doi:10.6041/j.issn.1000-1298.2022.09.029

基于 LightGBM 和处方数据的番茄病害诊断方法

徐　畅[1]　丁俊琦[1]　赵聃桐[1]　乔　岩[2]　张领先[1]

(1.中国农业大学信息与电气工程学院,北京 100083;2.北京市植物保护站,北京 100029)

摘要:为高效地挖掘植物病害处方数据并辅助精准诊断,以番茄病毒病、番茄晚疫病、番茄灰霉病 3 种病害为研究对象,构建基于贝叶斯优化 LightGBM 的番茄病害智能诊断模型,探索作物病害处方数据挖掘及其精准诊断。重点对处方原数据(文本数据标签和 One-hot 编码等)进行预处理,以基于 Wrapper 的递归特征消除法进一步提取作物病害处方数据的特征;利用基于 LightGBM 算法构建番茄病害诊断模型,并与 K 近邻(KNN)、决策树(DT)、支持向量机(SVM)、随机森林(RF)、梯度提升决策树(GDBT)、AdaBoost 和 XGBoost 常见机器学习模型运行结果进行比较分析并进行优化;设计基于 LightGBM 模型的 Android 手机端植物医生病害诊断 APP。实验结果表明,基于贝叶斯优化的 LightGBM 模型综合诊断准确率可达到 89.11%,比其他 7 种机器学习模型的诊断准确率平均高 3.65 个百分点;同时特征选择后的 LightGBM 模型在保证模型准确率的基础上降低了前期数据收集难度,模型综合准确率提高至 89.34%,其中番茄病毒病的诊断精确度和 F1 值均达到 96% 以上,运行时间减少了 47.73%;最后通过番茄叶霉病和番茄早疫病两种病害对本文模型进行了泛化能力测试,实验结果表明该模型具有较强的泛化能力和实用性。基于 LightGBM 模型设计的 APP 可以实现用户人群友好的交互式可视化且满足实际诊断需求。

关键词:番茄病害;处方数据;病害诊断;LightGBM;Android

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1000-1298(2022)09-0286-09　　**OSID**:

Tomato Disease Diagnosis Method Based on LightGBM and Prescription Data

XU Chang[1]　DING Junqi[1]　ZHAO Dantong[1]　QIAO Yan[2]　ZHANG Lingxian[1]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*
2. *Beijing Plant Protection Station*, *Beijing* 100029, *China*)

Abstract: Aiming at the problem of how to efficiently mine prescription big data and assist in accurate diagnosis, tomato virus disease, tomato late blight and tomato gray mold were selected as the research objects, and an intelligent diagnosis model of tomato disease based on Bayesian optimization LightGBM was constructed to explore the data mining and accurate diagnosis of crop disease prescription. The primary data (text data label and One-hot coding, etc.) were preprocessed, and the features of crop disease prescription data were further extracted by recursive feature elimination method based on Wrapper. The tomato disease diagnosis model was constructed based on LightGBM algorithm, and compared with the running results of K-nearest neighbor (KNN), decision tree (DT), support vector machine (SVM), random forest (RF), gradient boosting decision tree (GDBT), AdaBoost and XGBoost common machine learning models. An Android mobile terminal plant doctor disease diagnosis APP was designed based on LightGBM model. The experimental results showed that the comprehensive diagnosis accuracy of LightGBM model based on Bayesian optimization can reach 89.11%, which was 3.65 percentage points higher than that of other seven machine learning models on average. At the same time, the LightGBM model after feature selection reduced the difficulty of data collection in the early stage on the basis of ensuring the accuracy of the model, and the comprehensive accuracy of the model was improved to 89.34%. Among them, the diagnostic accuracy of tomato virus disease and F1-score could reach more than 96%, and the running time was reduced by 47.73%. Finally, the generalization ability of the proposed model was tested by tomato leaf mildew and tomato early blight, and the experimental results

收稿日期:2021-09-24　修回日期:2021-12-01

基金项目:国家自然科学基金项目(62176261)和现代农业产业技术体系北京市叶类蔬菜创新团队建设项目(BAIC07-2022)

作者简介:徐畅(1998—),女,博士生,主要从事农业信息化技术研究,E-mail:xc199873@126.com

通信作者:张领先(1970—),男,教授,博士生导师,主要从事农业信息化技术研究,E-mail:zhanglx@cau.edu.cn

indicated that the model had strong generalization ability and practicability. The APP designed based on LightGBM model can realize user friendly interactive visualization and meet the actual diagnostic needs.
Key words: tomato disease; prescription data; disease diagnosis; LightGBM; Android

0　引言

作物处方数据包含作物信息、环境信息、病害信息及其诊断知识，同时基于宿主、病原体和环境的传统流行病学和植物病理学知识为处方数据分析提供了新的研究视角。如何有效挖掘作物处方数据多源信息间的内在关系以及辅助精准诊断是一个亟待解决的问题。

基于处方数据挖掘的作物病害诊断问题可转化为计算机领域的多分类解决方法。机器学习模型在分类方面具有计算时间短、精度高、可移植性强的优点[1]。而现实中数据量的不断增加和数据的多元化，尤其是作物处方数据的复杂性和专业性，使传统的分类算法不能很好地满足现有数据的处理以及实际问题的解决需求。集成学习模型在分类问题上显示出了极大的优势[2]，利用基础算法的多样性可以提高集成模型的分类准确率、泛化能力和鲁棒性[3]。传统的 Boosting[4] 算法（如 GBDT 和 XGBoost）需要对每一个特征扫描所有的样本点来选择最佳的切分点，因此在效率和可扩展性上不能再满足相应的需求。为了解决在大样本、高维度数据环境下算法耗时长的问题，LightGBM[5] 使用了如下两种解决方法：梯度单边采样（Gradient-based one-side sampling，GOSS），对样本进行采样来计算梯度，而不是使用所有的样本点计算梯度；互斥特征捆绑（Exclusive feature bundling，EFB）方法将某些特征捆绑在一起降低特征的维度，并非对所有的特征进行扫描，从而使得寻找最佳切分点的消耗减少。因此可以在大幅降低样本处理时间复杂度的基础上，保证 LightGBM 的精度稳定或提升[6]。目前，LightGBM 模型已成功应用于不同领域[7-9]。

番茄是一种病害高发的作物[10]，有效防治病害是番茄种植中的重要工作[11]。对此，本文以番茄病毒病、番茄晚疫病、番茄灰霉病 3 种常见病害为研究对象，构建基于 LightGBM 的番茄病害智能诊断模型，探索一种针对处方多维大数据特点的番茄病害诊断方法。

1　诊断模型建立

基于 LightGBM 和处方数据的番茄病害诊断流程如图 1 所示。主要步骤如下：①获取番茄病害处方原始数据。②选取实验数据。③对数据进行统计分析，了解不同区域、不同生长期番茄病害发展情况。④对数据进行预处理（编码等）。⑤利用 RFECV 和 GDBT 结合的方法进行特征选择。⑥将优选后的特征作为诊断模型的输入，比较 8 种常见分类模型，以最优模型作为番茄病害诊断模型。⑦输出最终诊断结果并计算准确率。

1.1　作物病害处方数据集构建

本研究数据来源于北京市植物保护站提供的处方数据库。北京市在全国首次引入国际应用生物科学中心的植物智慧解决方案和植物诊所先进理念，

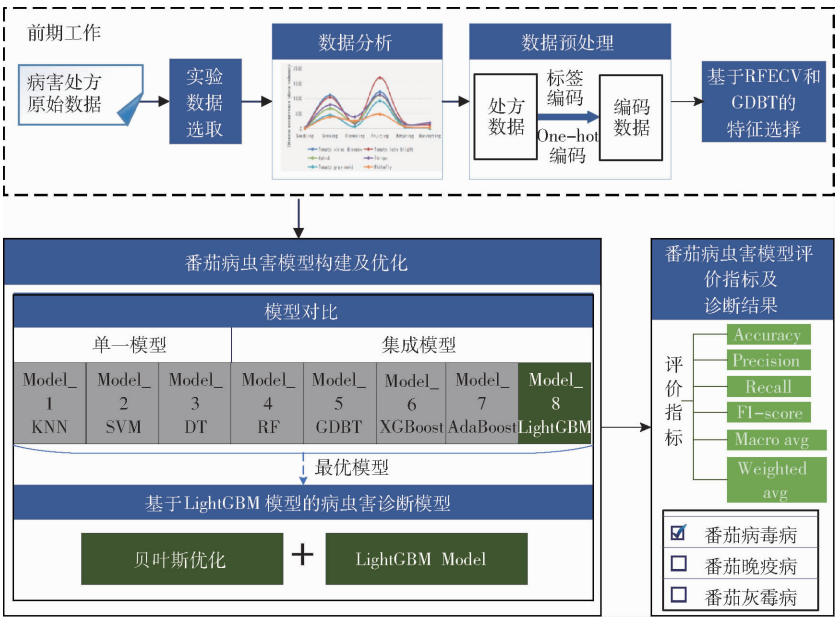

图 1　基于 LightGBM 和处方数据的番茄病害诊断流程

Fig. 1　Flowchart of tomato disease diagnosis based on LightGBM and prescription data

先后建立植物诊所 115 家,服务范围覆盖京郊 161 个乡镇,覆盖率达到 88.95%。以植物医生开具处方的形式为生产者提供病害智慧诊断和防控技术咨询。植物诊所依托病害动态防控解决方案库和标准化处方形成机制,植物医生现场面对面为农户开具病害诊断防控处方。基于双向信息流的原则,开出包括问诊农户、植物诊所、作物和症状、诊断结果和防控建议的标准化处方。同时遵循有害生物综合治理(Integrated pest management,IPM)原则,包括预防在内,以物理防控、生物防控为主的绿色、综合防控配套措施。植物医生开具的所有处方经过“植物医生填写录入—区级数据管理员初步协调验证—市级数据管理员二次协调验证”的三级数据验证程序,合格率达 90% 以上,并导入处方数据库,处方开具流程如图 2 所示。目前已经形成了 127 种病害绿色防控技术体系及 20 余万条处方数据知识库。

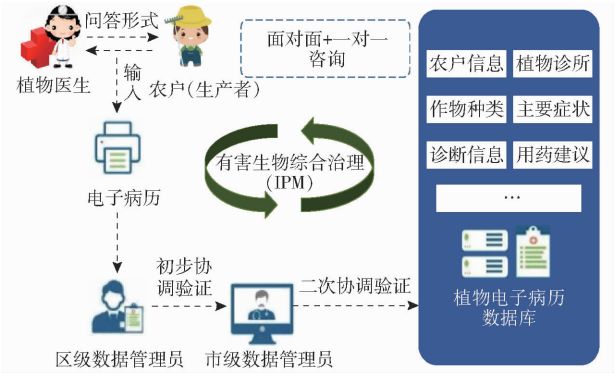

图 2 处方开具流程

Fig. 2 Prescribing process

本研究以番茄病毒病、晚疫病、灰霉病 3 种病害作为研究对象,选取 2019 年 3 月 25 日—2020 年 11 月 19 日番茄多种病害处方数据作为作物病害处方信息的数据源。每条处方数据包含所属区、时间、发育阶段、受害部位、发生面积、发生比重、主要症状、问诊记录、田间症状分布、诊断结果、开具农药名称、开具农药数量等共计 7 344 条处方数据,每种病害统计结果如表 1 所示。

表 1 实验数据统计结果

Tab. 1 Experimental data statistical results

病害名称	数据量/条	占比/%
番茄病毒病	2 607	35.5
番茄晚疫病	3 248	44.2
番茄灰霉病	1 489	20.3
总计	7 344	100

1.2 实验数据预处理

本研究数据预处理流程包括对源数据文件整理、转换,数据清洗(删除重复值、缺失值处理、一致化处理和异常值处理),数据统计,最后对输出数据进行编码(标签编码和 One-hot 编码),如图 3 所示。

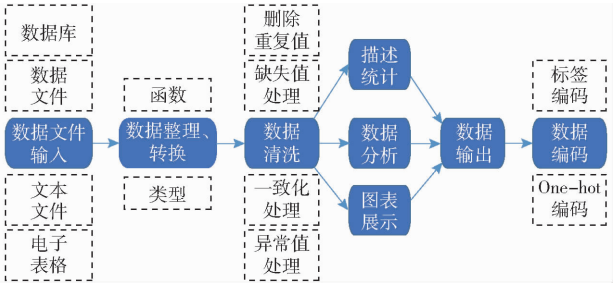

图 3 处方数据预处理流程图

Fig. 3 Prescription data pre-processing flowchart

1.2.1 文本数据标签编码

根据对处方数据的分析,将所属区域属性对应的顺义、海淀、延庆、密云、怀柔、平谷、通州、朝阳、昌平、丰台、大兴、房山 12 个区标注为 R1 ~ R12,将发育阶段属性对应的苗期、生长期、开花期、结果期、成熟期、收获期 6 个阶段标注为标签 S1 ~ S6,将发生比重属性对应的轻度发生、中度发生、重度发生 3 个比重标注为标签 P1 ~ P3。

1.2.2 文本数据 One-hot 编码

One-hot 编码是将类别变量转换为机器学习算法易于处理的数据,有利于进行损失函数或准确率计算。由于生产者对于受害部位、主要症状和田间症状分布的描述展示为多项选择,因此对以上 3 个属性的多选值进行 One-hot 编码,将受害部位的根、茎基部、茎、嫩叶/树枝、叶片、花、果实/谷粒、整株植物、嫩芽 10 个选项值和萎蔫、矮化、落叶、花叶、水疱状、腐烂等 27 个选项值以及局部分布、分散分布、线条分布、田地边缘、均匀分布等 9 个选项值进行编码。

1.3 特征选择方法

在作物病害处方数据集中,大量无关、冗余或噪声特征的存在不仅会带来维数升高的问题,还会直接影响分类器性能[12]。特征选择通过删除无关的冗余数据,为提高学习准确性提供了一个有效的解决方法,可以有效减少计算时间[13]。常用的特征选择方法可以分为过滤式(Filter)、封装式(Wrapper)和嵌入式(Embedded)[14]。而 Wrapper 方式的优势在于将特征选择问题转化为特征子集搜索问题,采用学习器分类性能作为特征子集的评价标准[15]。GDBT 是一种迭代的决策树算法,通过多轮迭代降低偏差,达到提高分类精度的目的[16]。结合子算法的优势,混合方法通常比传统方法更稳健,因此可以作为本研究处方数据的特征选择方法。

选取 Wrapper 方式的递归特征消除(RFE)与交

叉验证结合(RFECV)的方法对作物病害处方数据特征进行提取。此时的交叉验证(CV)为不同列(特征)的组合求均值,用于求得最优的特征数量。RFECV 算法对具体作物病害处方数据特征提取分为 2 个阶段:①RFE 阶段:将编码后的原始处方数据作为模型的输入变量,通过不断训练 GDBT 模型,根据 feature_importances_属性对特征进行重要性评级,每次训练完成后删除最不重要的一个或多个特征,然后对筛选出的特征集合再次进行训练,直到遍历完整个特征集,最终筛选出重要的特征变量个数为 d。②CV 阶段:对于数量为 d 的特征集合,其所有子集的个数为 2^d-1,将所有由不同特征数量构成的子集依次输入 GDBT 分类器,根据最高分类准确率输出的特征子集即为最优特征子集。RFECV 算法流程图如图 4 所示。

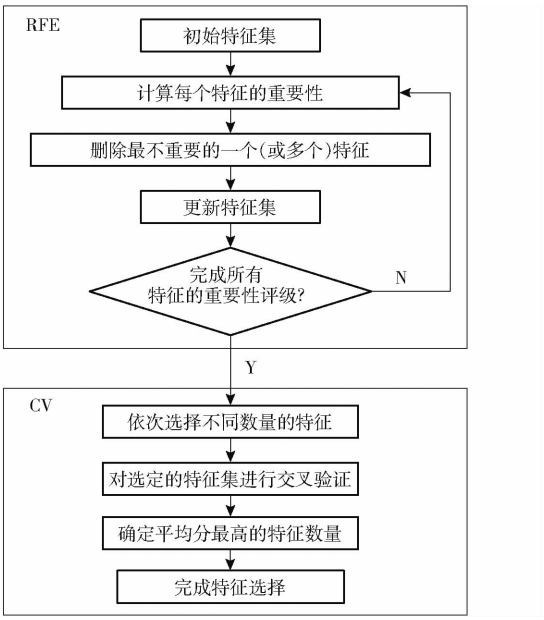

图 4　RFECV 算法流程图

Fig. 4　Flowchart of RFECV

1.4　基于 LightGBM 和处方数据的番茄病害诊断模型

1.4.1　LightGBM 算法

集成学习方法是指将多个学习模型组合,以获得更好的处理效果,使组合后的模型具有更强的泛化能力。LightGBM 是基于 Boosting 的集成量级高效梯度提升树,具有高效、低内存、高准确率的优点,同时支持并行化学习,可以处理大规模数据。LightGBM 相比于其他的 Boosting 集成方法,增加了梯度单边采样(GOSS)和互斥特征绑定(EFB)算法。

GOSS 算法的主要思想是:从减少样本角度,排除大部分小梯度的样本,仅用剩下的样本计算信息增益。由于梯度大的样本点会贡献更多的信息增益,为了保持信息增益评估的精度,GOSS 保留所有的梯度较大的实例,在梯度小的实例上使用随机采样。

EFB 算法的主要思想是:从减少特征角度,将互斥特征绑定在一起,保证信息完整性的同时提升计算效率。对于如何合并互斥特征的问题,LightGBM 的解决办法是利用直方图(Histogram)算法,将连续的特征离散化为 k 个离散特征,同时构造一个宽度为 k 的直方图用于统计信息(含有 k 个 bin)。利用直方图算法无需遍历所有数据,只需要遍历 k 个 bin 即可找到最佳分裂点。

1.4.2　基于贝叶斯优化算法的模型参数优化

贝叶斯优化算法[17](Bayesian optimization algorithm,BOA)是用于求解表达式未知的函数极值问题的方法,在参数组合寻优问题上被广泛应用。此方法可以利用之前已搜索的信息确定下一个搜索点,提高结果的质量以及搜索的速度,因此比网格搜索和随机搜索更为有效,具有迭代次数少、参数粒度小等优点。贝叶斯优化算法的核心由两部分构成:①先验函数(Prior function,PF):对目标函数进行建模,即计算每一点处的函数值均值 $\mu(x)$ 和方差 $\delta(x)$,通常用高斯过程回归实现。②采集函数(Acquisition function,AC):通过采集函数确定下一个采样点[18],采样点的选择要综合考虑利用和探索,以保证最大限度地找到全局最优解。采集函数主要包括期望改善(Expected improvement,EI)、概率改善(Probability of improvement, PI)、置信区间上界(Upper confidence bound,UCB)等方法,本实验选取 UCB 函数作为采集函数,其数学表达式为

$$\mathrm{UCB}(x)=\mu(x)+\varepsilon\delta(x) \tag{1}$$

式中　ε——权重

其中,参数 ε 用于平衡采样点的选择,有助于找到全局最优值。贝叶斯优化算法具体流程为:

选择 n_0 个采样点,计算 $f(x)$ 在采样点处的值

$n=n_0$

while $n\leqslant N$ do

　　根据当前采样数据 $D=\{(x_i,f(x)),i=1,2,\cdots,n\}$ 更新 $p(f(x)\mid D)$ 的均值和方差

　　根据 $p(f(x)\mid D)$ 均值和方差计算采集函数 $u(x)$

　　根据采集函数的极大值确定下一个采样点 $x_{n+1}=\underset{x}{\operatorname{argmax}}u(x)$

　　计算下一个采样点处的函数值:$y_n=f(x_{n+1})$

　　$n=n+1$

end while

return:$\operatorname{argmax}(f(x_1),f(x_2),\cdots,f(x_N))$ 以及对应的 y

算法首先初始化 n_0 个候选解,通常在整个可行域内

均匀地选取一些点。然后开始循环,每次增加一个点,直至找到 N 个候选解。每次寻找下一个点时,用已经找到的 n 个候选解建立高斯回归模型,得到任意点处的函数值的后验概率。然后根据后验概率构造采集函数,寻找函数的极大值点作为下一个搜索点。接下来计算在下一个搜索点处的函数值。算法最后返回 N 个候选解的极大值作为最优解。

1.4.3 基于 LightGBM 的番茄病害诊断模型

利用贝叶斯优化算法对 LightGBM 进行参数寻优时,以 LightGBM 的不同超参数组合作为自变量 x,以 5 折交叉验证评估得到的准确率(Accuracy)作为贝叶斯框架的输出 y。基于贝叶斯优化的 LightGBM 的番茄病害诊断模型具体步骤如图 5 所示。

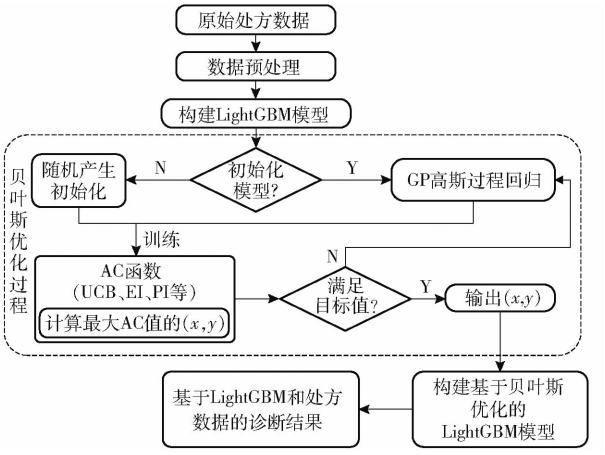

图 5 基于 LightGBM 的番茄病害诊断模型原理图

Fig. 5 Principle of tomato disease diagnosis model based on LightGBM

1.5 评价指标

对于二分类问题,常见的评价指标是精确率、召回率、F1 值和准确率[19]。

F1 值可以同时兼顾分类模型的精确率和召回率,是分类问题常见的评价指标,F1 值越高,表示分类器的综合性能越好。在本研究中对于番茄的某一种病害的诊断是二分类问题,对于番茄多种病害的诊断是多分类问题。对于多分类问题,其评价指标可扩展为宏平均和加权平均。宏平均是所有类别的评价指标(精确率、召回率和 F1 值)的算术平均值,但是此评价方法忽略了样本之间可能存在不平衡问题,因此加权平均在计算各个评价指标时先乘以该类在总样本中的占比再进一步求和。

2 实验结果与分析

2.1 特征选择结果分析与讨论

图 6 为通过 RFECV 结合 GDBT 得到的特征选择个数与模型交叉验证准确率的关系,本研究将原始数据的 50 个特征变量输入 GDBT 模型,前期交叉验证准确率基本一直保持上升趋势,说明特征越多,模型的准确率越高,模型准确率随着特征的增加先增加后趋于平稳,经实验得到选择的最佳特征个数为 32 个,包括所属区域、发育阶段、发生面积、发生比重、受害部位(整株植物、果实/谷粒、根等 5 个特征)、主要症状(丛枝、干枯、叶斑等 19 个特征)、田间分布症状(仅个别植株、分散分布等 4 个特征),作为番茄病害诊断模型的输入。

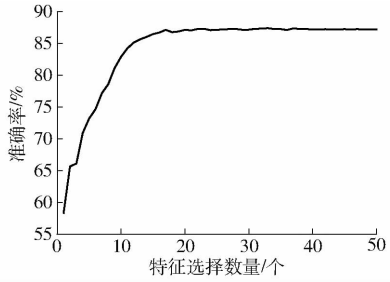

图 6 基于 RFECV 和 GDBT 的特征选择结果

Fig. 6 Feature selection results based on RFECV and GDBT

2.2 模型结果与分析

2.2.1 基于 LightGBM 的番茄病害诊断模型参数优化

通过构建 LightGBM 模型对 3 种番茄病害进行诊断,并同时基于 KNN、决策树、SVM、GDBT、随机森林、AdaBoost 和 XGBoost 进行建模以对比分析不同模型的分类性能,进一步验证本文提出的基于 LightGBM 的模型诊断能力。利用贝叶斯优化方法对本研究建立的 LightGBM 模型的 10 个重要参数进行优化,经过 5 折交叉验证获得最优超参数,其他参数均为默认值。待优化参数、选择范围及最终结果如表 2 所示。训练集和测试集的比例设置为 7∶3。

表 2 参数优化结果

Tab. 2 Parameter optimization results

待优化参数	数值范围	最优数值
colsample_bytree	(0.5,1)	0.56
min_child_samples	(2, 200)	9
num_leaves	(5, 1 000)	884
subsample	(0.6, 1)	0.7
max_depth	(2,10)	6
n_estimators	(10, 1 000)	1 000
reg_alpha	(0,10)	1.08
reg_lambda	(0,10)	4.8
min_gain_to_split	(0,1)	0.15
learning_rate	(0,1)	0.1

2.2.2 基于 LightGBM 的番茄病害诊断模型结果分析

根据表 3 可知,基于 LightGBM 的番茄病害诊断模型在番茄病毒病、番茄晚疫病和番茄灰霉病 3 个类别数据上均达到较好的分类效果。其中,在番

表 3　基于 LightGBM 的番茄病害诊断模型实验结果

Tab. 3　Experimental results of tomato disease diagnosis model based on LightGBM　%

特征选择	类别	精确率	召回率	F1 值
特征选择前	番茄病毒病	97.27	94.68	95.95
	番茄晚疫病	86.96	90.41	88.65
	番茄灰霉病	79.73	76.75	78.21
	准确率		89.11	
	宏平均	87.98	87.28	87.61
	加权平均	89.15	89.11	89.11
特征选择后	番茄病毒病	97.27	94.80	96.02
	番茄晚疫病	87.09	90.72	88.87
	番茄灰霉病	80.50	76.97	78.70
	准确率		89.34	
	宏平均	88.29	87.50	87.86
	加权平均	89.37	89.34	89.32

茄病毒病上的分类效果最佳，精确率和召回率分别达到 97.27% 和 94.68%，同时综合评价指标 F1 值可达到 95.95%。相比于番茄病毒病、番茄晚疫病和番茄灰霉病的分类效果稍显逊色，其原因在于两种病害均是实际田间环境中常见而难以区分的病害。两者在发病初期在果实上以灰白色霉层出现，发病后期在叶片上以深棕色斑点出现。特别是番茄晚疫病，此病害蔓延迅猛，短期内可造成毁灭性伤害，因此利用基于 LightGBM 的病害诊断模型的优势，对番茄病害及时防治具有重要现实意义。

2.2.3　对比实验结果分析

基于 LightGBM 的番茄病害诊断模型结果较优。根据表 4 中准确率、宏平均和加权平均 3 个评价指标的结果可知，LightGBM 模型整体的分类性能较好。与表 4 的常用机器学习方法相比较，无论是单一的机器学习模型（KNN、DT 和 SVM），还是基于 bagging 集成框架的 RF 算法和基于 Boosting 集成框架的其它算法（AdaBoost、GDBT 和 XGBoost），LightGBM 模型的表现最佳，准确率平均高于其它模型 3.65 个百分点。图 7 为 LightGBM 与其他 7 种机器学习方法的比较。

2.2.4　特征选择结果分析

特征选择后的 LightGBM 模型在保证模型准确率的基础上降低了前期数据收集难度。本研究利用 RFECV 和 GDBT 结合的特征选择方法，进一步减轻前期数据采集的困难，提高模型运行效率。由图 8 可以看出，本文提出的特征选择算法的可靠性和稳定性。由图 8 可知，所有机器学习模型均可在一定程度上减少运行时间，实验结果表明特征选择前后每个模型运行时间平均降低 20.37%，LightGBM 模型的效果最为显著，时间消耗减少了 47.73%，模型准确率提升至 89.34%。

表 4　特征选择前算法实验结果对比

Tab. 4　Comparison of algorithm experimental results before feature selection　%

算法	类别	精确率	召回率	F1 值
KNN	番茄病毒病	90.10	88.85	89.47
	番茄晚疫病	81.02	85.92	83.40
	番茄灰霉病	74.08	66.45	70.06
	准确率		82.94	
	宏平均	81.74	80.41	80.98
	加权平均	82.84	82.94	82.81
DT	番茄病毒病	94.55	94.55	94.55
	番茄晚疫病	85.88	86.24	86.06
	番茄灰霉病	74.34	73.68	74.01
	准确率		86.62	
	宏平均	84.92	84.82	84.87
	加权平均	86.59	86.62	86.60
SVM	番茄病毒病	93.93	92.14	93.03
	番茄晚疫病	83.14	88.43	85.70
	番茄灰霉病	77.07	69.30	72.98
	准确率		85.80	
	宏平均	84.71	83.29	83.90
	加权平均	85.75	85.80	85.69
RF	番茄病毒病	97.12	93.92	95.49
	番茄晚疫病	84.70	91.24	87.85
	番茄灰霉病	80.64	72.15	76.16
	准确率		88.25	
	宏平均	87.49	85.77	86.50
	加权平均	88.31	88.25	88.17
AdaBoost	番茄病毒病	97.59	82.26	89.27
	番茄晚疫病	69.73	92.70	79.59
	番茄灰霉病	72.35	41.89	53.06
	准确率		78.45	
	宏平均	79.89	72.28	73.97
	加权平均	80.24	78.45	77.56
GDBT	番茄病毒病	96.99	93.79	95.36
	番茄晚疫病	82.70	93.22	87.65
	番茄灰霉病	82.78	65.35	73.04
	准确率		87.66	
	宏平均	87.49	84.12	85.35
	加权平均	87.83	87.66	87.39
XGBoost	番茄病毒病	97.51	94.30	95.88
	番茄晚疫病	85.57	90.30	87.87
	番茄灰霉病	79.25	74.56	76.84
	准确率		88.48	
	宏平均	87.45	86.39	86.86
	加权平均	88.54	88.48	88.46

2.2.5　LightGBM 诊断模型泛化能力测试结果分析

基于 LightGBM 的番茄病害诊断模型具有一定实用性且泛化能力强。为验证本研究提出模型的泛化能力和实用性，数据集增加番茄叶霉病和番茄早疫病 2 种番茄病害种类，数据量分别为 390 条和

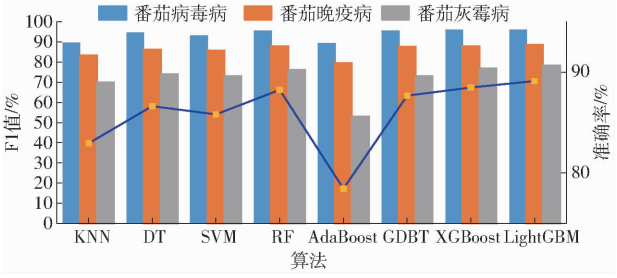

图 7 各算法在 3 种病害上的 F1 值和模型准确率

Fig. 7 F1-score of each algorithm on three kinds of diseases and model accuracy

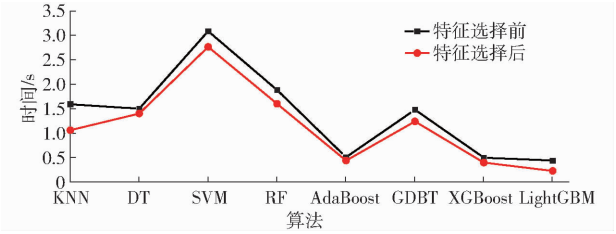

图 8 特征选择前后算法时间对比

Fig. 8 Comparison of algorithm time before and after feature selection

947 条。由表 5 可以看出，本研究提出的多分类模型在番茄病害四分类和五分类的实验效果都较佳，其中番茄病毒病、晚疫病、灰霉病和叶霉病 4 种病害分类的准确率达到 88.37%。

表 5 四分类和五分类的实验结果

Tab. 5 Experimental results of four-class classification and five-class classification

病害种类	类别	精确率/%	召回率/%	F1 值/%
四分类	番茄病毒病	77.28	71.46	74.26
	番茄晚疫病	95.11	95.11	95.11
	番茄灰霉病	97.44	94.21	95.80
	番茄叶霉病	86.59	89.85	88.19
	准确率		88.37	
	宏平均	89.11	87.66	88.34
	加权平均	88.25	88.37	88.27
五分类	番茄病毒病	71.89	70.91	71.40
	番茄晚疫病	95.18	94.93	95.05
	番茄灰霉病	97.22	97.22	97.22
	番茄叶霉病	77.58	84.13	80.72
	番茄早疫病	63.68	47.33	54.30
	准确率		81.31	
	宏平均	81.11	78.90	79.74
	加权平均	80.89	81.31	80.91

3 Android 客户端构建与应用

3.1 设计目标

为研究和掌握作物多种病害发生的规律和特点，根据处方大数据分析总结的经验，及时准确为生产者和决策者提供病害预警提示、对症防治、科学用药和辅助决策，结合 LightGBM 模型，设计适用于普通农户行为习惯、简明方便的“植物健康”APP 界面模式。

3.2 结构与功能

针对植物医生-农户诊断服务流程，基于 Android 系统应用程序开发技术[20]，选用 Android Studio 平台，采用 Android MVC 设计模式，此模型具有耦合性低的特点，使得 View（视图）层和 Model（模型）层可以很好地分离，达到解耦的目的，减少模块代码间的相互影响，利于开发人员维护。同时使用轻量级 SQLite 数据库，实现无服务器、零配置、事务性的 SQL 数据库引擎，可以按应用程序需求进行静态或动态连接。

根据设计目标，本系统设计了 2 个主要功能模块，分别是用户功能模块和管理员功能模块（图 9）。

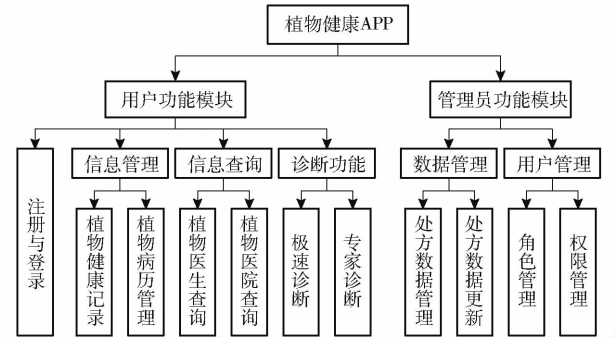

图 9 植物病害诊断系统结构与功能图

Fig. 9 Functional and structure diagram of plant disease diagnosis system

3.2.1 用户功能模块

用户功能模块主要包括注册与登录、信息管理（植物健康记录和植物病历管理）、信息查询（植物医生查询和植物医院查询）、诊断功能（极速诊断和专家诊断），图 10 分别是“植物健康”APP 的首页、极速诊断农户填写信息界面和极速诊断结果界面。农户可登录 APP 首页选择诊断方式，本研究主要应用于极速诊断，对于缺乏病害知识和经验的农户，通过简单填写病害相关信息，可以及时获得基于

(a) 首页界面 (b) 问诊界面 (c) 诊断结果界面

图 10 “植物健康”APP 手机端系统相关功能界面

Fig. 10 Related function interfaces of “Plant Health” APP mobile terminal system

LightGBM 和处方数据的病害诊断模型提供准确的诊断结果。对于处方数据库中不存在的特殊疾病特征，此 APP 还提供在线专家解答，农户与专家的问答结果和处方过程中产生的数据也将作为快速诊断新的处方数据库。

3.2.2　管理员功能模块

管理员功能模块包括数据管理（处方数据管理和处方数据更新）和用户管理（角色管理和权限管理）等基本功能。

4　结论

（1）以 RFECV 和 GDBT 结合的方法对番茄病害的关键特征进行选择，最终原始的 50 个特征简化为 32 个特征，作为番茄病害诊断模型的输入。实验结果表明特征选择前后每个模型运行时间平均降低 20.37%，其中基于贝叶斯优化的 LightGBM 模型在保证准确率的基础上运行时间降低了 47.73%。

（2）相比于 KNN、DT、SVM、RF、GDBT、AdaBoost、XGBoost 7 种常见机器学习方法，构建的基于 LightGBM 的番茄病害诊断模型分类性能最佳，平均高于其它模型 3.65 个百分点，综合准确率达到 89.34%，对于番茄病毒病诊断效果最佳，精确率和召回率分别达到 97.27% 和 94.80%，同时 F1 值可达到 96.02%。最后通过番茄叶霉病和番茄早疫病 2 种番茄病害验证了该模型的实用性和泛化能力。本研究提出的模型可以满足农户对于番茄病害诊断的实际需求。

（3）结合基于 LightGBM 的番茄病害诊断模型，本研究构建了“植物健康”APP 手机端系统，可视化展现了作物病害处方数据挖掘及诊断的实际应用场景，为实现基于处方数据的高效番茄病害诊断提供了新方法。

参　考　文　献

[1] 杨剑锋，乔佩蕊，李永梅，等. 机器学习分类问题及算法研究综述[J]. 统计与决策，2019，35(6)：36－40.
YANG Jianfeng，QIAO Peirui，LI Yongmei，et al. A review of machine-learning classification and algorithms[J]. Statistics & Decision，2019，35(6)：36－40.（in Chinese）

[2] MAHADEVAN A，AROCK M. A class imbalance-aware review rating prediction using hybrid sampling and ensemble learning [J]. Multimedia Tools and Applications，2021，80：6911－6938.

[3] DONG X，YU Z，CAO W，et al. A survey on ensemble learning[J]. Frontiers of Computer Science，2020，14(2)：241－258.

[4] 李诒靖，郭海湘，李亚楠，等. 一种基于 Boosting 的集成学习算法在不均衡数据中的分类[J]. 系统工程理论与实践，2016，36(1)：189－199.
LI Yijing，GUO Haixiang，LI Ya'nan，et al. A Boosting based ensemble learning algorithm in imbalanced data classification [J]. Systems Engineering Theory and Practice，2016，36(1)：189－199.（in Chinese）

[5] MENG Q. LightGBM：a highly efficient gradient Boosting decision tree[C]// Neural Information Processing Systems，2017.

[6] MA X，SHA J，WANG D，et al. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning [J]. Electronic Commerce Research and Applications，2018，31：24－39.

[7] XIE Y，HUANG H，HE X，et al. Using hyperparameter Bayes optimized LightGBM for frequency prediction of auto insurance [J]. Journal of Nonlinear and Convex Analysis，2021，22(10)：2139－2153.

[8] 顾桐，许国良，李万林，等. 基于集成 LightGBM 和贝叶斯优化策略的房价智能评估模型[J]. 计算机应用，2020，40(9)：2762－2767.
GU Tong，XU Guoliang，LI Wanlin，et al. Intelligent house price evaluation model based on ensemble LightGBM and Bayesian optimization strategy[J]. Journal of Computer Applications，2020，40(9)：2762－2767.（in Chinese）

[9] 尹航，吕佳威，陈耀聪，等. 基于 LightGBM－SSA－ELM 的新疆羊舍 CO_2 浓度预测[J]. 农业机械学报，2022，53(1)：261－270.
YIN Hang，LÜ Jiawei，CHEN Yaocong，et al. Prediction of CO_2 concentration in Xinjiang breeding environment of mutton sheep based on LightGBM－SSA－ELM[J]. Transactions of the Chinese Society for Agricultural Machinery，2022，53(1)：261－270.（in Chinese）

[10] 李磊，许敏，王美琴. 生防菌解淀粉芽胞杆菌 Ht－q6 可湿性粉剂的研制及对番茄病害的田间防效[J]. 中国生物防治学报，2018，34(5)：738－745.
LI Lei，XU Min，WANG Meiqin. Preparation of biocontrol agent *Bacillus amyloliquefaciens* Ht－g6 wettable powder and the field control against tomato diseases[J]. Chinese Journal of Biological Control，2018，34(5)：738－745.（in Chinese）

[11] 张燕，田国英，杨英茹，等. 基于 SVM 的设施番茄早疫病在线识别方法研究[J]. 农业机械学报，2021，52(增刊)：125－133，206.
ZHANG Yan，TIAN Guoying，YANG Yingru，et al. Online detection method of tomato early blight disease based on SVM[J]. Transactions of the Chinese Society for Agricultural Machinery，2021，52(Supp.)：125－133，206.（in Chinese）

[12] 许召召，李京华，陈同林，等. 融合 SMOTE 与 Filter－Wrapper 的朴素贝叶斯决策树算法及其应用[J]. 计算机科学，2018，45(9)：65－69，74.

XU Zhaozhao, LI Jinghua, CHEN Tonglin, et al. Naive Bayesian decision tree algorithm combining SMOTE and Filter - Wrapper and it's application[J]. Computer Science, 2018, 45(9):65 - 69,74. (in Chinese)

[13] 李郅琴, 杜建强, 聂斌, 等. 特征选择方法综述[J]. 计算机工程与应用, 2019, 55(24):10 - 19.
LI Zhiqin, DU Jianqiang, NIE Bin, et al. Summary of feature selection methods[J]. Computer Engineering and Applications, 2019, 55(24):10 - 19. (in Chinese)

[14] CAI J, LUO J, WANG S, et al. Feature selection in machine learning: a new perspective[J]. Neurocomputing (Amsterdam), 2018, 300:70 - 79.

[15] CHRYSOSTOMOU K, CHEN S, LIU X. Combining multiple classifiers for Wrapper feature selection[J]. International Journal of Data Mining, Modelling and Management, 2008, 1(1):91 - 102.

[16] FRIEDMAN J H. Greedy function approximation: a gradient Boosting machine[J]. Annals of Statistics, 2001, 29(5):1189 - 1232.

[17] RYAN E G, DROVANDI C C, MCGREE J M, et al. A review of modern computational algorithms for Bayesian optimal design[J]. International Statistical Review, 2016, 84(1):128 - 154.

[18] 崔佳旭, 杨博. 贝叶斯优化方法和应用综述[J]. 软件学报, 2018, 29(10):3068 - 3090.
CUI Jiaxu, YANG Bo. Survey on Bayesian optimization methodology and applications[J]. Journal of Software, 2018, 29(10):3068 - 3090. (in Chinese)

[19] 崔建双, 车梦然. 基于多分类支持向量机的优化算法智能推荐系统与实证分析[J]. 计算机工程与科学, 2019, 41(1):153 - 160.
CUI Jianshuang, CHE Mengran. An intelligent recommendation system for optimization algorithms based on multi-classification support vector machine and its empirical analysis[J]. Computer Engineering & Science, 2019, 41(1):153 - 160. (in Chinese)

[20] 叶海建, 郎睿. 基于 Android 的自然背景下黄瓜霜霉病定量诊断系统[J]. 农业机械学报, 2017, 48(3):24 - 29.
YE Haijian, LANG Rui. Cucumber downy mildew severity quantifying diagnosis system suitable for natural backgrounds based on Android[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(3):24 - 29. (in Chinese)

(上接第 276 页)

[29] 周瑜,刘俊涛,白翔. 形状匹配方法研究与展望[J]. 自动化学报,2012,38(6):889 - 910.
ZHOU Yu, LIU Juntao, BAI Xiang. Research and perspective on shape matching[J]. Acta Automatica Sinica, 2012,38(6):889 - 910. (in Chinese)

[30] 张堂辉,田捷,何余良,等. 基于相似度直方图的混合指纹匹配方法[J]. 计算机学报,2005,28(10):1728 - 1733.
ZHANG Tanghui, TIAN Jie, HE Yuliang, et al. A combined fingerprint matching algorithm based on similarity histogram[J]. Acta Automatica Sinica, 2005,28(10):1728 - 1733. (in Chinese)

[31] 胡新荣,欧阳祥,刘军平. 基于改进哈希算法的图像相似度计算[J]. 武汉纺织大学学报,2020,33(2):74 - 80.
HU Xinrong, OUYANG Xiang, LIU Junping. Image similarity calculation based on improved Hash algorithm[J]. Journal of Wuhan Textile University, 2020,33(2):74 - 80. (in Chinese)

[32] 邢致恺,贾鹤鸣,宋文龙. 基于莱维飞行樽海鞘群优化算法的多阈值图像分割[J]. 自动化学报,2021,47(2):363 - 377.
XING Zhikai, JIA Heming, SONG Wenlong. Levy flight trajectory-based salp swarm algorithm for multilevel thresholding image segmentation[J]. Acta Automatica Sinica, 2021,47(2):363 - 377. (in Chinese)

[33] 赵凤,郑月,刘汉强, 等. 多种群联合的多目标进化自适应阈值图像分割算法[J]. 计算机应用研究,2018,35(6):1858 - 1862.
ZHAO Feng, ZHENG Yue, LIU Hanqiang, et al. Multi-population cooperation-based multi-objective evolutionary algorithm for adaptive thresholding image segmentation[J]. Application Research of Computers, 2018,35(6):1858 - 1862. (in Chinese)

[34] 刘媛媛,王跃勇,于海业,等. 基于多阈值图像分割算法的秸秆覆盖率检测[J]. 农业机械学报,2018,49(12):27 - 35,55.
LIU Yuanyuan, WANG Yueyong, YU Haiye, et al. Detection of straw coverage rate based on multi-threshold image segmentation algorithm[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018,49(12):27 - 35,55. (in Chinese)

[35] 石坤泉,魏文国. 采用双边滤波的冷轧铝板表面缺陷图像去噪方法的研究[J]. 表面技术, 2018,47(9):317 - 323.
SHI Kunquan, WEI Wenguo. Image denoising method of surface defect on cold rolled aluminum sheet by bilateral filtering[J]. Surface Technology, 2018,47(9):317 - 323. (in Chinese)

[36] KINGMA D, BA J. Adam: a method for stochastic optimization[J]. arXiv:1412.6980,2014.

[37] HUANG H, XU H H, WANG X H, et al. Maximum F1-score discriminative training criterion for automatic mispronunciation detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(4): 787 - 797.