农业资源与环境学报 中文核心期刊 中国科技核心期刊
JOURNAL OF AGRICULTURAL RESOURCES AND ENVIRONMENT
欢迎投稿 http://www.aed.org.cn

基于决策树模型的耕地地力与玉米丝黑穗病发生关系研究

陈丽，崔运鹏，王末，牛永春，徐爱国，刘珂艺，刘娟，侯颖

引用本文:
陈丽, 崔运鹏, 王末, 等. 基于决策树模型的耕地地力与玉米丝黑穗病发生关系研究[J]. 农业资源与环境学报, 2021, 38(5): 928-936.

在线阅读 View online: https://doi.org/10.13254/j.jare.2021.0193

陈丽, 崔运鹏, 王末, 等. 基于决策树模型的耕地地力与玉米丝黑穗病发生关系研究[J]. 农业资源与环境学报, 2021, 38(5): 928-936.

CHEN L, CUI Y P, WANG M, et al. A study of relationships between cultivated land fertility and corn head smut using a decision tree model [J]. *Journal of Agricultural Resources and Environment*, 2021, 38(5): 928-936.

开放科学 OSID

基于决策树模型的耕地地力与玉米丝黑穗病发生关系研究

陈丽[1,2], 崔运鹏[1,2*], 王末[1,2], 牛永春[3], 徐爱国[3], 刘珂艺[1,2], 刘娟[1,2], 侯颖[1,2]

(1.中国农业科学院农业信息研究所, 北京 100081; 2.农业农村部农业大数据重点实验室, 北京 100081; 3.中国农业科学院农业资源与农业区划研究所, 北京 100081)

摘　要:为了探究耕地地力与玉米丝黑穗病发生关系,本研究以515个主要玉米种植县域为研究区域,选取有机质(Organic matter,OM)、全氮(Total nitrogen,TN)、全磷(Total phosphorus,TP)、全钾(Total potassium,TK)、有效磷(Available phosphorus,AP)、速效钾(Available potassium,AK)和pH 7种耕地地力因子,分别利用分类与回归树(Classification and regression tree,CART)模型、随机森林(Random forest,RF)模型和极端随机树(Extremely randomized trees,ERT)模型构建了玉米丝黑穗病发生程度与耕地地力因子关系模型,并进行了3种模型效果比较。结果表明:RF和ERT模型总分类性能明显优于CART模型,3个模型在病害发生程度1级(GⅠ)上的查准率(Precision,Pr)、查全率(Recall,Re)、F1 score(F1)值均较高,分类效果比病害发生程度2级(GⅡ)要好,但考虑到准确监测病害高发情况、减少高发病情况在分类预测中漏分机率对开展病害防治的重要性,确定ERT模型为最佳优选分类器。耕地地力特征变量与病害发生程度重要性分析表明,玉米丝黑穗病发生程度与耕地地力因子AP、TK、pH和TP具有一定的相关性。研究结果为深入探究耕地地力对玉米丝黑穗病影响机理提供了线索和支撑。

关键词:玉米丝黑穗病;耕地地力;分类与回归树模型;随机森林模型;极端随机树模型

中图分类号:S435.131.42　　**文献标志码:**A　　**文章编号:**2095-6819(2021)05-0928-09　　**doi**: 10.13254/j.jare.2021.0193

A study of relationships between cultivated land fertility and corn head smut using a decision tree model

CHEN Li[1,2], CUI Yunpeng[1,2*], WANG Mo[1,2], NIU Yongchun[3], XU Aiguo[3], LIU Keyi[1,2], LIU Juan[1,2], HOU Ying[1,2]

(1. Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China; 2. Key Laboratory of Big Agri-data, Ministry of Agriculture and Rural Affairs, Beijing 100081, China; 3. Institute of Agricultural Resources and Regional Plannig, Chinese Academy of Agricultural Sciences, Beijing 100081, China)

Abstract: To explore the relationship between cultivated land fertility and corn head smut occurrence, 515 main corn growing counties were selected, and the relationship between corn head smut incidence and cultivated land fertility was modeled using classification and regression tree(CART), random forest(RF) and extreme randomized trees(ERT) models, with organic matter(OM), total nitrogen(TN), total phosphorus(TP), total potassium(TK), available phosphorus(AP), available potassium(AK), and PH as characteristic variables. The classification accuracy of the different models were compared. The comprehensive classification performance of RF and ERT was obviously better than that of CART. The three models had higher Precision(Pr), Recall(Re), and F1 score(F1) values on disease grade 1(GⅠ), and the classification effect was better than that of disease grade 2(GⅡ). Considering the importance of accurately monitoring the high incidence of diseases and reducing leakage rates in classification and prediction for disease prevention and control, the ERT model, with its higher Re value on GⅡ was finally determined to be the best classifier. Additionally, importance analysis between the variables

收稿日期:2021-04-02　　**录用日期:**2021-06-22

作者简介:陈丽(1985—),女,助理研究员,山东聊城人,从事农业大数据管理、挖掘分析研究。E-mail:chenli02@caas.cn

*__通信作者:__崔运鹏　E-mail:cuiyunpeng@caas.cn

基金项目:国家重点研发计划项目(2017YFD0200607);中国农业科学院创新工程项目(2021CX013)

Project supported: The National Key Research and Development Program of China (2017YFD0200607); The Agricultural Science and Technology Innovation Program of CAAS(2021CX013)

characteristic of cultivated land fertility and degree of disease occurrence showed that the incidence of corn head smut correlated appreciably with AP, TK, pH, and TP. The results provide clues and support for further mechanistic research into effects of cultivated land fertility factors on corn head smut.

Keywords: corn head smut; cultivated land fertility; classification and regression tree model; random forest model; extremely randomized trees model

玉米是我国重要的粮食和饲料作物,2019年播种面积4 127万 hm^2(6.19亿亩),产量2.6亿t,仅次于美国,同时玉米在医药、化工等领域也具有广泛用途[1]。近年来,我国玉米生产区,尤其是东北、华北和黄淮主产区普遍采用多年连作、高肥、密植、免耕、秸秆还田等耕作措施,导致生产环境生态失衡、土壤环境恶化,玉米土传病害连年普发[2]。玉米丝黑穗病是由丝轴黑粉菌(*Sphacelotheca reiliana*)引起的一种最常见土传病害,在我国部分地区(尤其北方)危害严重,每年因该疾病造成的玉米产量损失可达30万t[3]。

土传病害的病原菌存活于土壤中或者土壤表面,并借土壤环境进行生长繁殖[4-5],土壤非生物因子的变化可对病原菌生存与致病能力产生较大影响[6]。已有研究表明,土壤pH可以通过引起土壤微生物群落变化,间接地影响青枯菌在土壤中的存活[7];土壤有效磷和速效钾会对盛花期油菜叶部菌核病发病产生较大影响[8];此外,土壤质地以及养分元素的种类、含量、形态也会影响病原菌的存活[9]。土传病原菌种类繁多,其对不同寄主和环境因子的侵害能力和程度不同。玉米丝黑穗病的发生除受耕作措施影响外,还与土壤的温度、湿度有着密切关系,但土壤地力条件对病菌积累和病害发生的影响却鲜有研究和报道。

当前对作物病害发生、发展影响因素的研究常用方法包括Pearson相关性分析、Spearman相关性分析和逐步回归分析等方法。针对生物数据多维性、不稳定性、变量间相互影响等特点,通过建立线性相关关系进行病害主要影响因子筛选和二者间相关性分析,对于处理多维度、非线性的数据,往往不能达到对变量间复杂相互关系进行深入探究的目的[10]。决策树及其优化、改进算法(本研究统称决策树算法)是数据挖掘与知识发现领域中基于逻辑的经典的分类预测方法之一,能够较好地处理变量间的非线性关系,对噪声信息和数据缺失具有较好的稳健性和鲁棒性。此外,决策树属于白箱模型,整个计算过程可见,树状结构简单直观,非常便于用户理解和解释,在农作物病情监测预警、病害诊断识别以及种植面积提取等方面都表现出较好的应用效果[11-12]。综上,本研究以耕地地力因子为自变量,玉米丝黑穗病发生程度为因变量,分别采用分类与回归树模型(Classification and regression tree,CART)、随机森林模型(Random forest,RF)和极端随机树模型(Extremely randomized trees,ERT)构建病害发生与耕地地力关系模型,从宏观尺度对玉米丝黑穗病发生程度的地力影响因子进行判断和预测分析,以期为玉米丝黑穗病的科学防治提供依据。

1 材料与方法

1.1 数据来源

玉米是一种喜温作物,对土壤条件要求不十分严格,主要种植于我国东北、华北和西南地区。广泛的玉米种植以及我国复杂多样的土壤类型分布,造成玉米生产耕地地力条件区域差异明显。因此,基于耕地地力空间异质性和数据的可获得性,本研究以县域为研究单元,选取了515个主要玉米种植县作为研究区域(图1)。研究所用的数据主要包括县域玉米丝黑穗病病情数据和耕地地力数据。其中,玉米丝黑穗病病情数据来源于农业农村部全国农业技术推广服务中心和2008—2014年全国农业植保专业统计资料。耕地地力数据主要有3个来源:有机质(Organic matter,OM)、全氮(Total nitrogen,TN)、有效磷(Available phosphorus,AP)、速效钾(Available potassium,AK)和pH来自全国农业技术推广服务中心编著的《测土配方施肥土壤基础养分数据集(2005—2014)》;全磷(Total phosphorus,TP)和全钾(Total potassium,TK)来自科技基础性工作项目“1∶5万土壤图籍编撰及高精度数字土壤构建(二期工程)”的土壤剖面数据库;对于无法通过上述途径获取完整数据的县域,通过国家科技基础条件平台——国家地球系统科学数据中心(http://www.geodata.cn)全国第二次土壤普查工作成果进行数据补充。本研究对剖面分层数据经计算0~20 cm加权平均值后,按算术平均值计算县域平均值。

1.2 数据预处理

1.2.1 缺省数据处理

CART、RF和ERT支持数据缺失情况下的运算,但

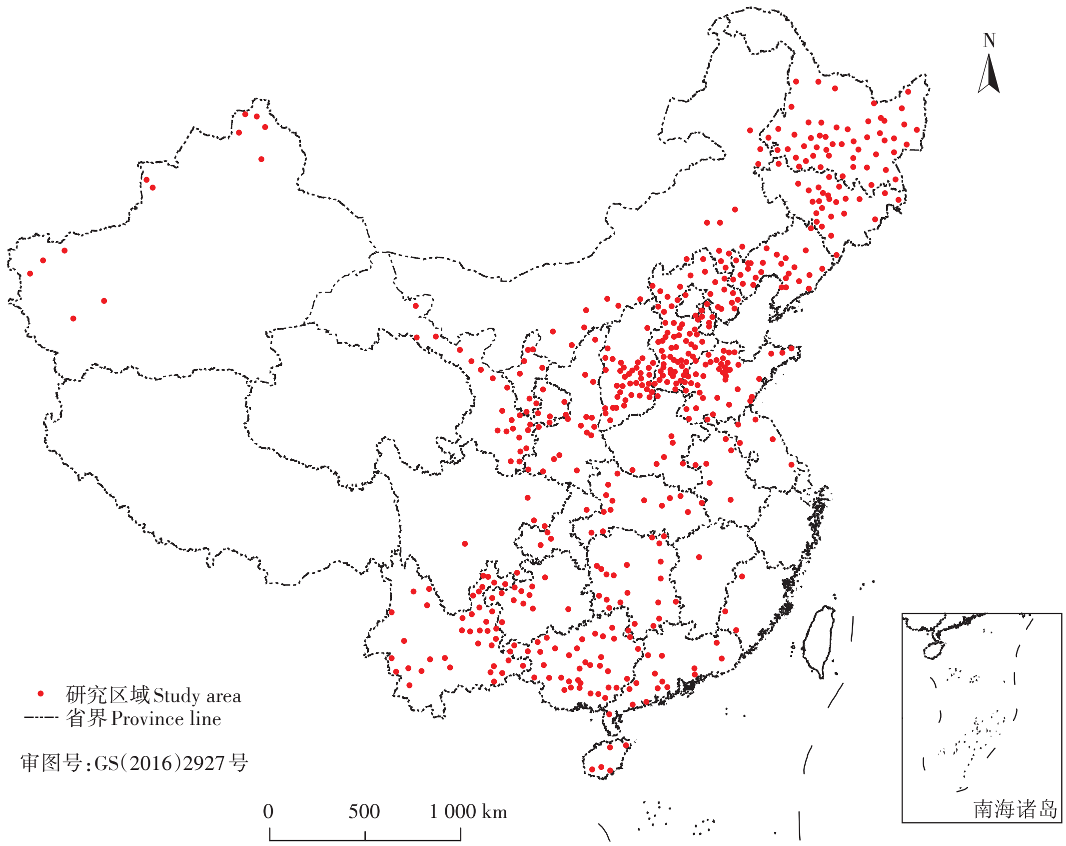


图1 研究区域空间分布

Figure 1 Distribution map of the study area

考虑到本研究区域空间不连续性和土壤地力因子的空间异质性,缺省值数据处理采取直接删除的方式。

1.2.2 不平衡数据修正

数据不平衡问题广泛存在,不平衡性是指数据中目标变量分布不均衡。使用决策树等传统的机器学习方法进行分类,往往更偏向于多数类,从而对少数类的分类精度造成负面影响[13]。可以通过基于数据的修正和基于算法的修正两种方式进行处理,常用方法有过采样法(Oversampling)、欠采样法(Undersampling)和阈值移动法(Threshold-moving)等[14]。

玉米丝黑穗病的发生具有典型的时空异质性,我国一般年份玉米丝黑穗病的田间发病率不超过10%[15],黑龙江主栽品种低于5%。因此,针对研究数据集不平衡性问题,本研究采用过采样法的代表性算法SMOTE(Synthetic minority oversampling technique)对训练集中少数类样本进行插值,生成人工样本,实现数据集均衡的目的。SMOTE算法的步骤:①对少数类训练样本集中的每个样本x_j计算其与其他样本之间的欧式距离,找到其K近邻样本,其中j=1,2,…,N_j,N_j表示少数类训练样本的数目;②从K近邻样本中随机选择一个样本$\hat{x}_j$,计算其与训练样本x_j之间的特征差值$d_j=\hat{x}_j-x_j$;③乘以一个随机数δ,$\delta\in[0,1]$;④生成人工少数类样本$x_{new}=x_j+\delta\times d_j$;⑤重复操作步骤①~④,直到训练样本集达到平衡。

1.3 病害发生程度分级与特征变量选取

1.3.1 病害发生程度分级

玉米丝黑穗病在不同年份、不同地域的病害发生程度不同,多处于2%~8%之间。根据我国农业行业标准《玉米抗病虫性鉴定技术规范 第3部分:玉米抗丝黑穗病鉴定技术规范》(NY/T 1248.3—2006)[16]中田间玉米发病程度分级标准,得出我国玉米主产地区丝黑穗病平均发生情况,结果显示不同病害发生程度分级下区县数目差异较大,且大部分区县发病率在5%以下。考虑到数据的不均衡性,级别划分过多可能会带来预测准确性的降低,同时,为了减少模型在高发病类别预测中漏分机率,提高风险防范能力,因此,采用一种空间分级算法——自然断裂点法(Jenks)对模型构建的病害发生程度进行分级。分级结果为:病害发生程度1级(GⅠ),发病率≤2.5%;病害发生程度2级(GⅡ),发病率>2.5%。2.5%的发病率处于偏轻向

中等病情的过渡阶段，也较符合提前预判预防的实际需求。

1.3.2 特征变量选取

《耕地地力调查与质量评价技术规程》（NY/T 1634—2008）中耕地地力因子共分为6类60小项[17]，分别是气象、立地条件、剖面性状、土壤理化性状、障碍因素和土壤管理，其中与土传病害发生关系密切的主要为土壤理化性状和土壤管理，通过文献调研、咨询玉米病害专家意见，并结合数据的可获取性，本研究选取OM、TN、TP、TK、AP、AK和pH值7个因子作为特征变量。为了得到更好的模型训练效果，本研究在建模前对特征变量进行了Z-score归一化处理。

1.4 模型构建

1.4.1 CART模型

CART是以Gini系数的减少量为测度指标，选取使Gini系数减少量最大的特征变量和分割属性阈值作为切分点，对训练样本数据集进行分组构造的二叉决策树模型，并通过剪枝降低复杂度、避免过拟合、提高树的可解释性[18]。本研究的剪枝策略为预剪枝，最大树深度设置为5。

CART模型最优特征选择：

Gini系数表示特征变量的不确定度，其定义为：

$$\mathrm{Gini}(S)=1-\sum_{i=1}^{m}P^2(C_i) \tag{1}$$

式中：S表示训练样本数据集，其包含样本数为N；C_i为样本分成的不同类别；m为样本分成的类别数；$i\in\{1,2,3,\cdots,m\}$；$P(C_i)$为样本归为第i类的概率，$P(C_i)=S_i/N$，S_i为归为第i类样本类别C_i的训练样本个数，且$N=\sum_{i=1}^{m}S_i$。

如果训练样本集S根据特征A是否取某一可能值a被分割成S_1和S_2两个子样本集，则在特征A的条件下，集合S的Gini系数为：

$$\mathrm{Gini}(S,A)=(\frac{N_1}{N})\times\mathrm{Gini}(S_1)+(\frac{N_2}{N})\times\mathrm{Gini}(S_2) \tag{2}$$

式中：N_1、N_2分别表示S_1和S_2两个子样本集包含的样本个数。

Gini系数的减少量为：

$$\Delta\mathrm{Gini}(S)=\mathrm{Gini}(S)-\mathrm{Gini}(S,A) \tag{3}$$

1.4.2 RF模型

RF是一种多棵树集成分类器，它以CART为元分类器，通过集成学习的思想将多棵树集成在一起，增强分类能力，提高分类精度。RF通过自助法（Bootstrap）构建多个相对独立的决策树，每棵决策树训练集的获得采用有放回的随机抽样，抽样生成的袋外数据可以进行预测分类正确率的评估以及模型泛化能力的估计[19]。相比于仅构建一棵树的CART模型，RF无需进行剪枝操作，通过多数投票机制进行最终的决策。RF虽然能够有效避免陷入过度拟合和局部最优，但对于处理不平衡数据问题仍存在不足：一是少数类训练样本由于本身数据量较少，被选中的概率就更低，从而加剧不平衡性；二是少数类训练样本占比较低，训练出来的决策树不能很好地体现占有量少的少数类特点[20]。

1.4.3 ERT模型

ERT同样是一种多棵树集成分类器，与RF不相同的是，为减少每个基分类器的偏差，ERT中的每棵树都是利用整个训练样本进行学习，而在分割树节点时，ERT使用特征随机子集（包括分裂特征和分割值）来训练每个基分类器，表现出极强的随机性[21]。

1.5 模型精度验证

独立的样本数据对模型进行验证能够更好地体现实际模型的精度[22]。本研究从原始样本集中随机选择1/10数据作为测试样本，利用准确率（Accuracy，Ac）、查准率（Precision，Pr）、查全率（又称召回率，Recall，Re）和F1 score（F1）等指标对模型预测总精度和类别精度进行验证。其中，模型总精度Pr、Re和F1计算采用宏平均（Macro average）法。

受试者工作特征曲线（Receiver operating characteristic curve，ROC）源于雷达信号探测分析技术，后被引入数据挖掘和机器学习领域进行分类器性能评价，其显示了输出阈值在所有可能值的范围内移动时模型的敏感性（正确分类的阳性观察值的比例，即真正例率或查准率）和特异性（正确分类的阴性观察值的比例，即真反例率）[23]。ROC曲线以不同阈值下的真正例率（灵敏性）为纵坐标，假正例率（1-特异性）为横坐标绘制而成。对于多分类问题，通常采用One VS Rest和One VS One方法转化为二分类进行计算，可以单独绘制各类别的ROC曲线，也可以绘制出总ROC曲线。AUC（Area under ROC curve）表示ROC曲线下的面积，通过AUC值可以直观地判断不同分类器优劣和泛化性能，AUC值通常介于0.5~1之间，AUC值越大，分类器分类质量越好[24]。

1.6 特征重要性判断

耕地地力特征对病害发生预测的相对重要性可以通过分割树节点的特征使用的相对顺序（即深度）来评估，在树顶部使用的特征会对更大一部分输入样

本的最终预测决策作出贡献。Scikit-learn通过将特征贡献的样本比例与数据集纯度减少相结合得到特征的重要性[25]，并且对该结果进行了归一化处理，计算后的各特征重要性取值在0~1之间，总和为1，且取值越高，其对应的特征对预测决策的贡献越大。

1.7 模型参数设置及训练步骤

本研究中CART、RF和ERT模型构建和评价均使用了Scikit-learn中包含的相应模块，3种模型和SMOTE算法的实现均采用Python编程语言。RF和ERT算法运行时需要设置2个参数，即决策树数目（Ntree）和节点用来分裂时抽取的特征个数（Mtry）。通过大量试验，本研究Ntree设置为300，Mtry设置为总特征的平方根。需要说明的是，Scikit-learn中RF和ERT预测结果的实现是取每个基分类器预测概率的平均，而不是让每个基分类器对类别进行投票。

具体训练步骤如下：①对原始数据进行清洗、格式化和缺省值处理，生成标准数据集；②训练数据集与测试数据集随机划分，调用Scikit-learn中的sklearn.model_selection.train_test_split方法；③使用SMOTE算法对训练集进行随机过采样；④通过调用Scikit-learn中相关API，基于训练集数据样本，通过多次参数调试，构建CART、RF和ERT 3种算法模型分类器，并对测试集数据进行分类预测；⑤利用Ac、Pr、Re和F1等指标进行模型精度评价，并调用Matplotlib.pyplot模块的相关方法绘制出ROC曲线，进行分类器性能评价。

2 结果与分析

2.1 玉米丝黑穗病发病情况空间分布

参照农业行业标准NY/T 1248.3—2006玉米丝黑穗病病情分级标准，进行统计和空间分异分析。由表1可以看出，2008—2014年，玉米丝黑穗病发病程度为轻微、偏轻、中等、偏重的县域分别有304、161、44、6个，未出现严重发病的县域，大部分（90%）玉米主产地区的玉米丝黑穗病发病率在5%以下，玉米品种的不断更新换代以及科学有效地防治对于控制该病发生、发展起到重要作用。病情偏重的6个县域分别是涞源县、青龙满族自治县、五常市、应城市、兴城市和子长县，均出现部分年份玉米丝黑穗病高发的情况，表明小范围、区域性的玉米丝黑穗病防治仍不容忽视（图2）。随着我国玉米播种面积持续增加，加强玉米病虫害监测与防控，减少其引发的玉米产量损失，对于玉米安全生产意义重大。

表1 玉米丝黑穗病发病情况统计

Table 1 Statistics on the incidence of corn head smut

病情级别 Disease level	发病率区间 Incidence interval	病情描述 Disease Description	县域数目 Number of counties	占比 Percentage/%
1	(0,1%]	轻微	304	59
3	(1%,5%]	偏轻	161	31
5	(5%,10%]	中等	44	9
7	(10%,40%]	偏重	6	1
9	(40%,100%]	严重	0	0

注：515个县域玉米发病率取2008—2014年的平均值。

Note: The incidence of corn head smut in 515 counties is the average from 2008 to 2014.

2.2 模型精度比较

为了探讨决策树模型在本研究中的适用性，比较不同模型分类效果，在相同的计算机软硬件环境下，进行了最优模型筛选，以及基于同一测试数据集的分类验证。由表2可以看出，RF分类的准确率、查准率分别为0.843、0.818，ERT为0.824、0.828，均高于CART，基于多棵树的集成分类器的整体分类效果明显优于基于单棵树的分类器分类效果；就集成分类器内部而言，ERT查准率略高于RF，但查全率低于RF，因此通过二者的平衡分数F1得分可以看出，RF综合性能相对较好。具体到不同的病害发生程度，决策树模型在GⅠ上的预测分类Pr、Re和F1值明显大于GⅡ，表明3个模型在GⅠ上使用均呈现出较好的分类效果。然而GⅠ类玉米丝黑穗病病情轻微，对玉米产量影响有限，而准确监测病害中高发情况、减少模型在GⅡ类分类预测中漏分机率，对及时开展预防措施有重要意义。因此，综合考虑衡量，ERT在GⅡ类上的Re最大，能较好地减少中高发病类别预测中

表2 三种模型分类精度比较

Table 2 Comparison of classification accuracy of three models

模型 Models	准确率 Accuracy	查准率 Precision			查全率 Recall			F1得分 F1 score		
		整体 Total	GⅠ	GⅡ	整体 Total	GⅠ	GⅡ	整体 Total	GⅠ	GⅡ
CART	0.765	0.727	0.829	0.625	0.727	0.829	0.625	0.727	0.829	0.625
RF	0.843	0.818	0.947	0.539	0.743	0.857	0.778	0.768	0.899	0.636
ERT	0.824	0.828	0.974	0.385	0.679	0.822	0.833	0.709	0.892	0.526

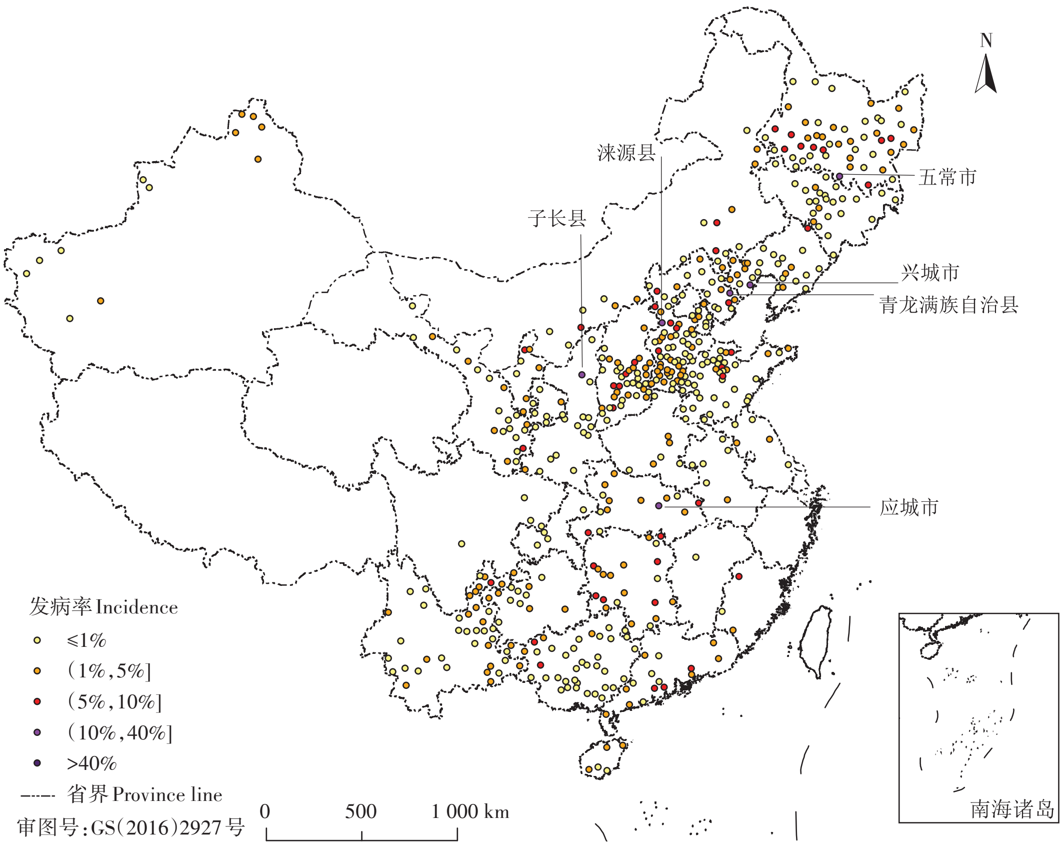


图2 研究区玉米丝黑穗病病情空间分布

Figure 2 Spatial distribution of corn head smut in the study area

漏分机率,提高风险防范能力,模型整体分类效果较佳。且最终确定3个模型选择优先级为ERT>RF>CART。

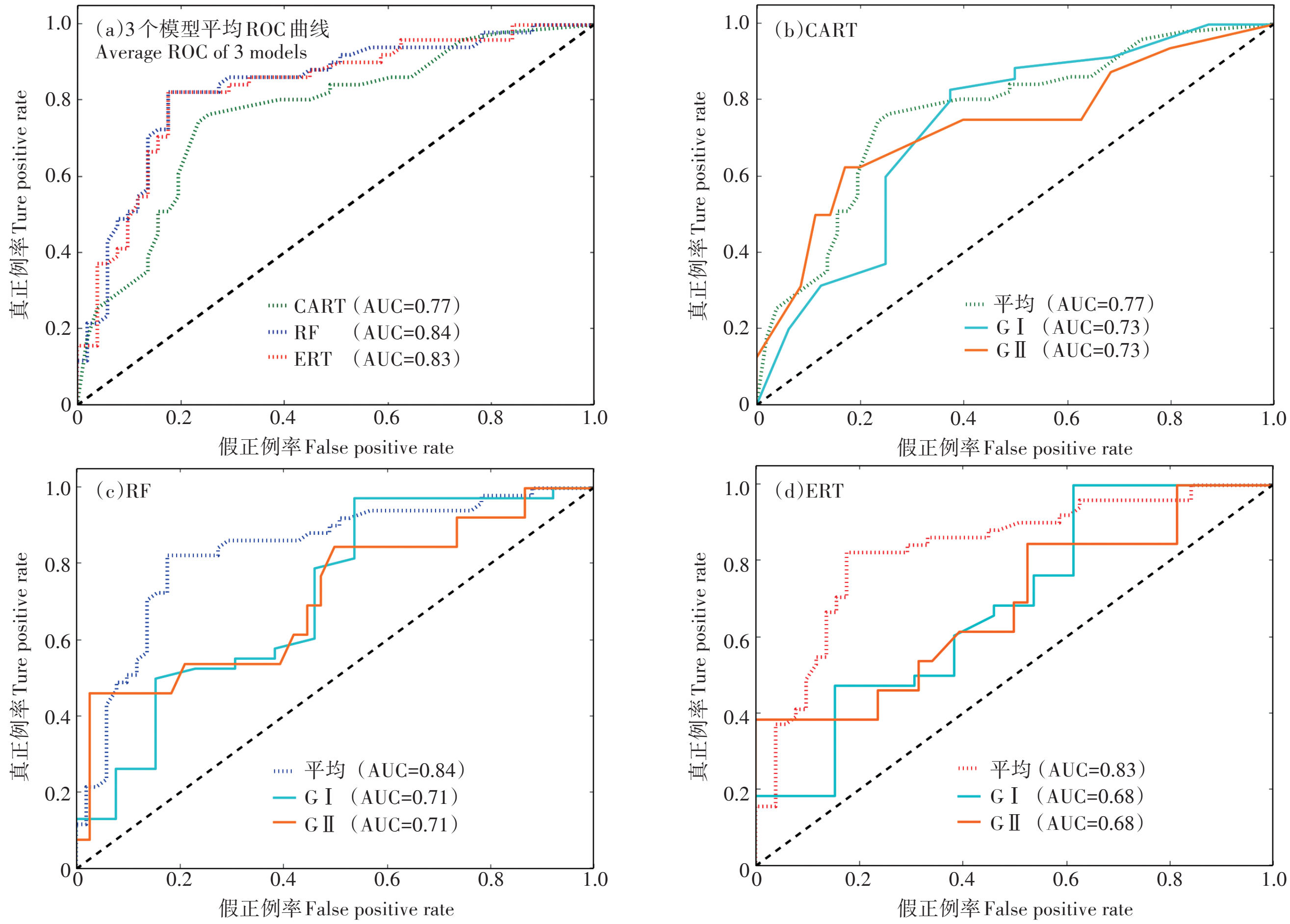
采用ROC曲线和AUC进一步对模型精度、泛化性能进行比较。由图3a可以看出,3个模型的平均ROC曲线均位于坐标点(0,0)和(1,1)连线的左上方,模型分类效果好于随机分类;由于3个模型在空间上存在交叉,难以一般性地推断各模型优劣性,通过CART、RF和ERT 3个模型曲线下面积AUC值(0.77、0.84、0.83),可以判断RF、ERT模型整体性能优于CART模型,RF和ERT二者的性能差距不明显。具体到各个模型(图3b、图3c、图3d),玉米丝黑穗病发病程度GⅠ类和GⅡ类的ROC曲线也均位于坐标点(0,0)和(1,1)连线的左上方,模型预测结果优于“随机猜测”;与各模型的平均ROC曲线相比,虽然在某些阈值上表现出较优的性能,但通过GⅠ类、GⅡ类以及模型整体AUC值比较可以看出,类别整合后的模型整体性能得到提升。

2.3 耕地地力特征变量与病害发生程度重要性分析

耕地地力特征对病害发生预测的相对重要性可以通过分割树节点的特征使用的相对顺序(即深度)来评估,由图4可以看出,耕地地力特征在3种模型预测病害发生程度中的贡献存在一定差异,CART中TP、pH和TK 3个特征的重要性相对较大,RF中各特征重要性程度相对均衡,pH、AP、TP和TK略显突出,ERT中AP、TK、pH和TP 4个特征重要性相对较大,综合来看,耕地地力特征中的TP、pH、AP、TK与玉米丝黑穗病发生程度之间具有较明显的相关性。本研究基于县域单元调查统计数据,从数据挖掘和空间异质性角度研究得出玉米丝黑穗病发生程度与耕地地力因子AP、TK、pH和TP具有较强的相关性,为进一步研究耕地地力对玉米丝黑穗病发生的影响机理提供了线索,但具体影响机理、影响程度以及判定准确性还需进一步从试验角度进行检验和探索。

3 讨论

目前,田间作物的病害防控主要依赖于化学品使用和田间管理措施,不仅费时、昂贵,还会造成环境问题。随着对成本效益和环境无害控制手段的日益重视,从宿主抗性、生物制剂、生物炭以及气候环境变化



坐标点(0,0)和(1,1)连线对应于“随机猜测”模型；AUC指ROC曲线、X轴以及坐标点(1,0)与坐标点(1,1)之间连线组成部分的面积

The line between points (0,0) and (1,1) corresponds to the "random guess" model; AUC refers to the area of the region enclosed by ROC, X axis, and the line between the point (1,0) and (1,1)

图3 3种模型ROC曲线与AUC比较

Figure 3 ROC and AUC of three different methods

等方面开展的致病机理、流行规律[26-27]等研究已成为热点。土传病害受到土壤中多种因素的影响，研究显示：油菜根肿病的发生与pH、TP、TK、碱解氮和速效铁等相关性显著[28]；硒对油菜菌核病发病有较强的影响，其可以提高植物抗性和控制病原活性[29]；改变土壤条件会影响土壤微生物群落结构、提高作物抗性，直接或间接抑制病害发生[30-31]。本研究通过CART、RF和ERT 3种决策树模型，揭示出玉米丝黑穗病发生程度与耕地地力因子AP、TK、pH和TP具有一定的相关性，与丁伟[32]研究得到的氮磷钾肥不同施用量、不同组合配比方式影响玉米丝黑穗病的结果基本一致，其田间试验显示氮肥、磷肥和钾肥对玉米丝黑穗病有一定影响，其中磷肥对病害的控制作用尤为明显，这也间接验证了本研究结果。但本研究未能进一步探究耕地地力因子对玉米丝黑穗病的影响机制，今后有必要从正负相关性、相关程度及机理方面开展深入研究。

在机器学习或统计学习模型中，算法对于数据扰动的鲁棒性是评价模型效果的重要一环。一些最广泛使用的机器学习算法，如人工神经网络和仅构建单棵树的决策树算法，存在输入数据的微小变化而导致模型输出巨大变化的不稳定性和不可靠风险[33-34]。集成学习是一种强大的机器学习方法，它集成了大量的基础模型来生成最终的输出，克服了个体搜索过程中的缺陷，同时在没有足够的数据提供训练时，表现出比单个模型更好的效果和泛化性能[35]。通过对比3个模型精度也可以看出，集成分类器RF和ERT总分类性能明显优于CART，在GⅠ上的分类效果也整体表现优异；但考虑到准确监测病害中高发情况、减少模型在GⅡ类分类预测中漏分机率对开展病害预防的重要性，将3个模型在GⅡ类上的Re大小作为模型效果评判的重点关注指标，综合衡量后确定3个模型

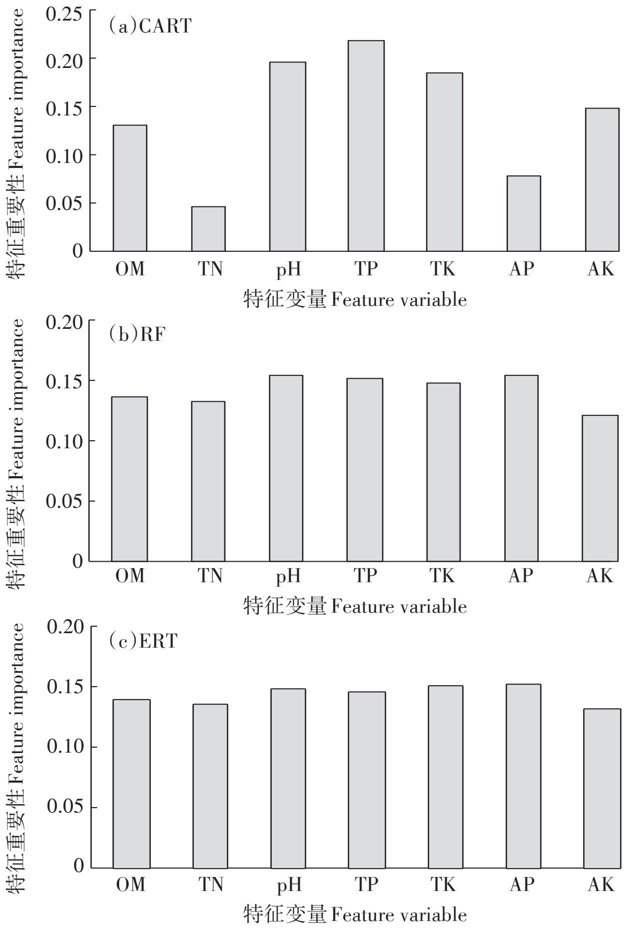


图4 CART、RF和ERT模型特征变量重要性

Figure 4 Feature variable importance of CART, RF and ERT models

的选择优先级为ERT>RF>CART。

此外,本研究基于县域单元调查统计数据,从数据挖掘和空间异质性角度开展研究,数据在空间和时间上的粒度相对粗糙,一定程度上限制了病害发生程度的细化分级和特征向量构建,可能对模型分类效果和特征向量重要性判断造成一定影响,下一步拟基于典型地区微观地块尺度数据进行验证和分析。

4 结论

本研究以县域为研究单元,利用CART、RF和ERT 3种机器学习算法尝试构建了病害发生与耕地地力因子关系模型,对影响玉米丝黑穗病发生程度的耕地地力因子进行判断,并对模型优劣进行比较。研究结论如下:

(1)2008—2014年,我国大部分玉米主产地玉米丝黑穗病为轻微和偏轻发病状态,发病率5%以下地区占比达90%,病情偏重的6个县域分别是涞源县、青龙满族自治县、五常市、应城市、兴城市和子长县。

(2)基于多棵树的集成分类器(RF和ERT模型)分类性能明显优于单棵树分类器(CART模型),分类预测效果较好。但考虑到准确监测病害高发情况、减少中高发病情况在分类预测中漏分机率对开展病害防治的重要性,确定ERT模型为最佳优选分类器。

(3)玉米丝黑穗病发生程度与耕地地力因子AP、TK、pH和TP具有一定的相关性,这为进一步研究耕地地力对玉米丝黑穗病发生的影响机理提供了线索。

致谢:

感谢农业农村部全国农业技术推广服务中心周阳、李春广为本研究提供数据。

参考文献:

[1] 王振营, 王晓鸣. 我国玉米病虫害发生现状、趋势与防控对策[J]. 植物保护, 2019, 45(1):6-16. WANG Z Y, WANG X M. Current status and management strategies for corn pests and diseases in China[J]. *Plant Protection*, 2019, 45(1):6-16.

[2] 刘杰, 姜玉英, 曾娟, 等. 2015年玉米重大病虫害发生特点和趋势分析[J]. 中国植保导刊, 2016, 36(10):53-58. LIU J, JIANG Y Y, ZENG J, et al. Analysis on the occurrence characteristics and trend of maize major diseases and insect pests in 2015[J]. *China Plant Protection*, 2016, 36(10):53-58.

[3] ZHAO X R, YE J R, WEI L, et al. Inhibition of the spread of endophytic *Sporisorium reilianum* renders maize resistance to head smut[J]. *The Crop Journal*, 2015, 3(2):87-95.

[4] MITSUBOSHI M, KIOKA Y, NOGUCHI K, et al. Evaluation of suppressiveness of soils exhibiting soil-borne disease suppression after long-term application of organic amendments by the co-cultivation method of pathogenic *Fusarium oxysporum* and indigenous soil microorganisms[J]. *Microbes and Environments*, 2018, 33(1):58-65.

[5] BITTNER R J, ARELLANO C, MILA A L. Effect of temperature and resistance of tobacco cultivars to the progression of bacterial wilt, caused by *Ralstonia solanacearum*[J]. *Plant and Soil*, 2016, 408(1/2):1-12.

[6] 韦中, 王佳宁, 江高飞, 等. 土传病原细菌的生存与致病权衡[J/OL]. 土壤学报, 2021. DOI:10.11766/trxb202008310399. WEI Z, WANG J N, JIANG G F, et al. Survival-virulence trade-off of soil-borne pathogenic bacteria[J/OL]. *Acta Pedologica Sinica*, 2021. DOI:10.11766/trxb202008310399.

[7] WANG R, ZHANG H C, SUN L G, et al. Microbial community composition is related to soil biological and chemical properties and bacterial wilt outbreak[J]. *Scientific Reports*, 2017, 7(1):343-353.

[8] 苏跃, 吴沿友. 喀斯特地区土壤环境对油菜菌核病的影响[J]. 中国植保导刊, 2016, 36(3):42-46. SU Y, WU Y Y. Effect of soil factors on rape sclerotiniose in karst areas in Guizhou Province[J]. *China Plant Protection*, 2016, 36(3):42-46.

[9] 马超, 杨欣润, 江高飞, 等. 病原青枯菌土壤存活的影响因素研究进展[J/OL]. 土壤学报, 2021. DOI:10. 11766/trxb202008140252. MA C, YANG X R, JIANG G F, et al. Research progresses on key factors affecting survival of *Ralstonia solanacearum* in soils[J/OL]. *Acta Pedolog-*

ica Sinica, 2021. DOI:10. 11766/trxb202008140252.

[10] 程娇, 龚静莲, 汪深, 等. 基于随机森林算法构建白眉野草螟监测预警模型[J]. 植物保护学报, 2019, 46(3):549-555. CHENG J, GONG J L, WANG S, et al. Establishment of monitoring and forecasting model for eastern grass veneer *Agriphila eneociliella* based on the random forest algorithm[J]. *Journal of Plant Protection*, 2019, 46(3):549-555.

[11] 纪思琪, 吴芳, 李乃祥. 基于决策树的蔬菜病害静态预警模型[J]. 天津农学院学报, 2017, 24(2):77-80. JI S Q, WU F, LI N X. Static early warning model of vegetable diseases based on decision tree[J]. *Journal of Tianjin Agricultural University*, 2017, 24(2):77-80.

[12] 王连喜, 徐胜男, 李琪, 等. 基于决策树和混合像元分解的江苏省冬小麦种植面积提取[J]. 农业工程学报, 2016, 32(5):182-187. WANG L X, XU S N, LI Q, et al. Extraction of winter wheat planted area in Jiangsu Province using decision tree and mixed-pixel methods [J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2016, 32(5):182-187.

[13] GUO H X, LI Y J, SHANG J, et al. Learning from class-imbalanced data: Review of methods and applications[J]. *Expert Systems with Applications*, 2016, 73:220-239.

[14] BARUA S, ISLAM M M, YAO X, et al. Mwmote-majority weighted minority oversampling technique for imbalanced data set learning[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2013, 26 (2):405-425.

[15] 高立起, 石爱丽, 董立, 等. 玉米丝黑穗病对玉米产量损失程度研究[J]. 农学学报, 2014, 4(5):12-15. GAO L Q, SHI A L, DONG L, et al. Study on loss rate of the com yield by the effect of head smut[J]. *Journal of Agriculture*, 2014, 4(5):12-15.

[16] 农业部. 玉米抗病虫性鉴定技术规范 第3部分:玉米抗丝黑穗病鉴定技术规范:NY/T 1248.3—2006[S]. 北京:中国农业出版社, 2006. Ministry of Agriculture of the PRC. Rules for evaluation of maize for resistance to pests Part 3:Rule for evaluation of maize for resistance to head smut:NY/T 1248.3—2006[S]. Beijing:China Agriculture Press, 2006.

[17] 农业部. 耕地地力调查与质量评价技术规程:NY/T 1634—2008 [S]. 北京:中国农业出版社, 2008. Ministry of Agriculture of the PRC. Rules for soil quality survey and assessment:NY/T 1634—2008 [S]. Beijing:China Agriculture Press, 2008.

[18] 马慧琴. 基于多源数据的小麦白粉病遥感监测与预测模型研究 [D]. 南京:南京信息工程大学, 2017. MA H Q. Remote sensing monitoring and forecasting models of wheat powdery mildew based on multi-source data[D]. Nanjing:Nanjing University of Information Science & Technology, 2017.

[19] 曹正凤. 随机森林算法优化研究[D]. 北京:首都经济贸易大学, 2014. CAO Z F. Study on optimization of random forests algorithm [D]. Beijing:Capital University of Economics and Business, 2014.

[20] 吴琼, 李运田, 郑献卫. 面向非平衡训练集分类的随机森林算法优化[J]. 工业控制计算机, 2013, 26(7):89-90. WU Q, LI Y T, ZHENG X W. Optimized random forests algorithm for imbalanced training sets[J]. *Industrial Control Computer*, 2013, 26(7):89-90.

[21] AHMAD M W, REYNOLDS J, REZGUI Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees[J]. *Journal of Cleaner Production*, 2018, 203:810-821.

[22] 唐翠翠, 黄文江, 罗菊花, 等. 基于相关向量机的冬小麦蚜虫遥感预测[J]. 农业工程学报, 2015, 31(6):201-207. TANG C C, HUANG W J, LUO J H, et al. Forecasting wheat aphid with remote sensing based on relevance vector machine[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2015, 31(6):201-207.

[23] ROBIN X, TURCK N, HAINARD A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves[J]. *BMC Bioinformatics*, 2011, 12:77.

[24] 张海娟, 陈勇, 黄烈健, 等. 基于生态位模型的薇甘菊在中国适生区预测[J]. 农业工程学报, 2011, 27(增刊1):413-418. ZHANG H J, CHEN Y, HUANG L J. Predicting potential geographic distribution of *Mikania micrantha* planting based on ecological niche models in China[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2011, 27(Supp1):413-418.

[25] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in Python[EB/OL]. (2011) [2021-04-02]. https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees.

[26] LI X H, WANG Z H, GAO S R, et al. Analysis of QTL for resistance to head smut(*Sporisorium reiliana*) in maize[J]. *Field Crops Research*, 2008, 106(2):148-155.

[27] 马艳, 王光飞. 生物炭防控植物土传病害研究进展[J]. 中国土壤与肥料, 2014(6):18-24. MA Y, WANG G F. Review of biochar utilization on soil-borne disease control[J]. *Soil and Fertilizer Sciences in China*, 2014(6):18-24.

[28] 孙超超. 我国油菜主产区土壤环境条件对根肿病发生的影响研究 [D]. 北京:中国农业科学院, 2014. SUN C C. Effects of soil factors on clubroot disease in main rapeseed production area[D]. Beijing:Chinese Academy of Agricultural Sciences, 2014.

[29] LIU K, CAI M M, HU C X, et al. Selenium(Se) reduces *Sclerotinia* stem rot disease incidence of oilseed rape by increasing plant Se concentration and shifting soil microbial community and functional profiles[J]. *Environmental Pollution*, 2019, 254(Part B):113051.

[30] ROS M, ALMAGRO M, JUAN A F, et al. Approaches for the discrimination of suppressive soils for *Pythium irregulare* disease[J]. *Applied Soil Ecology*, 2020, 147:103439.

[31] CHEN S, QI G F, MA G Q, et al. Biochar amendment controlled bacterial wilt through changing soil chemical properties and microbial community[J]. *Microbiological Research*, 2020, 231:126373.

[32] 丁伟. 肥料对玉米丝黑穗病的影响研究[D]. 北京:中国农业科学院, 2014. DING W. Effects of fertilizer on maize head smut[D]. Beijing:Chinese Academy of Agricultural Sciences, 2014.

[33] AHMAD M W, REYNOLDS J, REZGUI Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees[J]. *Journal of Cleaner Production*, 2018, 203:810-821.

[34] WANG Z Y, WANG Y R, ZENG R C, et al. Random forest based hourly building energy prediction[J]. *Energy and Buildings*, 2018, 171 (15):11-25.

[35] FAN C, XIAO F, WANG S W. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques[J]. *Applied Energy*, 2014, 127:1-10.