese)
- [2] KEGELMEYER J W. Extraction of Cloud Statistics from Whole Sky Imaging Cameras: SNL-CA[R]. Livermore, CA (United States): Sandia National Lab., 1994.
- [3] KRIEBEL K T, SAUNDERS R W, GESELL G. Optical Properties of Clouds Derived from Fully Cloudy AVHRR Pixels[J]. Beiträge zur Physik der Atmosphäre, 1989, 62: 165-171.
- [4] 李俊杰,傅俏燕.“高分七号”卫星遥感影像自动云检测[J].航天返回与遥感,2020,41(2):108-115.
LI Junjie, FU Qiaoyan. Automatic Cloud Detection of GF-7 Satellite Imagery[J]. Spacecraft Recovery & Remote Sensing, 2020, 41(2): 108-115. (in Chinese)
- [5] SHANG H, CHEN L, LETU H, et al. Development of a Daytime Cloud and Haze Detection Algorithm for Himawari-8 Satellite Measurements over Central and Eastern China[J]. Journal of Geophysical Research Atmospheres, 2017, 122(6): 3528-3543.
- [6] ZHU Z, WANG S, WOODCOCK C E. Improvement and Expansion of the Fmask Algorithm: Cloud, Cloud Shadow, and Snow Detection for Landsats 4-7, 8, and Sentinel 2 Images[J]. Remote Sensing of Environment, 2015, 159: 269-277.
- [7] ZHU Z, WOODCOCK C E. Automated Cloud, Cloud Shadow, and Snow Detection in Multitemporal Landsat Data: An Algorithm Designed Specifically for Monitoring Land Cover Change[J]. Remote Sensing of Environment, 2014, 152: 217-234.
- [8] 张永宏,杨晨阳,陶润喆,等.基于FY-4A数据的青藏高原多时相云检测方法[J].遥感技术与应用,2020,35(2):389-398.
ZHANG Yonghong, YANG Chenyang, TAO Runzhe, et al. Multi-Temporal Cloud Detection Method for Qinghai-Tibet Plateau Based on FY-4A Data[J]. Remote Sensing Technology and Application, 2020, 35(2): 389-398. (in Chinese)
- [9] LIU Y, ACKERMAN S A, MADDUX B C, et al. Errors in Cloud Detection over the Arctic Using a Satellite Imager and Implications for Observing Feedback Mechanisms[J]. Journal of Climate, 2010, 23(7): 1894-1907.
- [10] HEIDINGER A K, EVAN A T, FOSTER M J, et al. A Naive Bayesian Cloud Detection Scheme Derived from CALIPSO and Applied within PATMOS-X[J]. Journal of Applied Meteorology and Climatology, 2012, 51(6): 1129-1144.
- [11] DEVASTHALE A, JOHANSSON E, KARLSSON K, et al. Advancing the Uncertainty Characterisation of Cloud Masking in Passive Satellite Imagery: Probabilistic Formulations for NOAA AVHRR Data[J]. Remote Sensing of Environment, 2015, 158: 126-139.
- [12] CHEN G, DONGCHEN E. Support Vector Machines for Cloud Detection over Ice-Snow Areas[J]. Geo-spatial Information Science, 2007, 10(2): 117-120.
- [13] LI P, DONG L, XIAO H, et al. A Cloud Image Detection Method Based on SVM Vector Machine[J]. Neurocomputing, 2015, 169: 34-42.
- [14] FU H, SHEN Y, LIU J, et al. Cloud Detection for FY Meteorology Satellite Based on Ensemble Thresholds and Random Forests Approach[J]. Remote Sensing, 2018, 11(1): 44.
- [15] WANG C, PLATNICK S, MEYER K, et al. A Machine-Learning-Based Cloud Detection and Thermodynamic-Phase Classification Algorithm Using Passive Spectral Observations[J]. Atmospheric Measurement Techniques, 2020, 13(5): 2257-2277.
- [16] CHEN N, LI W, GATEBE C, et al. New Neural Network Cloud Mask Algorithm Based on Radiative Transfer Simulations[J].

- [Remote Sensing of Environment](#), 2018, 219: 62-71.
- [17] LIU C, YANG S, DI D, et al. A Machine Learning-Based Cloud Detection Algorithm for the Himawari-8 Spectral Image[J]. *Advances in Atmospheric Sciences*, 2021, 39: 1994-2007.
- [18] 冯权泷, 陈泊安, 李国庆, 等. 遥感影像样本数据集研究综述[J]. *遥感学报*, 2022, 26(4): 589-605.
FENG Quanlong, CHEN Boan, LI Guoqing, et al. A Review for Sample Datasets of Remote Sensing Imagery[J]. *National Remote Sensing Bulletin*, 2022, 26(4): 589-605. (in Chinese)
- [19] 胡凯, 严昊, 夏旻, 等. 基于迁移学习的卫星云图云分类[J]. *大气科学学报*, 2017, 40(6): 856-863.
HU Kai, YAN Hao, XIA Min, et al. Satellite Imagery Cloud Classification Based on Transfer Learning[J]. *Transactions of Atmospheric Sciences*, 2017, 40(6): 856-863. (in Chinese)
- [20] 范学伟, 郑有飞, 王立稳. 基于Himawari-8气象卫星的东亚夏季冰云云顶特征分布[J]. *气象科学*, 2021, 41(1): 50-59.
FAN Xuewei, ZHENG Youfei, WANG Liwen. The Ice Cloud Top Characteristic Distributions over East Asia in Summer Based on Himawari-8 Meteorological Satellites[J]. *Journal of the Meteorological Sciences*, 2021, 41(1): 50-59. (in Chinese)
- [21] TAKAHITO I, RYO Y. Algorithm Theoretical Basis for Himawari-8 Cloud Mask Product[J]. *Meteorological Satellite Center Technical Note*, 2016, 61: 1-17.
- [22] 姚愚, 李蕊, 郑建萌, 等. 1961—2017年云南季节变化特征分析[J]. *气象科学*, 2020, 40(6): 849-858.
YAO Yu, LI Rui, ZHENG Jianmeng, et al. Studies on the Characteristics of Seasonal Variation of Yunnan Province during 1961-2017[J]. *Journal of the Meteorological Sciences*, 2020, 40(6): 849-858. (in Chinese)
- [23] BREIMAN L. Random Forests[J]. [Machine Learning](#), 2001, 45(1): 5-32.
- [24] AMIT Y, GEMAN D. Shape Quantization and Recognition with Randomized Trees[J]. *Neural Computation*, 1997, 9(7): 1545-1588.
- [25] GEURTS P, ERNST D, WEHENKEL L. Extremely Randomized Trees[J]. [Machine Learning](#), 2006, 63(1): 3-42.

作者简介

张辉, 男, 1987年生, 2010年获贵州大学电气工程专业学士学位, 工程师。研究方向为输电线路检修试验与运行管理。
Email: 182174124@qq.com。

通讯作者

韩旭, 男, 1993年生, 2019年获中国矿业大学摄影测量与遥感专业硕士学位, 工程师。研究方向为云遥感与定量遥感反演。Email: 228239634@qq.com。

(编辑: 庞冰)