

基于随机森林模型的小麦白粉病遥感监测方法

Remote sensing monitoring of wheat powdery mildew based on random forest model

李健丽^{1,2} 董莹莹² 师越² 朱溢佺¹ 黄文江^{2*}

(1. 首都师范大学数学科学学院, 检测成像北京高校工程研究中心, 北京 100048;

2. 中国科学院遥感与数字地球研究所, 数字地球重点实验室, 北京 100094)

Li Jianli^{1,2} Dong Yingying² Shi Yue² Zhu Yining¹ Huang Wenjiang^{2*}

(1. Beijing Higher Institution Engineering Research Center of Testing and Imaging, School of Mathematical Sciences, Capital Normal University, Beijing 100048, China; 2. Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China)

小麦白粉病是危害我国小麦生产的主要病害, 及时高精度地监测其发生发展状况对有效制定病害防控策略十分重要。基于目测手查的传统病害调查方式, 耗时费力且无法实时监测。随着空间遥感技术的发展, 利用遥感开展病害监测已成为及时获取作物病害信息的重要方式。目前作物病害监测方法大致分为统计模型和物理模型, 前者灵活易用、操作简单; 后者能定量分析病害发生情况, 但输入变量较多且结构复杂。统计模型在作物病害方面应用广泛, 其中支持向量机(support vector machine, SVM)模型在小样本、非线性及高维模式识别中表现出特有的优势, 分类能力突出(聂臣巍等, 2016), 但仍需要调整很多参数且对大规模训练样本耗时; 而随机森林(random forest, RF)模型善于处理大规模的输入变量且需调整参数较少, 在遥感影像分类、叶面积反演等领域应用中具有良好的分类性能(李欣海, 2013), 但在作物病害测报方面的应用极少。本试验引入随机森林模型监测小麦白粉病, 以期提高大数据下的监测精度, 为有效防控该病害提供基础依据。

1 材料与方法

1.1 材料

小麦白粉病野外实测调查数据: 于2014年5月下旬在河北省石家庄市的藁城市、晋州市、赵县开展野外调查, 共获取87个采样点的GPS坐标、小麦白粉病病情等级实测调查数据。

遥感影像及气象数据: 2014年5月6日、5月22日的Landsat TM8影像数据自地理空间数据云主页(<http://www.gscloud.cn/>)下载, 空间分辨率为30 m;

2014年5月17日至22日逐日气象站点温、湿度数据来源于世界243个国家的天气主页(<http://rp5.ru/>)。

1.2 方法

RF模型构建: RF模型利用bootstrap重采样方法从原始样本中抽取多个样本, 对每个bootstrap样本进行决策树建模, 组合成多棵决策树进行预测, 并通过多数投票法决定最终分类结果的联合预测模型(Breiman, 2001)。模型建立主要包括核函数的创建、决策树数量(ntree)和内部节点随机选择属性个数(mtry)的选取。核函数的创建主要基于R语言Random Forest 4.6-6软件包RF函数。通过对不同样本子集的多次分析和仿真试验比较发现, ntree在超过200时, 误差趋于稳定。因此建模时ntree选择200, mtry参数取默认值。训练数据集中特征的个数定为14。

SVM模型构建: SVM的基本思想是, 在线性可分情况下, 在原空间寻找2类样本的最优分类超平面; 在线性不可分情况下, 加入松弛变量, 通过使用非线性映射将低维输入空间的样本映射到高维属性空间使其变为线性情况, 并在该特征空间中寻找最优分类超平面。模型的建立主要是核函数的创建, 即基于R语言e1071软件包svm函数中的径向基函数。

数据建模及模型评价标准: 采用研究区2014年5月下旬遥感影像反演所得数据以及空间气象数据, 首先使用ENVI 4.8软件对遥感数据进行辐射定标、大气校正和几何校正等预处理, 然后提取种植面积、植被指数等。地表温度采用大气校正法反演, 站点气象数据使用ArcMap 9.3软件进行空间插值处理。本试验选用归一化植被指数、归一化水分植被

指数、差值植被指数、地表温度、大气温度及大气湿度等特征变量作为小麦白粉病监测模型的输入变量,通过独立样本 t 测验法筛选出与小麦白粉病的发病情况在0.05水平显著相关的14个特征变量作为最终的输入变量,以小麦白粉病发病与健康为等级作为模型的输出变量。采用均方根误差RMSE、决定系数 R^2 、Pearson相关系数对建模结果进行对比分析,并以错、漏分误差、总体精度及Kappa系数评价对小麦白粉病的监测结果。

2 结果与分析

2.1 研究区域小麦白粉病发病情况

2014年5月下旬小麦灌浆期的白粉病发病空间分布图显示,赵县发病面积较大,晋州市次之,藁城市最少。RF模型和SVM模型监测的发病面积在赵县与藁城市基本一致,在赵县分别占当地小麦种植面积的

93.85%和96.29%,在藁城市分别占14.21%和12.83%,而在晋州市监测结果差异较大,RF模型显示发病区域主要分布在晋州东北部及中南部,占该市小麦种植面积的42.14%,SVM模型显示发病区域主要分布在晋州市南部,占该市小麦种植面积的15.04%(图1)。

2.2 监测模型的拟合优度及精度分析

拟合优度评价结果表明,RF模型的Pearson相关系数、 R^2 及RMSE的监测结果均优于SVM模型。RF模型的总体精度为85.71%,Kappa系数为71.49%,均大于SVM模型的71.42%和42.73%。且RF模型的漏分误差为10.00%和18.18%,小于SVM模型的30.00%和27.27%;错分误差为18.18%和10.00%,小于SVM模型的30.00%和27.27%。表明RF模型对小麦白粉病病情的监测结果优于SVM模型,在小麦白粉病病情监测方面具有较高的可行性。

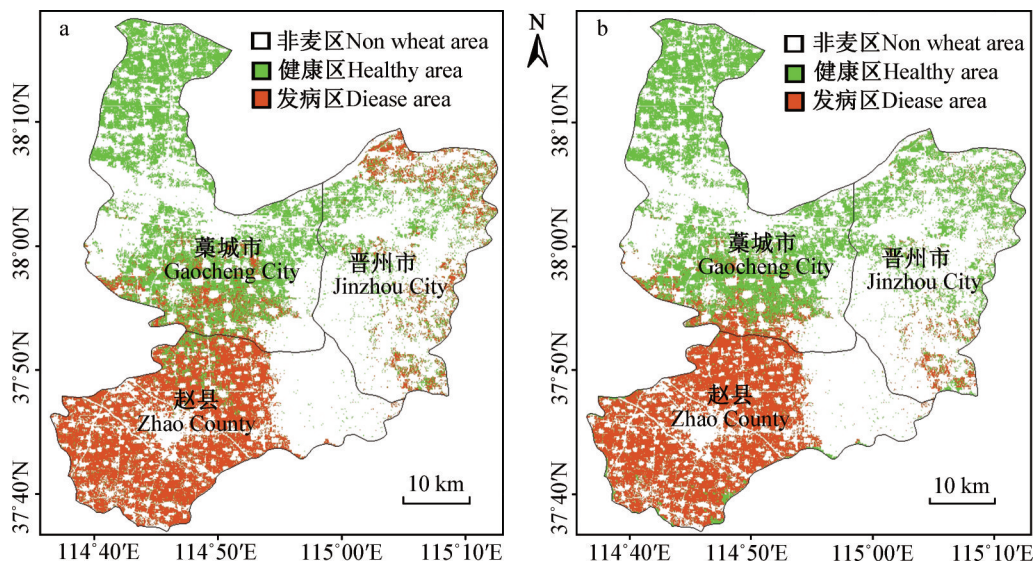


图1 小麦白粉病RF模型(a)和SVM模型(b)的监测结果

Fig. 1 Monitoring results of wheat powdery mildew with random forest (a) and support vector machine (b) models

3 讨论

RF模型与SVM模型对小麦白粉病的监测结果在晋州市东北部存在差异,分析原因可能是晋州市野外调查数据集中,存在数据噪声与不平衡数据状态,SVM模型精度受到一定影响,而RF模型对于缺失数据和非平衡数据有较强的鲁棒性,因而监测结果显示RF模型的精度优于SVM模型。从2种模型的数学原理出发,RF方法是集成学习算法,有助于将多个弱学习器组合形成强学习器,善于处理大数据样本,且行、列随机采样过程的引入使其具有良好的噪声容忍性,能有效避免过拟合的发生;而SVM方法在小样本数据集上具有较明显的优势,对于大数据量而言,SVM模型拟合结果并不理想。

参考文献 (References)

- Breiman L. 2001. Random forests. *Machine Learning*, 45(1): 5–32
- Li XH. 2013. Using “random forest” for classification and regression. *Chinese Journal of Applied Entomology*, 50(4): 1190–1197 (in Chinese) [李欣海. 2013. 随机森林模型在分类与回归分析中的应用. *应用昆虫学报*, 50(4): 1190–1197]
- Nie CW, Yuan L, Wang BT, Jin XL, Huang WJ, Zhang JC, Yang GJ. 2016. Monitoring wheat powdery mildew based on integrated remote sensing and meteorological information. *Acta Phytopathologica Sinica*, 46(2): 285–288 (in Chinese) [聂臣巍, 袁琳, 王保通, 金秀良, 黄文江, 张竞成, 杨贵军. 2016. 综合遥感与气象信息的小麦白粉病监测方法. *植物病理学报*, 46(2): 285–288]

(责任编辑: 李美娟)