

实现了早期识别还是模型过拟合？基于多光谱无人机图像对欧洲云杉小蠹虫害林木进行早期识别并对比机器学习与植被指数模型的表现

Langning Huo, Eva Lindberg, Jonas Bohlin, Henrik Jan Persson

瑞典农业大学, 瑞典

随着全球气候变暖的加剧, 森林害虫的爆发越来越频繁, 对森林造成了巨大的破坏。在早期阶段发现和清除森林中的树干蛀虫对于避免虫害扩散来说十分重要。利用遥感技术可以有效地探测和绘制森林死亡区域, 但在虫害早期林木光谱变化较小, 因此实现早期检测仍然具有挑战性。

本研究利用多光谱无人机图像评估了欧洲云杉小蠹虫(*Ips typographus*, L.)早期染病林木的可探测性。研究的科学问题和目标是(1)检验林木在受害前是否因活力降低表现出光谱异常, (2)量化林木在受害早期阶段可探测性, (3)对比使用单一植被指数(VI)与使用机器学习模型在早期识别上的精度区别, 并对比模型在未训练区域数据上的表现。(4)测试我们在之前的研究中提出的MR_DSWI2指数的表现。

该研究使用了四个时段的多光谱无人机图像, 覆盖了瑞典南部6个小班的24个样地。四个时段分别为2021年5月(受害前)、6月(染病早期)、8月(染病早期和中期)和10月(染病中期和晚期)。研究每周对997棵云杉进行了健康状况清查, 其中有208棵受害, 记录了这些林木染病开始的周数。研究采用标记控制分水岭分割方法对单株树冠进行分割, 计算树冠平均反射率及10个植被指数, 并将林木依据侵染时长进行分组, 分别计算每组染病林木的可探测性。研究分别使用了随机森林分类(RF)和线性判别分析(LDA), 各自设置三组自变量, 包括(1)所有波段, (2)所有计算的植被指数, (3)MR_DSWI2使用的四个波段。研究还利用MR_DSWI2、NDRE2和NGRDI指数分别建立了3个LDA模型。为了检验潜在的过拟合以及检验模型在未训练区域上的适用性, 研究使用了两种设置训练和测试数据集的方法并进行了精度比较。方法A使用5个小班的单木数据对模型进行训练, 使用剩余小班的单木进行测试。这种方法测试了模型在未训练区域的性能, 过度拟合的模型会产生较低的精度。方法B不区分小班, 随机分配90%的树木进行训练, 10%的树木进行测试。方法B通常用于分离训练和测试数据集, 但精度无法提现模型在未训练区域的表现。如模型出现过拟合, 则方法A和方法B的验证精度将出现差异。研究精度用Kappa系数表示。

研究结果表明:(1)当使用维度更多、复杂度更高的模型(即RF模型)时, 训练区域的验证精度较高, 而未训练区域的验证精度较低, 表明这些模型的过拟合更为突出。方法B, 即使用与训练数据相同区域的数据对模型进行验证, 无法表明模型在未训练区域的适用性。因此, 我们建议使用方法A来测试模型的性能, 并对结果进行了进一步的讨论。(2)基于在未训练区域进行测试时的结果(方法A), 无论是简单线性模型还是随机森林模型, 都无法将健康林木与即将染病的林木进行成功分类, 因此我们认为染病林木在染病前没有表现出可探测的易染病特性。(3)基于在未训练区域进行测试时的结果(方法A), 没有模型可以将健康林木与染病不超过5周的林木进行成功区分, 意味着林木在染病的前5周期间是几乎无法通过多光谱无人机进行识别的。林木染病的可探测率在染病10-13周期间较高, 当使用最佳分类模型(MR_DWSI2 LDA)时, kappa系数中位数为0.67-0.75。染病19-22周时可探测性进一步升高。(4)在所有分类模型中, 基于MR_DWSI2指数的线性判别模型在不同染病阶段的准确率最高且最稳定, 其次是使用4波段的线性判别模型和使用NGRDI指数的线性判别模型。