DOI:10.13652/j.issn.1003-5788.2020.05.014

基于电子鼻与 LightGBM 算法判别葡萄酒品种的研究

Research on discriminating wine varieties based on electronic nose and LightGBM algorithm

乔 森 张 磊 母芳林

QIAO Miao ZHANG Lei MU Fang-lin

(河北工业大学人工智能与数据科学学院,天津 300130)

(*School of Artificial Intelligence, Hebei University of Technology, Tianjin* 300130, *China*)

摘要:针对葡萄酒的鉴别问题,通过电子鼻采集 7 种葡萄酒的气味信息,应用 LightGBM 算法对葡萄酒的气味特征进行学习,并运用 TPE 超参数优化算法对 LightGBM 算法超参数进行自适应寻优,以 5 折交叉验证为指标评估模型的性能。试验结果表明 LightGBM 建立的判别模型对葡萄酒样本的判别准确率为 96.62%,优于传统的支持向量机、随机森林、神经网络,验证了 LightGBM 在葡萄酒品种鉴别中的优越性。

关键词:葡萄酒;电子鼻;LightGBM;TPE

Abstract: Aiming at the problem of wine identification, the odor information of 7 kinds of wine was collected through the electronic nose, the LightGBM algorithm was used to learn the odor characteristics of the wine, and the TPE hyperparameter optimization algorithm is used to adaptively optimize the HyperGB parameter of the LightGBM algorithm. Verification is an indicator to evaluate the performance of the model. The experimental results showed that the discrimination model established by LightGBM had a 96.62% accuracy rate for wine samples, which was superior to traditional support vector machines, random forests, and neural networks. It verifies the superiority of LightGBM in wine variety identification and provides wine identification a fast, reliable and effective analysis method is also suggested, and more excellent algorithms can be introduced into the field of wine smell data mining machines.

Keywords: wine; electronic nose; light gradient boosting machine (LightGBM); tree parzen estimator (TPE)

葡萄酒是一种极具风格和个性化的酒精饮料,不同产地、不同年份、不同品种和不同工艺条件的葡萄酒均具有显著的特征[1],其中葡萄品种是决定葡萄酒品质的重要因素。目前,鉴别不同品种的葡萄酒的方式主要还是利用品评专家的感官鉴定[2]来实现,受到个人经验和条件的限制;而现有的仪器分析[3]、理化分析[4]等只能从某一或者某几个侧面反映葡萄酒的品质。

随着电子鼻技术的发展,很多研究人员开始利用这一技术对葡萄酒品质特征进行快速判别。张振等[5]利用表面声波型电子鼻对不同年份的黄酒样品进行采样,并利用主成分分析法和典型判别分析对气体数据进行分析,成功区分了 4 种酒龄黄酒样品。许春华等[6]利用电子鼻指纹分析系统对张裕干白和长城干红的气味进行鉴别,并采用主成分分析和线性判别分析法对传感器响应信号进行分析,实现了对葡萄酒的风味评价。刘奕彤等[7]利用电子鼻检测技术有效地鉴别了西拉、马瑟兰和美乐 3 种品种干红葡萄酒的香气差异。宫雪[8]利用电子鼻对不同葡萄品种酿造葡萄酒进行检测,结合主成分分析方法和线性判别分析探索电子鼻的识别能力,结果显示,电子鼻能很好地识别与分区葡萄酒的品种。

LightGBM 是一种集成学习算法,具有较优的数据分类能力,不易过拟合,在食品安全[9]、信用评级[10]、电力评估[11]、疾病预测[12]等方面可实现快速准确的判别,但目前尚未见其在葡萄酒品种鉴别中的相关报道。研究拟提出一种 LightGBM 结合电子鼻检测的葡萄酒品种快速、准确识别方法,以期为葡萄酒检测引入性能优异的算法。

1 材料与方法

1.1 材料与仪器

1.1.1 试验材料

赤霞珠、马瑟兰、西拉、梅洛、蛇龙珠、佳美、品丽珠 7 个品种干红葡萄酒样品:华夏产区 2018 年产的原酒,每

作者简介:乔森,女,河北工业大学在读硕士研究生。

通信作者:张磊(1977—),男,河北工业大学教授,博士。

E-mail:zhanglei@hebut.edu.cn

收稿日期:2020-02-20

种样品 100 瓶,中粮华夏长城葡萄酒有限公司。

1.1.2　主要仪器

便携式电子鼻:PEN3 型,由 10 个金属氧化物气体传感器矩阵(如表 1 所示)、气体采集装置和信号处理单元组成,德国 Airsense 公司。

表 1　PEN3 传感器名称与性能描述

Table 1　Sensor names and performance descriptions for PEN3

序号	传感器名称	性能描述
A	W1C	芳香成分,苯类敏感
B	W5S	灵敏度大,对氮氧化合物灵敏
C	W3C	对氨类、芳香成分灵敏
D	W6S	对氢化物有选择性
E	W5C	对短链烷氢、芳香成分灵敏
F	W1S	对甲烷类灵敏
G	W1W	对硫化物灵敏,对烃和硫的有机成分较灵敏
H	W2S	对醇类、醛酮类灵敏
I	W2W	对芳香成分、有机硫化物灵敏
J	W3S	对长链烷怪灵敏

1.2　试验方法

1.2.1　试验环境控制　室内温度 22～25 ℃,湿度 50%～55%。用移液器取每个酒样 300 mL 并将酒样装于 500 mL烧杯中,用保鲜膜密封,并使其与小瓶中的空气静置平衡 10 min,使样品气体能充分挥发在密闭烧杯中,待气体达到饱和平稳状态后进行正式试验。

1.2.2　电子鼻采样　采用直接顶空吸气法,气体采集前以 300 mL/min 的速率吸取经由活性炭处理的洁净空气,对电子鼻的气室和气道进行清洗,清洗时间为 60 s;检测时,将进气针与补气针同时插入保鲜膜密封的烧杯中,电子鼻内置气泵开始工作,以 300 mL/min 的速率吸取样品气体,采集间隔时间 1 s,采样时间为 90 s;为避免试验过程中人为操作造成的偶然性误差,确保样品的准确性与可靠性,对同一样品进行 3 次重复试验。每次采集后的气体信息以文本方式保存到计算机内,以便进行后续的数据分析处理。

1.3　建模方法

1.3.1　LightGBM 算法　LightGBM 算法是一种基于 GBDT 的数据模型,是将弱学习器组合成强大的学习器的集成学习算法[13]。算法中使用回归树作为弱学习器,通过使用每个预测结果与目标值的残差作为下一个学习的目标,获得当前残差回归树,每个树都学习所有先前树的结论与残差,将多个决策树的结果加在一起作为最终预测输出。利用直方图算法对特征进行预排序,并利用节点展开方式进行树的构建,是一种高效、高精度、高性能的分类算法。

1.3.2　支持向量机　支持向量机(SVM)是在分类分析中的监督式演算法,利用分离超平面将两种或多种类别资料做区分[14]。当资料为线性可分时,支持向量机透过决策平面将不同类别资料进行区分,资料与决策平面的距离成为边界,距离越大越能够明确的区分资料。面对非线性的分类问题时,先计算每个资料与决策边界的最小距离,再将所有的距离加总求最大值,得到区分线为分离超平面。

1.3.3　随机森林　随机森林(RF)是以决策树为元分类器,通过随机方式建立"森林"对样品进行训练并预测的一种分类器[15]。使用拔靴法将数据随机进行取后放回的动作,在数据取出后使用特征袋法随机选取训练数据集特征来生成决策树,重复这样的动作建立出每棵独立的决策树,最后对多颗决策树进行投票对分类结果进行评断。

1.3.4　BP 神经网络　神经网络是由人工神经元所组成,以人工神经元来模仿生物神经元的功能,再由人工神经连接成网络,进而达到模仿生物神经网络的目的[16]。在多层神经网络中,由于隐藏层没有理想输出值,只能透过计算最后一个隐藏层中的误差来估计上一层的理想输出值后来计算上一层的误差,通过这种方式一层一层的反向分析传递到第一层,称之为反向传输神经网络(BPNN)。

1.3.5　TPE 超参数寻优　以 TPE 算法对 LightGBM 超参数进行自适应寻优,假设 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 代表模型中选择的超参数,$\Lambda_1,\Lambda_2,\cdots,\Lambda_n$ 代表每个超参数的选择域;则模型的超参数选择域空间定义为 $\Lambda=\Lambda_1\times\Lambda_2\times\cdots\times\Lambda_n$,假设训练中的损失函数 $L(\cdot)$,当 $\lambda\in\Lambda$ 的超参数使用 k 折交叉验证方法时,超参数的优化问题可以表示为最小化公式:

$$f(\lambda)=\frac{1}{k}\sum_{i=1}^{k}L(\lambda,D^{i}_{\text{train}},D^{i}_{\text{validation}}),\tag{1}$$

式中:

$f(\lambda)$——k 次损失函数的平均值;

k——交叉验证次数;

D^{i}_{train}——在训练集上训练;

$D^{i}_{\text{validation}}$——在验证集上验证;

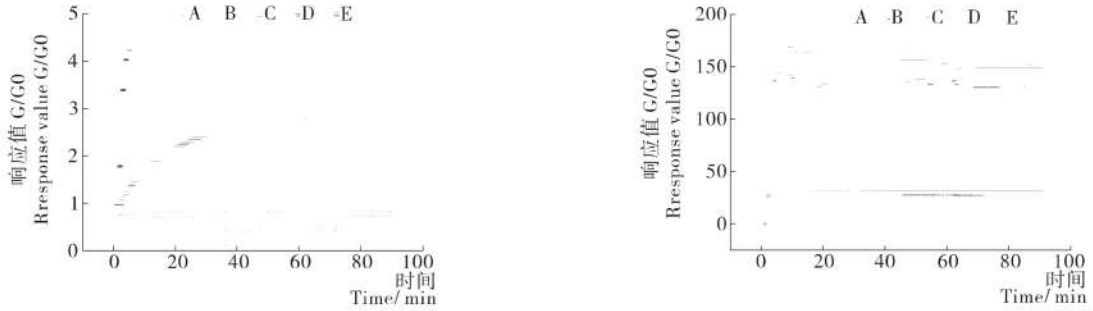
$L(\lambda,D^{i}_{\text{train}},D^{i}_{\text{validation}})$——损失值。

TPE 算法利用概率模型代理复杂优化函数[17],概率模型中引入了待优化目标的先验,模型能有效减少不必要的采样,是考虑历史参数的一种搜索方法。TPE 使用顺序模型全局优化(SMBO)方式进行超参数寻优[18],利用预期改进法(EI)作为优化准则,使用以往的超参数推荐下一次的超参数。

2　不同品种葡萄酒的识别与分析

2.1　电子鼻响应信号曲线

由图 1 可观察到，电子鼻响应值的变化趋势呈现一定的规律，在 90 s 的检测过程中，传感器的响应值先突然升高，偏离原有基线，随着检测时间的延长，传感器的响应值基本达到稳定状态，其中 B、F、G、H、I 5 个传感器对葡萄酒气味响应明显，G、F 响应值更是高于 150，表明葡萄酒中存在甲烷类、烃和硫的有机成分。其他 5 种传感器响应值都在 5 以下，没有变化或者变化不明显。通过观察响应曲线，电子鼻设备能对葡萄酒进行检测，但想要对每种品种进行建模分析，需要对数据进行进一步的处理。



（a）响应值较低的传感器　　（b）响应值较高的传感器

图 1　传感器响应图

Figure 1　Sensor response diagram

通过二维多项式拟合传感器响应曲线，其表达式：

$$y = A_0 + A_1 x + A_2 x^2, \tag{2}$$

式中：

y ——传感器吸附过程的响应值；

A_0、A_1、A_2——多项式曲线拟合系数；

x ——传感器吸附时间，s。

采用每条拟合曲线的模型 3 个系数 A_0、A_1、A_2 作为单个传感器特征值。

2.2　LightGBM 判别模型的建立

试验中，共采用到 2 100（7 种品种×100 瓶×3 次平行试验）组葡萄酒的气味信息数据，每组数据具有 30（10 个传感器×3 个特征值）维特征。LightGBM 算法经 Python2.7 实现，采用 TPE 超参数寻优算法对模型超参数进行选择，其中参数表述、取值范围、最终取值情况由表 2 所示。采用 5 折交叉验证方法进行判别准确性评估，将 2 100 组资料分为 5 个子集，每次轮流挑选 1 个子集（420 组）资料作为验证，剩下的 4 个子集（1 680 组）数据作为训练资料，最后将 5 次的资料辨别率取平均作为整体的辨别率。

从表 3 可以看出，LightGBM 方法作为改进的集成算法在葡萄酒的气味数据挖掘中表现出了优秀的判别准确性。在 5 折交叉验证中，每次的判别准确率均高于 95%，并在第 3 次交叉验证中准确率高达 98.10%，提升了最终的平均准确率，并且 5 折交叉验证避免了判别的偶然性与单一性，有力地说明了 LightGBM 模型的适用性。

2.3　不同算法性能的比较

为验证所提的 LightGBM 在葡萄酒鉴别中的分类优越性，选择支持向量机（SVM）、随机森林（RF）、神经网络（BPNN）3 种在电子鼻检测中常用的分类算法进行结果的验证与比较。为保证各算法达到最优的效果，同样采取 TPE 超参数寻优方法对模型进行优化，采用 5 折交叉验证对模型进行分类准确性的判别。

由表 4 可知，4 种算法对葡萄酒鉴别准确率均高于 90%，说明电子鼻结合模式识别能有效地判别葡萄酒中葡萄的品种；LightGBM 算法取得了最高的判别准确率，说明LightGBM模型通过不断拟合前一棵树的误差能有

表 2　超参数信息

Table 2　Hyperparametric information

参数名	取值范围	参数解释	最终超参数值
colsample_by_tree	(0.6,1.0)	构造每棵树时列的子样本比	0.9
learning_rate	(0.005,0.2)	学习率	0.15
max_depth	(1,30)	树的最大深度	24
min_child_samples	(1,15)	子节点中需要的最小数据数量	1
num_leaves	(5, 15)	叶子数	7
reg_alpha	(0,1)	L1 权值正则化项	0.12
reg_lambda	(0,1)	L2 权值正则化项	0.6

表 3　LightGBM 模型的 5 折交叉验证的测试集判别准确率

Table 3　5-fold cross-vaidation method for accuracy of LightGBM model in test set

交叉验证次数	准确率/%	交叉验证次数	准确率/%
第 1 次	96.19	第 4 次	97.38
第 2 次	95.95	第 5 次	95.48
第 3 次	98.10	平均值	96.62

表 4　基于 5 折交叉验证的不同算法测试集准确率

Table 4　5-fold cross-vaidation method for accuracy of different algorithm model in test set　%

交叉验证次数	LightGBM	SVM	RF	BPNN
第 1 次	96.19	92.14	94.48	92.38
第 2 次	95.95	91.19	93.10	91.43
第 3 次	98.10	89.52	94.48	90.24
第 4 次	97.38	90.00	94.05	93.10
第 5 次	95.48	89.29	93.57	92.86
平均准确率	96.62	90.43	93.97	92.00

效提高分类准确率。其次为随机森林算法,说明对于特征值与特征向量进行随机选取构建的"森林"能多气味数据进行较全面的训练与学习,但因没考虑每棵树产生的误差其分类效果劣于 LightGBM。通过比较得知,经典的支持向量机算法和神经网络算法在验证集上的效果相对较差,支持向量机平均判别准确率最低为 90.53%,并且在第 5 次交叉验证中准确率为 89.29%,在 420 个验证集中有 45 个被判别错误,其分类效果不佳。说明支持向量机在对葡萄酒气味信息进行分类时无法寻找到最优的分线性映射函数,无法对多品种的葡萄酒数据构建最优的分类超平面。相较于支持向量机,神经网络展现了较优良的分类效果,在 5 折交叉验证中其分类准确率均高于 90%,并且平均准确率为 92%仅次于随机森林算法,说明误差反向传播的神经网络算法通过不断减小误差能达到较好的分类效果,然而每次训练样本仅为 1 680 个,神经网络无法得到最优的训练,固其分类效果欠佳。

3　结论

利用电子鼻对赤霞珠、马瑟兰、西拉、梅洛、蛇龙珠、佳美、品丽珠 7 种葡萄酒的气味进行采集。通过观察传感器响应曲线提出二次多项式拟合方法对曲线进行拟合,提取多项式 3 个系数作为 90 s 传感器信号的特征值,大大地降低了特征值的维度。然后,提出 LightGBM 算法对不同品种葡萄酒进行区分,并利用 TPE 参数寻优方法对算法进行改进,最后对比支持向量机、随机森林、反向传输神经网络算法的分类效果,结果表明 LightGBM 模型的 5 折交叉验证平均准确率为 96.62%,分类准确度最高,验证了所提算法在葡萄酒品种鉴别中的优越性。

试验探索了电子鼻和 LightGBM 模型在葡萄酒品种检测中的可行性,为提高判别准确率后续将进一步探索电子鼻数据,通过特征选择方法选取更具代表的葡萄酒气味特征对其进行分析。

参考文献

[1] 郑青. 不同陈酿年份、葡萄品种及葡萄产地葡萄酒香气成分的研究[D]. 南昌:南昌大学, 2015: 3-8.

[2] 陶永胜, 彭传涛. 中国霞多丽干白葡萄酒香气特征与成分关联分析[J]. 农业机械学报, 2012, 43(3): 130-139.

[3] 陶永胜, 刘吉彬, 兰圆圆, 等. 人工贵腐葡萄酒香气的仪器分析与感官评价[J]. 农业机械学报, 2016, 47(2): 270-279, 315.

[4] 金新宇, 吴时敏, 黄明泉, 等. 2013—2018 年进口葡萄酒市场及理化指标分析[J]. 粮食与油脂, 2019, 32(10): 77-81.

[5] 张振, 李臻锋, 宋飞虎, 等. 电子鼻结合化学计量法用于检测黄酒酒龄[J]. 食品与机械, 2015, 31(3): 57-61, 118.

[6] 许春华, 肖作兵, 牛云蔚, 等. 电子鼻和电子舌在果酒风味分析中的应用[J]. 食品与发酵工业, 2011, 37(3): 163-167.

[7] 刘弈彤, 刘期成, 李红娟, 等. 烟台产区不同品种干红葡萄酒香气差异分析[J]. 酿酒科技, 2019(8): 40-47.

[8] 宫雪. 电子鼻和电子舌对葡萄酒的感官评价分析研究[D]. 杨凌: 西北农林科技大学, 2014: 17-58.

[9] 高亚男, 王文倩, 王建新. 集成模糊层级划分的 LightGBM 食品安全风险预警模型: 以肉制品为例[J/OL]. 食品科学. [2020-04-01]. http://kns.cnki.net/kcms/detail/11.2206.TS.20200330.1547.077.html.

[10] 马晓君, 沙靖岚, 牛雪琪. 基于 LightGBM 算法的 P2P 项目信用评级模型的设计及应用[J]. 数量经济技术经济研究, 2018, 35(5): 144-160.

[11] 周挺, 杨军, 周强明, 等. 基于改进 LightGBM 的电力系统暂态稳定评估方法[J]. 电网技术, 2019, 43(6): 1 931-1 940.

[12] 张渊, 冯聪, 李开源, 等. ICU 患者急性肾损伤发生风险的 LightGBM 预测模型[J]. 解放军医学院学报, 2019, 40(4): 316-320.

[13] FAN Jun-liang, MA Xin, WU Li-feng, et al. Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data[J]. Agricultural Water Management, 2019, DOI: 10.1016/j.agwat.2019.105758.

[14] LI Qiang, GU Yu. Classification of multiple Chinese liquors by means of a QCM-based E-Nose and MDS-SVM classifier[J]. Sensors, 2017, 17(2): 272-286.

[15] LI Qiang, GU Yu, WANG Nan-fei. Application of random forest classifier by means of a QCM-Based E-Nose in the identification of Chinese liquor flavors[J]. IEEE Sensors Journal, 2017, 17(6): 1 788-1 794.

[16] LIU Hui-xiang, LI Qiang, YAN Bin, et al. Bionic electronic nose based on MOS sensors array and machine learning algorithms used for wine properties detection[J]. Sensors, 2018, 19(1): 1-11.

[17] 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): 3 068-3 090.

[18] 李斌, 王卫星. NCA 萃取和贝叶斯优化调参对分类模型的改进[J]. 计算机应用与软件, 2019, 36(8): 281-287.