基于 Adaboost 模型的江苏省冬小麦产量预测

张顺航　张凤航　李金泽
（山东农业大学信息科学与工程学院，山东泰安 271000）

摘要　农作物产量预测是农业科学的一个重要问题，而气象特征的变化将对农作物的产量产生影响。本文根据 1981—2016 年江苏省的气象数据，研究影响农作物产量的关键气象特征，并利用机器学习中的 Adaboost 算法，对江苏省近几年的小麦产量进行预测。结果表明 Adaboost 模型预测准确率较高，从而为农业生产中的正确决策提供参考。

关键词　冬小麦产量预测；气象特征；机器学习；Adaboost 模型；江苏省
中图分类号　S127　　**文献标识码**　A　　**文章编号**　1007-5739(2019)12-0248-02

我国是农业大国，粮食产量连年增长，农业生产量和生产方式都进入到了新的阶段，农业生产新要求也随之产生。在农业生产中，注重农产品品质、关注农业灾害预警能力等已迫在眉睫[1-3]。2009 年曾有专家在联合国做出预测，到 2050 年世界粮食产量可能需要翻一倍才足以养活全球人口。而在众多影响农作物产量的因素中，气候变化对农业产生直接影响[4]。因此，通过气象因素预测农作物产量尤为关键。但由于农作物在不同生长阶段对气象因素有着不同的要求，使得建立气象因素与农作物产量的关系模型较为困难[5]。此外，虽然目前已有较多时间序列模型，如灰色模型预测、最小二乘法、BP 神经网络、高斯过程，但在面对实际而复杂的分类问题时效果并不理想[6]。

1　Adaboost 模型

Adaboost 算法是 boosting 算法的一种，属于一种迭代算法。它的核心思想是用同一个训练集训练不同的弱分类器，然后把这些弱分类器联合起来形成强分类器，有效地解决了单个分类器面对复杂问题时精度不足的问题。通过江苏省历年冬小麦的产量和江苏省历年的气象因素构建 Adaboost 模型，以平均温度、降水量、平均湿度、日照时数、有效积温等气象因素作为输入，通过特征提取筛选出对产量影响较大的特征向量，利用交叉验证方法评估模型，然后进行参数调优，最终实现较高精度的冬小麦产量预测[7]，总流程见图 1。

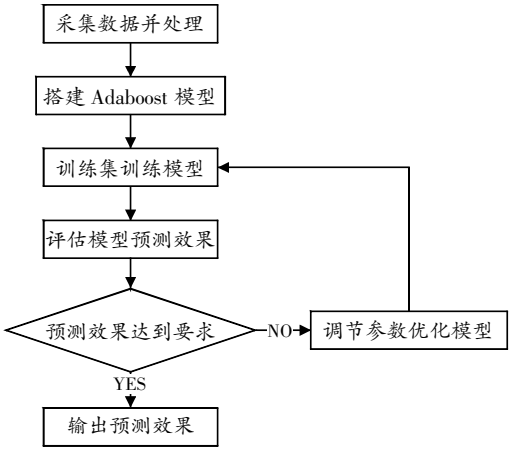

图 1　Adaboost 模型总流程图

2　实例分析

2.1　数据准备

预测农作物产量是实现精细农业的重要措施之一，而影响冬小麦产量的主要气象因素有温度、湿度、降水量、日照时数、有效积温等因素。

为了获得江苏省历年气象数据，收集了包括江苏省南京、无锡、淮安、常州等在内的 23 个站点 1981—2016 年共 36 年每天的气象数据，包括平均温度、降水量、辐射量、日照、平均湿度、平均风速等数据。然后将这些站点的各类气象数据取月平均值，再取所有站点的月平均值为江苏省每类气象数据的月平均值。根据相关资料可知，江苏省冬小麦从 10 月开始播种，翌年 5 月收获[8-11]，所以选取这几个月份的气象数据作为训练集的输入集。

从国家统计局官网上查找江苏省冬小麦近 30 年的产量数据作线性拟合。从图 2 可以看出，圆点表示实际产量值，实线为拟合曲线。用(实际产量-拟合结果)/拟合结果，如果结果大于 0 表示这一年增产(用 1 表示)，如果结果小于 0 则表示这一年减产(用 0 表示)，据此得到训练集的因变量。

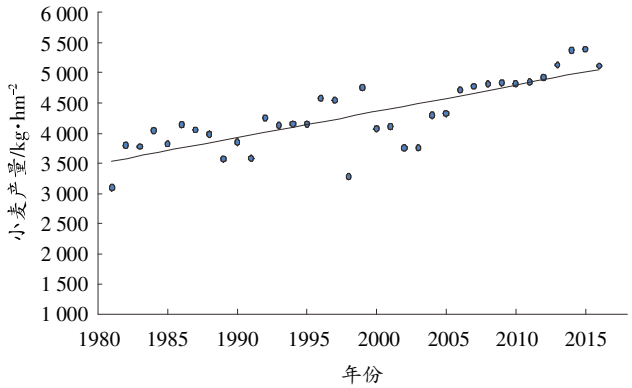

图 2　产量线性拟合图

2.2　建立 Adaboost 模型

AdaBoost 是基于加性模型的算法，即基学习器的线性组合 $H(x)=\Sigma\alpha_t h_t(x)$，其中，$\alpha_t$ 为每个基学习器的权值，$h_t(x)$ 为每个基学习器的预测结果。建立过程主要分为 3 个部分：指数损失函数 $l_{exp}(H|D)$ 的降低、基学习器的权值 α_t 的更新和训练集样本分布的更新和训练集样本分布的更新和训练集样本分布 $D_t(x)$ 的更新。

2.2.1　指数损失函数。若为样本的实际标签值，$H(x)$ 为样本的预测标签值，设 $E_{x\sim D}[e^{-f(x)H(x)}]$ 为样本服从数据集分布 D 时，$e^{-f(x)H(x)}$ 的期望值，则可以表示出指数损失函数，若存在 $H(x)$ 使得损失函数可以最小化，其中 $y\epsilon\{-1,1\}$，则可以求出

收稿日期　2019-03-10

$$\mathrm{sign}(H(x))=\begin{cases}1, P(f(x)=1|x)>P(f(x)=-1|x)\\ -1, P(f(x)=1|x)<P(f(x)=-1|x)\end{cases}$$
$$=arg\ max\ P(f(x)=y|x)$$

这证明指数损失函数是分类任务 0/1 损失函数的替代函数。

2.2.2 基学习器权值 α_t 的更新。h_t 和 α_t 当基学习器 $h_t(x)$基于分布 D_t 产生后,可求得基学习器的权重 α_t 应使得 $\alpha_t h_t$ 最小化指数损失函数,进行偏导数并置零,得到

$$\frac{\partial l_{\exp}(\alpha_t h_t|D_t)}{\partial \varepsilon_t}=-e^{-\alpha t}(1-\varepsilon_t)+e^{\alpha t}\varepsilon_t=0;$$

$$\alpha_t=\frac{1}{2}1n\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)。$$

这样就得到了基学习器的权值更新公式。

2.2.3 训练集样本分布 $D_t(x)$的更新。获得基学习器 $h_{t-1}(x)$后,样本分布将进行调整,可以得到理想的基学习器,其中 $h_t(x)$将在分布 $D_t(x)$下最小化分类误差,因而 $h_t(x)$应该基于分布 $D_t(x)$来训练。

2.3 特征选择与参数调节

用训练集进行特征选择,选择 6 个对江苏省冬小麦产量影响最大的气象因素,分别为 3 月份降水量、1 月份平均湿度、3 月份平均湿度、1 月份平均温度、5 月份平均温度和 3 月份日照时数。然后是参数调节,Adaboost 的参数主要有 3 个,基分类器循环次数 n_estimators、学习速度 learning_rate 和模型提升准则 algorithm。基分类器循环次数过多,模型容易过拟合;循环次数过少,模型容易欠拟合。学习速度如果过大,则容易错过最优值;如果过小,则收敛速度会很慢。模型提升准则有 2 种方式 SAMME 和 SAMME.R,前者是对样本集预测错误的概率进行划分,后者是对样本集的错分率进行划分。先用模型的默认参数对冬小麦产量进行预测,评估预测效果,然后采用网格搜索法对模型进行调参。最终调整的参数为 n_estimators=40,learning_rate=0.8,algorithm='SAMME.R'。

2.4 结果分析

采用留一检验法训练模型并得到最终的预测结果,36 年中预测正确的年份多达 30 年,正确率为 83.3%。根据表 1 可知,在 2005—2016 年的 12 年时间里,预测错误的年份仅有 2 年,且 2010 年之后的预测值全部正确,说明该 Adaboost 模型对冬小麦的产量预测精度较高,尤其是对近几年小麦的增产减产情况全部预测正确。

表 1 2005—2016 年Adaboost模型预测结果

年份	预测值	真实值	预测值是否正确
2005	0	0	正确
2006	1	1	正确
2007	0	1	错误
2008	1	1	正确
2009	1	1	正确
2010	0	1	错误
2011	1	1	正确
2012	0	0	正确
2013	1	1	正确
2014	1	1	正确
2015	1	1	正确
2016	1	1	正确

3 结语

近几年,虽然机器学习成为了最热门的技术之一,但实际上由于机器学习需要大规模的训练集训练,所以实际应用范围有限。尤其是对于农业领域来说,可供使用的数据非常有限,只能用小数据集进行研究。常规的方法在用小数据集做预测与分类时精度低且结果不稳定,而本文所用的 Adaboost 算法将弱分类器联合起来形成强分类器,有效地解决了单个分类器面对复杂问题时精度不足的问题,在实际应用中可行度较高。

4 参考文献

[1] 江显群,陈武奋.BP 神经网络与 GA-BP 农作物需水量预测模型对比[J].排灌机械工程学报,2018,36(8):762-766.
[2] 王晓喆,延军平,张立伟.河南省气候生产力时空分布及粮食产量预测[J].农业现代化研究,2011,32(2):213-216.
[3] 高蕾.基于 ARIMA 模型的安徽省粮食产量预测研究[J].合肥学院学报(社会科学版),2015,32(5):17-18.
[4] 朱新国,张展羽,祝卓.基于改进型 BP 神经网络马尔科夫模型的区域需水量预测[J].水资源保护,2010,26(2):28-31.
[5] 林绍森,唐永金.几种作物产量预测模型及其特点分析[J].西南科技大学学报(自然科学版),2005,20(3):55-60.
[6] 王兴,刘晶晶,阚苗苗,等.我国主要粮食作物产量预测模型及分布特征分析[J].长江大学学报(自科版),2014,11(4):76-79.
[7] 宰松梅,郭冬冬,温季,等.作物产量预测的 BP 神经网络模型研究[J].人民黄河,2010,32(9):71-72.
[8] 商兆堂,张旭晖,商舜,等.江苏省冬小麦生产潜力气候变化趋势评估[J].江苏农业科学,2018,46(12):245-249.
[9] 陈夏.江苏省冬小麦模型模拟优化研究及应用[D].南京:南京信息工程大学,2017.
[10] 姚金保,马鸿翔,张鹏,等.小麦宁麦 26 丰产性、稳产性及适应性分析[J].浙江农业学,2018,59(11):1966-1968.
[11] 王瑞峥,江洪,金佳鑫,等.黄淮海地区冬小麦物候对气候变化的响应及对产量的影响[J].江苏农业科学,2018,46(22):71-75.

(上接第 237 页)

是景、步步有景,着力形成全县“大景区”格局[7],打造全域旅游公园。

4.5 大力发展电子商务,推进乡村产业新时代

实施“互联网+农业”行动计划,持续拓展农村电商服务网点功能,加强农产品上行引导,完善农村电商物流体系建设,支持仓储冷库建设,扶持发展冷链运输,着力提高农业农村电商综合效益。巩固提升电商进农村全覆盖工作,加大绩溪县电商公共服务中心和物流配送中心投入,开展电商特色小镇、示范村创建,引导电商企业与专业合作社、贫困户深度合作,设立网销产品种养基地,加速推进“电商绩溪”创建工作。

5 参考文献

[1] 罗歆.贯彻新发展理念 努力实施乡村振兴战略[J].农村经济与科技,2018,29(11):257-258.
[2] 谭均梅,张源云.茂名市茂南区特色种植存在的问题及建议[J].现代农业科技,2018(17):58.
[3] 杨丽君,李宗阳.景谷县碧安乡高原特色农业发展现状及建议[J].现代农业科技,2017(3):260-261.
[4] 李荣琼,李珂,张小岚,等.昆明市高原特色农业发展现状及对策[J].现代农业科技,2017(5):260-262.
[5] 何勋.区域特色农业旅游发展动力机制研究[J].安徽农业科学,2011,39(28):17432-17435.
[6] 廖东海,张琼,秦桂芳,等.武陵山区特色农业产业发展探讨:以张家界七星椒产业发展为例[J].现代农业科技,2014(14):294-295.
[7] 赵宪军.保定市特色休闲观光农业的发展战略及模式选择[C]//中国农学会.循环农业与新农村建设:2006 年中国农学会学术年会论文集.北京:中国农学会,2006.