东北农业科学 2021,46(5):141-144
Journal of Northeast Agricultural Sciences DOI: 10.16423/j.cnki.1003-8701.2021.05.030

基于云平台的大米产地确证方法研究

王靖会[1],崔 浩[1],程娇娇[1],王艳辉[2],陈 雷[3],王朝辉[4]*
(1. 吉林农业大学信息技术学院,长春 130118;2. 吉林省长春市净月开发区福祉街道办事处,长春 130122;3. 吉林省长春市交警支队南关区大队,长春 130000;4. 吉林农业大学食品工程技术学院,长春 130118)

摘 要:针对大数据背景下地理标志大米产地真伪鉴别的算法模型与实现技术,以大米中矿物质元素含量数据为基础,运用Hadoop分布式集群技术,构建了基于MapReduce的并行化随机森林、支持向量机、人工神经网络与线性判别分析算法模型。结果表明,并行化随机森林模型的判别准确率为97.55%,与相同条件下并行化构建的支持向量机、人工神经网络与线性判别分析模型相比具有更好的产地判别精度,同时依托并行化随机森林模型构建的云平台能获取到较好加速比,不仅能够实现对未知地区大米数据进行准确的产地鉴别,而且能够通过提升数据量或计算节点数,更高效地处理大规模数据。

关键词:随机森林;并行化;MapReduce;加速比;产地确证
中图分类号:S126;TP311 **文献标识码**:A **文章编号**:2096-5877(2021)05-0141-04

Research on Confirmation Method of Rice Producing Area Based on Cloud Platform

WANG Jinghui[1], CUI Hao[1], CHENG Jiaojiao[1], WANG Yanhui[2], CHEN Lei[3], WANG Zhaohui[4]*
(1. *College of Information Technology, Jilin Agricultural University, Changchun* 130118; 2. *Fuzhi Sub-district Office, Jingyue Development Zone, Changchun* 130122; 3. *Changguan City Traffic Police Detachment, Nanguan District Brigade, Jilin Province, Changchun* 130000; 4. *College of Food Engineering and Technology, Jilin Agricultural University, Changchun* 130118,China)

Abstract: Aiming at the arithmetic model and implementation technology of authenticity and falsity identification of geographical indication rice origin under the background of large data, a parallel random forest arithmetic model based on MapReduce was constructed by using hadoop distributed cluster technology on the basis of mineral element content data in rice. The results show that the discriminant accuracy of parallel random forest model is 97.55%. Compared with the linear discriminant analysis model and parallel support vector machine under the same conditions, the parallel random forest model has better discriminant accuracy. At the same time, the parallel random forest model has a growing acceleration ratio, which can achieve rapid and accurate identification of unknown area data.

Key words: Random forest; Parallelization; MapReduce; Speedup ratio; Origin confirmation

近年来,国内掺假大米、勾兑大米问题突出,尤其是针对具有地理标志保护的大米真实性问题引起了社会的广泛关注。美国、加拿大、欧盟以及日本等国家相继建立了国家动物识别系统,并研究了牛肉、羊肉、家禽、水产品以及谷物类产品的食品可追溯系统[1-7]。农产品质量安全追溯技术在国外虽有较成熟的应用,2013年欧洲"马肉丑闻"事件中暴露出的溯源系统中标签错误、成分替代、产地虚假等食品真实性问题也同样引起了国内外研究人员的广泛关注[8]。

许多专家开始探索农畜产品与其生长区域的空间相关性,研究从源头进行产地确证、判别产品真伪的方法和技术。如Canizoa等[9]应用LDA、SVM、RF化学计量学方法对阿根廷葡萄籽进行地区分类,随机森林分类效果最佳,准确率达到93%。吴玥[10]采用随机森林、支持向量机与线性

收稿日期:2019-03-01
基金项目:吉林省重点科技研发项目(20180201051NY)
作者简介:王靖会(1974-),女,副教授,硕士,研究方向为数据挖掘与人工智能。
通讯作者:王朝辉,男,博士,副教授,E-mail: wzhjlndsp@aliyun.com

判别分析方法对梅河地区与相邻地区大米进行产地鉴别，随机森林算法判别正确率为96%。王靖会等[11]采用反向传播人工神经网络、随机森林与支持向量机方法对柳河与辉南地区大米进行产地鉴别，随机森林算法效果最佳，准确率达到100%。苏亚麟等[12]采用随机森林与支持向量机对南昌大米进行产地鉴别，随机森林算法效果最佳，准确率达到92%。现有的研究表明，随机森林、支持向量机、人工神经网络与线性判别分析被广泛应用于农产品的产地确证方面[9-12]。但是随着稻米产地源数据的大量增加，传统的单机算法在建模及产地预测时缺少足够的计算能力，运算效率急剧下降，很难满足大数据分析的要求。

云计算因其高效的计算能力和大数据存储能力被广泛应用于农业领域。随着农业数据的不断增加，云计算被大量用于数据的收集，保存，挖掘与根据数据进行预测。MapReduce作为云计算中的核心技术，由于具有高扩展性、高容错性与使用简便的特点，被广泛应用在算法的并行操作中[13-14]。本文旨在探讨不同并行化算法在产地确证中的预测性能及运行效率，选择随机森林、支持向量机、人工神经网络与线性判别分析四种机器学习算法，利用云计算技术中的MapReduce构建并行化大米产地确证模型，并通过实验比较，筛选最优的产地确证方法，测试构建的云平台在处理大规模数据时的运行效率。

1 材料与方法

1.1 数据来源

本文以梅河地区地标大米作为研究对象，柳河、辉南、延边地区的大米作为非梅河地区样本，共实地采集样本214个。为避免不平衡数据对实验结果造成影响，梅河地区样本数为108个，非梅河地区样本数为106个，具体采集情况见表1。

表1 稻米样本采集点分布信息表

梅河	样品数	柳河	样品数	辉南	样品数	延边	样品数
湾龙镇	25	罗通山镇	7	赵家街	10	民俗村	11
海龙镇	29	时家店	11	团林镇	7	朝阳川	10
黑头山镇	20	姜家店	8	兴德镇	10	东光镇	11
曙光镇	34	孤山子镇	10	光明镇	11	-	-

1.2 矿物质元素含量检测

本研究根据我国GB/T 14609-2008、GB/T 5009.91-2003、GB 5009.12-2010的食品检测标准，检测地标大米样品中铜(Cu)、锌(Zn)、铁(Fe)、锰(Mn)、钾(K)、钙(Ca)、钠(Na)、镁(Mg)、铅(Pb)、镉(Cd)10种矿物质元素的含量数据，使用新丰牌JLGJ4.5砻谷机与新丰牌HNMJ3碾米机分别进行稻米的去壳与糙米去糙操作，JXFM110锤式旋风磨进行大米样品的研磨操作，AA-6300原子吸收分光光度计进行样品的矿物质元素检测。其中石墨炉原子吸收分光光度法用于检测样品中铅(Pb)和镉(Cd)的含量，火焰原子吸收分光光度法用于检测样品中其他8种矿物质元素的含量。

1.3 数据库的建立

HDFS分布式文件系统用于存储基本数据，跟传统的数据库不同，HDFS分布式文件系统不需要对所有数据进行串行读取，而是并行读取存储在不同计算节点中的数据，提高了数据的读取速率。本文在HDFS分布式文件系统架构上，将经过标准化数据预处理的10种矿物质元素(Cu、Zn、Fe、Mn、K、Ca、Na、Mg、Pb、Cd)含量数据作为输入变量，地区(kind)属性作为输出变量，按7∶3比例划分训练集与测试集，建立产地确证数据库。

1.4 机器学习算法

(1)随机森林算法：是由多个决策树组成的分类器。以C4.5算法为例，依据各个属性的信息增益率构建多个决策树，形成随机森林模型。当有新样本输入到模型中时，样本数据的最终分类结果由投票操作最终决定。由于随机森林算法具有防止过拟合与较高的抗干扰能力，因此被广泛应用于农产品的产地确证中。(2)支持向量机：主要思想是通过核函数在特征空间上寻找间隔最大的最优超平面，用以进行空间样本点的分类。由于支持向量机可以有效避免分类模型出现过度适应的问题，因此可用于稻米的产地鉴别。(3)人工神经网络：是一种模仿生物神经网络结构和功能的数学或计算模型。它由大量神经元和节点相互连接组成，每组节点通过一定的权重值连接，最终根据不同的连接方式形成不同的神经网络形式，是一种非线性监督学习算法，被广泛应用于分类领域。(4)线性判别分析：是一种经典的线性学习方法，常用于二分类问题，其原理是根据给定的

训练集样本数据,依据一定方法将训练集数据投射到同一条直线上,尽可能地集中相同类别样本的投影点,并且分散不同样本的投影点。当新的样本数据进行分类时,将样本数据投射到相同的直线上,依据新样本对应的投影点位置进行分类,确定样本的类别。

1.5　云计算平台与分类模型

云计算平台采用3台4核16 G内存,规格为Ecs.cm4.xlarge的阿里云服务器构建Hadoop集群,操作系统均为CenOS 7.4,Hadoop版本为Hadoop 2.7.4,JDK版本为JDK 1.8,Tomcat版本为Tomcat 7.0,Mysql版本为Mysql 5.1,同时采用MapReduce技术并行构建随机森林模型、人工神经网络模型、支持向量机模型与线性判别分析模型。

1.6　模型评估方法

混淆矩阵通过矩阵的方式,描述样本在模型中的预测结果与真实类别之间的关系,常用于评估模型的预测及泛化能力。矩阵结果主要由真正例(True Positive,TP)、假正例(False Positive,FP)、真反例(True Negative,TN)、假反例(False Negative,FN)四部分组成,用以计算准确率、灵敏度与特异度三个性能评价指标,如公式(1~3)所示。本研究中混淆矩阵的设置如表2所示。

$$准确率 = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \quad \cdots\cdots (1)$$

$$灵敏度 = \frac{TP}{TP + FN} \times 100\% \quad \cdots\cdots (2)$$

$$特异度 = \frac{TN}{FP + TN} \times 100\% \quad \cdots\cdots (3)$$

表2　梅河和非梅河大米产区分类的混淆矩阵

分类	预测结果	
	梅河	非梅河
梅河	真正例(TP)	假反例(FN)
非梅河	假正例(FP)	真反例(TN)

2　结果与讨论

2.1　并行化分类算法的大米产地确证模型预测能力评估

本文采用MapReduce构建的并行化分类模型调优后参数分别为:随机森林模型的分类树个数值为30,特征数值为4,离散化参数为10;支持向量机模型的gamma值为0.625,cost值为1;人工神经网络模型的隐藏层神经元个数为9;线性判别分析模型经过递归特征消除后筛选出最优特征变量子集为(Cu,Ca,Cd,Mg,Mn,Fe,K,Pb,Na)。各模型性能的实验测试结果如表3所示,并行化随机森林模型的预测准确率、特异度、灵敏度均高于其他三种模型,预测能力最佳。

表3　模型精度与泛化能力测试结果　%

并行化模型	准确率	特异度	灵敏度
随机森林	97.55	96.90	98.21
支持向量机	93.56	92.63	95.13
线性判别分析	92.76	91.32	93.25
人工神经网络	87.63	82.23	93.52

2.2　云平台效率评估

平台的计算时间取决于数据通信时间和分类建模时间[15]。本文选择预测性能最好的并行化随机森林模型进行加速比实验,评估云平台的运算效率。在实验过程中通过改变集群的计算节点数与数据集规模大小验证模型的各并行评价指标[16-17]。由于集群资源的限制,本文的加速比实验通过改变3个节点数目测试传统单机串行随机森林算法运行时间与多个节点并行化随机森林算法运行时间的比值。为测试在不同数据量下模型性能的变化,将训练集数据条数扩充为184 W、368 W、552 W作为数据集D1、D2、D3,以此验证模型性能,不同数据集在不同节点中的运行时间结果见表4。

表4　不同数据集在不同节点中的集群运行时间结果

	D1	D2	D3
单机环境(s)	532	2 703	7 103
1个节点(s)	579	2 758	7 256
2个节点(s)	296	772	1 225
3个节点(s)	202	563	973

由加速比实验可知(图1),本文构建的并行化随机森林算法在处理数据集D1时,加速比接近线性加速比,在处理数据集D2与数据集D3时,加速比增长趋势已经超过线性加速比。这是由于数据集D1与D2、D3相比数据规模较小,集群相比

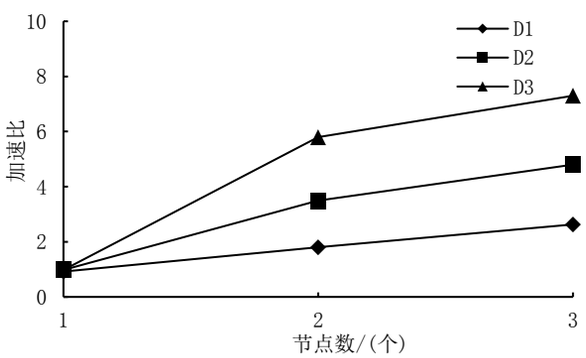

图1　平台加速比实验结果

传统单机需要更多的时间用于任务调度等方面，集群在处理较小规模的数据中计算优势暂不突出，但是加速比始终呈现增长趋势。同时在处理更大规模的数据集D2与D3时，随着数据集和节点数的不断增加，集群的任务调度时间所占比例减少，并行化执行任务比传统单机串行执行任务更少，加速比增长趋势已经远超线性加速比。综上随着数据集和计算节点的增加，传统随机森林算法需要更多的计算资源来降低运算时间，而本文构建的并行化随机森林算法，通过集群较强的计算能力提高自身运算速度，随着数据规模的不断增大，并行化随机森林算法的优势逐渐得到体现，满足了大规模数据量计算的需求。

3 结 论

本文依托并行化分类算法构建了大米产地确证云平台，其中利用MapReduce技术分别并行化构建了随机森林模型、支持向量机模型、人工神经网络模型与线性判别分析模型。实验结果表明，并行化的随机森林模型比其他三种并行化模型具有更好的产地鉴别效果。在加速比实验中，同时在集群和单机环境中分别运行随机森林模型，在集群环境下构建的并行化随机森林算法比传统单机串行的随机森林算法在处理大规模数据时具有更高的效率，优势更加明显。

参考文献：

[1] Cappai M G, Rubiu N G, Pinna W. Economic assessment of a smart traceability system (RFID+DNA) for origin and brand protection of the pork product labelled “suinetto di Sardegna” [J]. Computers and Electronics in Agriculture, 2018, 145: 248-252.

[2] Sheikha A F E, Hu D M. How to trace the geographic origin of mushrooms?[J].Trends in Food Science & Technology, 2018, 78 (AUG): 292-303.

[3] Gautam R, Singh A, K. Karthik, et al. Traceability using RFID and its formulation for a kiwifruit supply chain[J].Computers & Industrial Engineering, 2017, 103(JAN): 46-58.

[4] 关 静.大米质量安全溯源系统的设计与实现[D]. 哈尔滨：东北农业大学，2016.

[5] 张鉴滔.基于WebGIS的农产品产地管理与追溯系统研制[D]. 杭州：浙江大学，2012.

[6] 姜 爽，韩贵清，司震宇，等.第三方稻米溯源平台设计与实现[J]. 农业工程学报，2017，33(24)：215-221.

[7] 董 雪.有机蔬菜质量控制及可追溯体系研究综述[J]. 吉林农业科学，2010，35(3)：51-56.

[8] 张云雪.欧盟肉制品安全监管制度研究[D]. 重庆：西南大学，2014.

[9] Canizoa B V, Escuderoa L B, Péreza M B, et al. Intra-regional classification of grape seeds produced in Mendoza province (Argentina) by multi-elemental analysis and chemometrics tools[J]. Food Chemistry, 2018, 242: 272-278.

[10] 吴 玥.基于机器学习方法的地理标志大米产地确证技术研究[D]. 长春：吉林农业大学，2017.

[11] 王靖会，臧妍宇，曹 崴，等.基于机器学习方法的吉林大米产地确证模型研究[J]. 中国粮油学报，2018，33(9)：123-130.

[12] 苏亚麟，吕开云.基于随机森林算法的特征选择的水稻分类—以南昌市为例[J]. 江西科学，2018，36(1)：161-167.

[13] 张 鑫.随机森林算法的优化研究及在文本并行分类上的应用[D]. 南京：南京邮电大学，2018.

[14] 郑凯航.基于大数据技术的随机森林模型并行化设计及实现[D]. 太原：太原理工大学，2017.

[15] 夏吉安，杨余旺，曹宏鑫，等.云计算的蚕豆虫害可见-近红外光谱分类[J]. 光谱学与光谱分析，2018，38(3)：756-760.

[16] 于 延，王建华.基于云计算平台的随机森林算法的研究与实现[J]. 科技通报，2013，29(4)：50-52.

[17] 陈娉婷，官 波，沈祥成，等.大数据时代开放式农业信息知识库构建研究[J]. 东北农业科学，2018，43(5)：60-64.

（责任编辑：王丝语）

（上接第125页）的发展涉及到农业、商业、税收等各个部门，逐步走向统一化管理，各个部门之间要加强协作，建立适应市场经济发展的审批机制；三要发挥政府的调控职能。财政协调上，政府建立农业投入约束机制，调动各方资金支持苹果产业集群的发展，形成多元化投资体系；信贷协调上，政府要通过对惠农信贷资金的倾斜，促进苹果产业结构调整；税收协调上，政府要制定惠农政策，减轻龙头企业负担，平等纳税。四要发挥政府的监督协调职能，通过建立科学的规章制度，严格按照国家标准，加强对企业和产品的监督检查，保护消费者权益。

参考文献：

[1] 韩振兴，刘宗志，常向阳.山西省特色农业产业集群集中度和竞争力分析—以运城苹果、朔州羊肉、晋城大豆为例[J]. 中国农业资源与区划，2018，39(11)：99-109.

[2] 郝曦煜，梁 杰，郭文云，等.白城市特色食用豆产业发展优势分析[J]. 东北农业科学，2019，44(1)：87-90.

[3] 黄福江，高志刚.基于“钻石模型”的荷兰花卉产业集群要素分析及经验启示[J]. 世界农业，2016(2)：12-15.

[4] 刘丽娜.基于钻石模型的中国水产品出口竞争力分析—以福建省为例[J]. 世界农业，2017(6)：150-157.

[5] 杨建斌.陕西省西安市临潼区石榴产业发展研究[J]. 东北农业科学，2018，43(3)：47-51.

[6] 张 慧.汉中市城固县柑桔产业发展研究[J]. 东北农业科学，2018，43(6)：58-60.

（责任编辑：王丝语）