近红外光谱融合电子鼻数据对烟叶产地判别研究

汪阳忠[1],张鑫[1],蔡振波[1],黄雯[1],费婷[1],吴达[1],张旭峰[2],孟祥周[2],束茹欣[1]

(1.上海烟草集团有限责任公司 技术中心,上海 200082;2.同济大学 环境科学与工程学院,上海 200092)

摘　要:基于烟叶近红外光谱、Heracles 电子鼻及二者的融合数据,建立了云南、河南、福建和吉林 4 个省份的烟叶产地识别模型以及河南省内漯河、南阳、平顶山、许昌和驻马店 5 个地级市的烟叶产地识别模型.对于地理位置相距比较远的不同省份的烟叶,基于单一数据源就可以建立准确率比较高的产地识别模型.对于河南省内 5 个地级市的烟叶,其地理位置相距近,气候变化小,烟叶相似性高,仅基于单一信息源的数据,该产地识别模型的准确率偏低.为了提高河南省内 5 个地级市烟叶产地识别的准确率,将烟叶近红外光谱数据与 Heracles 电子鼻数据进行融合,由于增加了烟叶数据信息量,这 5 个产地的识别效果明显提升,其留一法内部交叉验证准确率为 98.26%,高于数据融合前单一数据源判别模型的 86.96%.研究表明 Heracles 电子鼻数据可以在不同的数据维度上,对近红外光谱数据进行信息量补充,为烟草品种溯源、质量监测、市场监督等方面提供新思路.

关键词:近红外光谱;Heracles 电子鼻;数据融合;支持向量机

中图分类号:O69　　**文献标志码**:A　　**文章编号**:1000-2367(2024)02-0104-07

烟草产地的准确分类对于烟草行业的质量控制和市场竞争具有重要意义,传统的基于经验和感官评价的分类方法存在主观性和不稳定性等问题,这可能导致分类结果的不准确性和不一致性.为此,近年来研究者基于烟叶近红外光谱(near-infrared spectroscopy,NIR)数据结合机器学习方法建立烟叶产地的快速识别模型.耿莹蕊等[1]基于 NIRS,采用灰狼算法优化参数,最终建立了 8 个烟叶产地的支持向量机算法(support vector machine,SVM)分类模型.鲁梦瑶等[2]基于卷积神经网络对烟叶近红外光谱数据进行处理,针对近红外光谱数据的特点,对卷积神经网络进行改进,建立了东北、黄淮、西南三大烤烟产区识别模型.束茹欣等[3]基于 NIR-PCA-SVM 联用技术建立了云南、河南、安徽、福建、贵州、吉林 6 个省产地识别模型.

在前期烟叶产地分类判别的研究中,由于这些产地属于不同的行政区域,其地理位置距离比较远,气候差异大,因此烟叶本身的差异也比较大,基于近红外光谱数据可以建立准确率比较高的产地识别模型.随着企业实际要求更加严格,生产中越来越关注同一省内不同地级市烟叶产地的识别,但由于这些地级市地理位置比较近,气候差异小,烟叶本身的差异相应地也比较小,仅利用近红外光谱单一数据源建立的地级市烟叶产地识别模型准确率就比较低.可能的原因是近红外光谱数据的信息量不能满足更精准的建模要求,或者是对近红外数据处理的机器学习算法还需改进[4].本文尝试补充更多源的信息数据,建立对于地理位置相距比较近的同一省内不同地级市产地的识别准确率高的模型.近两年来,电子鼻(electronic nose,EN)数据也被引入到烟草行业的快速检测中,并与近红外数据融合,展现出与近红外数据不同维度的信息内容,但相关研

收稿日期:2023-07-25;**修回日期**:2023-08-31.

基金项目:国家自然科学基金(42177378);国家烟草专卖局卷烟烟气重点实验室开放研究基金课题(2021-7).

作者简介:汪阳忠(1989－),男,福建泉州人,上海烟草集团有限责任公司工程师,主要从事烟草化学分析技术研究,E-mail:wangyz@sh.tobacco.com.cn.

通信作者:束茹欣(1974－),男,上海人,上海烟草集团有限责任公司高级工程师,主要从事卷烟配方、烟叶原料质量研究,E-mail:shurx@sh.tobacco.com.cn.

引用本文:汪阳忠,张鑫,蔡振波,等.近红外光谱融合电子鼻数据对烟叶产地判别研究[J].河南师范大学学报(自然科学版),2024,52(2):104-110.(Wang Yangzhong, Zhang Xin, Cai Zhenbo,et al.Classification of tobacco leave parts based on the fusion of near-infrared spectroscopy and Heracles electronic nose data[J].Journal of Henan Normal University(Natural Science Edition),2024,52(2):104-110.DOI:10.16366/j.cnki.1000-2367.2023.07.25.0004.)

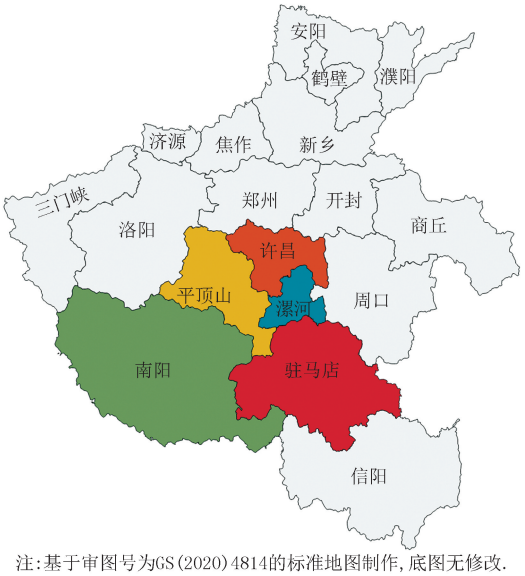
究工作还比较少.王文俊等[5]利用烟叶近红外光谱和电子鼻融合数据建立判别烟叶清香型、中间香型和浓香型 3 种香型风格的模式识别模型，比单一数据模型的准确率提高超过 12%.ZHANG 等[6]在烟叶 NIR 和 EN 数据融合的基础上，通过遗传算法选择出了建模变量，再利用支持向量机算法建立烟叶年份的分类模型，准确率提高也超过 10%.

为了建立准确率比较高的同一省内不同地级市烟叶产地的识别模型，本文尝试基于烟叶 NIR 和 EN 数据融合进行建模.为此采集了河南省漯河、南阳、平顶山、许昌和驻马店的烟叶近红外光谱数据和电子鼻数据，利用两者融合数据建立同一省内不同地级市烟叶产地的模式识别模型.本研究旨在探索烟叶产地识别的多维度数据分析方法，希望可以为烟草行业的发展和质量控制提供有力支持.

1　实验和算法

1.1　数据

收集了河南、云南、福建和吉林 4 个省份的烤后烟叶共 352 个，用于建立不同省份产地分类模型，其中云南省烟叶 111 个，河南省烟叶115 个，福建省烟叶 91 个，吉林省烟叶 35 个.这 352 个烟叶样本中，上部、中部和下部烟叶样本数据分别为：89、169 和 94 个.河南省的 115 个烤后烟叶中，包括漯河 27 个样本、南阳 15 个样本、平顶山 25 个样本、许昌 27 个样本和驻马店 21 个样本.这 115 个烟叶样本用来研究地级市产地分类模型，如图 1 所示，该 5 个地级市的地理位置非常接近，适合用于同一省份内小产地识别研究.



注：基于审图号为GS(2020)4814的标准地图制作，底图无修改.

图1　河南省5个地级市地图

Fig.1　The map of 5 cities in Henan

1.2　近红外光谱

对烤后烟叶进行研磨后，过 60 目筛，然后取 20 g烟叶粉末放置在内径大小为 5 cm 的样品杯中近红外扫描.实验使用了 Spotlight 400 傅立叶变换红外光谱仪，配置了漫反射积分球附件和 DTGS 检测器，该仪器由英国 PerkinElmer 公司生产.分辨率：4 cm^{-1}，扫描次数：32 次.

1.3　Heracles 电子鼻系统

Heracles 电子鼻系统是法国 Alpha MOS 公司生产的，其与 AlphaSoft，IMM-Pro 和 AroChemBase 一起专门设计用于帮助行业和实验室掌握和改善其产品的嗅觉质量.Heracles 电子鼻仪器是一种新型的气味分析手段，依据气相基本原理对顶空气体进行分析，通过机器学习等数据分析方法得到响应信息.样品中的挥发性化合物可以通过 Heracles 电子鼻系统精确分离出来，并可以通过 Arochembase 数据库进行定性分析.Heracles 电子鼻系统具有分析时间短、精确度高等特点[7].Heracles 电子鼻扫描是在室温常压下进行，取 1 g 烟叶粉末进行电子鼻扫描，烟叶粉末样品在孵化器中的加热震荡温度为 50 ℃，加热震荡时间选择 10 min.

1.4　算法

1.4.1　偏最小二乘算法(partial least squares，PLS)

本研究中，烟叶近红外光谱数据和电子鼻数据都具有高维度特征，即变量特征数远超样本数量，通常会造成维数灾难的问题.为此本文采用 PLS 算法作为降维方法.PLS 是一种常用的高维数据降维方法，通过建立原始数据与目标变量之间的线性关系，将高维数据转化为一组低维的潜在变量或因子[8-9].在降维过程中 PLS 能够提取与目标变量最相关的数据特征，实现数据的降维和压缩[10-11].

1.4.2　SVM 算法

SVM 是一种机器学习算法，用于分类和回归分析，通过构建最优的超平面来进行数据分类，具有良好的

线性和非线性分类能力.SVM 利用核函数将数据映射到高维特征空间,从而处理非线性关系[12],具有强鲁棒性、强泛化能力,并能处理高维和噪声大数据等优点.其训练过程通过优化算法和拉格朗日乘子法来找到最优的分离超平面.在预测阶段,新数据点被映射到特征空间并进行分类判断[13-14].

2 结果与讨论

2.1 近红外光谱和 Heracles 电子鼻数据

图 2 是不同省份产地的烟叶近红外光谱,对比不同省份产地的烟叶近红外光谱,云南省烟叶的吸光度信号明显更强一些,河南省烟叶的吸光度更弱一些.图 3 是河南省内部不同产地的烟叶近红外光谱数据,对比河南省内不同地级市烟叶近红外光谱,吸光度的差异主要体现在波数 4 100~5 000cm^{-1}范围之间.

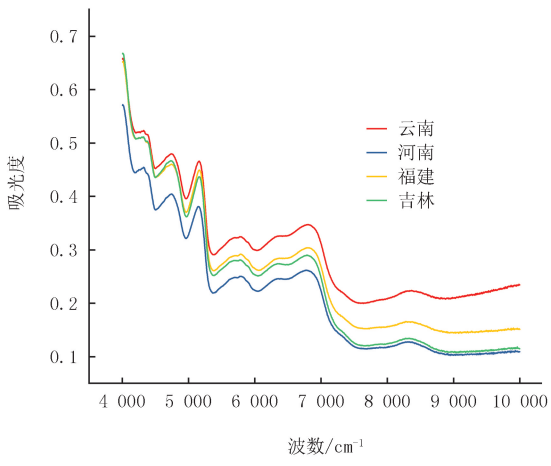


图2 不同省份产地烟叶近红外光谱

Fig.2 NIRS of different growing provinces of tobacco

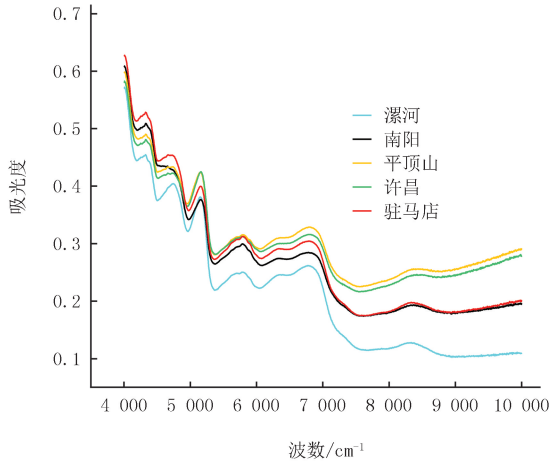


图3 河南内部不同产地烟叶近红外光谱

Fig.3 NIRS of different growing cities of tobacco in Henan

扫描得到的 Heracles 电子鼻数据如图 4 和图 5 所示.图 4 是不同省份烟叶 Heracles 电子鼻数据,图 5 是河南省内部不同产地的烟叶 Heracles 电子鼻数据.Heracles 电子鼻系统的 120 s 保留时间内,每 1 秒钟采集数据 100 个,总共采集了 12 000 个数据.利用不同颜色来代表不同省份或河南省内不同地区烟叶样品的电子鼻数据,从图 4 和图 5 可以看出,不同产地烟叶其响应值有着比较大的差异.

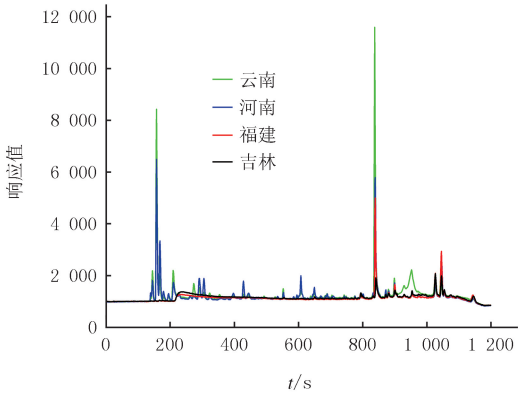


图4 不同省份产地的烟叶Heracles电子鼻数据

Fig.4 Heracles EN data of different growing provinces of tobacco

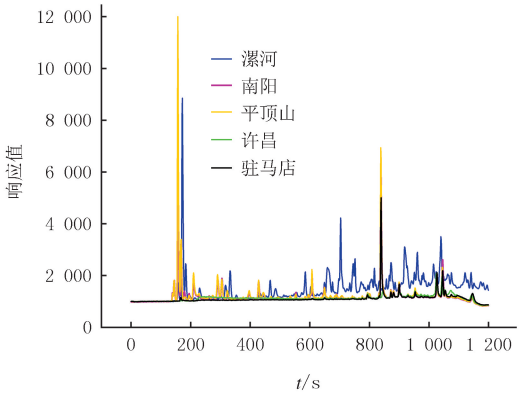


图5 河南内部不同产地烟叶Heracles电子鼻数据

Fig.5 Heracles EN data of different growing cities of tobacco in Henan

2.2 模型构建与参数优化

本工作的建模流程先采用 PLS 降维,再做 PLS 因子个数选择,最后构建烟叶产地 SVM 分类判别模型.建立 4 个省份和河南省 5 个地级市产地 SVM 分类判别模型的区别在于输入数据和产地信息的不同.输入数据包括近红外数据、电子鼻数据、近红外与电子鼻融合数据,产地信息包括 4 个省份产地与河南省 5 个地级市产地.

以河南省 5 个地级市产地的分类模型及其近红外光谱数据为例来说明本工作的建模流程.将河南省

5 个产地的近红外光谱数据进行 PLS 降维，并对 PLS 因子个数进行选择，选择标准是 SVM 分类模型的留一法交叉验证的准确率.本文没有利用更常用的 PCA 降维，而利用 PLS 降维，主要是因为 PLS 降维过程中应用到了目标信息，更有利于提高后续模型的分类准确率.在 PLS 因子个数选择的过程中，过少的 PLS 因子个数包含的信息量比较少，可能造成模型的“欠拟合”，导致模型准确率低.过多的 PLS 因子个数往往会包含过多的冗余信息，可能造成模型的“过拟合”，导致模型准确率也比较低.因此选择 PLS 因子个数时从 8 个开始，20 个结束.当 PLS 因子个数为 14 时，模型的留一法内部交叉验证准确率最高，为 98.26%，见图 6 所示.留一法内部交叉验证的流程大致是这样的：假设一个数据集有 N 个样本，将每一个样本作为测试样本，其他 $N-1$ 个样本作为训练样本.这样得到 N 个分类器，N 个测试结果.用这 N 个测试结果的平均值来衡量模型的性能.在利用 SVM 算法建立分类模型时，需要对算法的参数进行优化，其中两个重要的参数是核函数和惩罚因子.PLS 因子个数为 14，线性核函数和径向基核函数选择不同的惩罚因子，对比河南省 5 个产地 SVM 分类模型的留一法内部交叉验证准确率.由图 7 可见，PLS 因子个数为 14，选取线性核函数，惩罚因子取 30 时，模型留一法内部交叉验证准确率最高，为 98.26%.因此可确定河南省 5 个地级市产地的分类模型的 PLS 因子个数为 14，SVM 模型的核函数为线性、惩罚因子为 30.

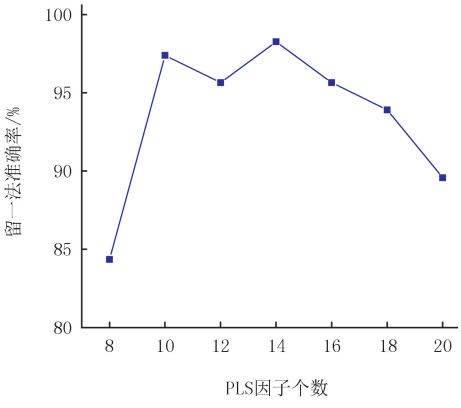


图6　PLS因子个数选择

Fig. 6　Selection results of PLS factors

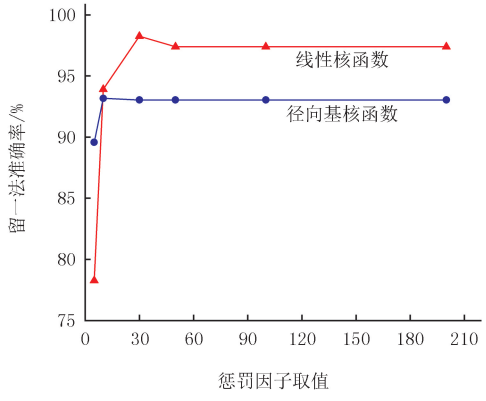


图7　参数选择

Fig. 7　Parameter optimization of SVM model

2.3　基于单一数据源的模型结果

2.3.1　两种数据源结果比较

利用 2.2 节的建模流程，分别构建基于近红外光谱、电子鼻数据的 4 个省份产地以及河南省 5 个地级市产地的分类模型，其结果如表 1 所示.仅基于近红外光谱数据的 4 个省份产地分类模型的建模准确率与留一法内部交叉验证准确率分别为 100.00%与 98.86%，仅基于电子鼻数据的 4 个省份产地分类模型的建模准确率与留一法内部交叉验证准确率分别为 95.45%与 92.33%.由此可以看出：不同省份产地的烟叶差异比较大，仅基于单一数据源即可得到准确率非常高的烟叶产地识别模型.而对于河南省内部 5 个地级市产地识别模型，无论是仅基于近红外光谱数据，还是仅基于电子鼻数据，其建模准确率非常高，但其留一法内部交叉验证准确率明显偏低.这说明仅基于一种数据源，获得烟叶的信息还比较少，模型也存在过拟合现象.

表 1　基于单一数据源的模型结果

Tab. 1　Model results based on singular data source

数据源	4 个省份产地模型		河南省 5 个地级市产地模型	
	建模准确率/%	留一法准确率/%	建模准确率/%	留一法准确率/%
近红外光谱数据	100.00	98.86	100.00	86.96
电子鼻数据	95.45	92.33	99.13	86.96

2.3.2　仅基于近红外光谱数据的 5 个产地模型留一法结果

仅基于近红外光谱数据的 5 个地级市产地分类模型的建模准确率与留一法内部交叉验证准确率分别为 100.00%与 86.96%(表 1).相较于省份产地分类模型，地级市产地模型的留一法内部交叉验证准确率下降了 11.90%.留一法内部交叉验证准确率见表 2，115 个样本中预报准确了 100 个.其中，漯河的准确率为

96.30%,南阳的准确率为 66.67%,平顶山的准确率为 92.00%,许昌的准确率为 81.48%,驻马店的准确率为 90.48%.可以看出基于 NIR 数据建立河南省内 5 个地级市的产地识别模型,其留一法内部交叉验证准确率还比较低,特别是南阳的准确率只有 66.67%.

表 2 基于近红外光谱数据 5 个地级市产地识别模型留一法内部交叉验证结果

Tab. 2 Confusion matrix for the leaving-one-out cross-validation of SVM model based on NIR data to identify 5 cities

地区	漯河	南阳	平顶山	许昌	驻马店	准确率/%
漯河	26	1	0	0	0	96.30
南阳	4	10	1	0	0	66.67
平顶山	1	0	23	1	0	92.00
许昌	0	0	2	22	3	81.48
驻马店	0	0	0	2	19	90.48

2.3.3 仅基于电子鼻数据的 5 个产地模型留一法结果

仅基于电子鼻数据的 5 个地级市产地分类模型的建模准确率与留一法内部交叉验证准确率分别为 99.13%与 86.96%(表 1).相较于省份产地分类模型,地级市产地模型的留一法内部交叉验证准确率显著下降了 5.37%.在留一法内部交叉验证中,电子鼻模型对于许昌的预测准确率偏低,仅有74.07%(表 3).

表 3 基于 Heracles 电子鼻数据 5 个地级市产地识别模型留一法内部交叉验证结果

Tab. 3 Confusion matrix for the leaving-one-out cross-validation of SVM model based on Heracles EN data to identify 5 cities

地区	漯河	南阳	平顶山	许昌	驻马店	准确率/%
漯河	25	2	0	0	0	92.59
南阳	1	13	1	0	0	86.67
平顶山	0	0	24	0	1	96.00
许昌	1	0	5	20	1	74.07
驻马店	1	0	0	2	18	85.71

对比表 2 和表 3,可以看出,仅基于单一近红外光谱数据模型对许昌的预测准确率较高,达到81.48%,但南阳的预测准确率较差,仅为 66.67%.仅基于单一电子鼻数据模型对许昌的预测准确率比较低,仅为 74.07%,但南阳的准确率高,为 86.67%.这两个模型的其他 3 个地级市的准确率则较为接近.通过对比近红外光谱与电子鼻的地级市分类模型结果可以看出,近红外光谱与电子鼻数据是从两个不同的维度来反映烟叶样本的信息特征,通过融合两种维度的数据,可以为模型提供更多的信息,进而增加模型准确率.

2.4 基于融合数据的模型结果讨论

无论是基于单一近红外光谱数据的烟叶产地识别模型,还是基于单一 Heracles 电子鼻数据的烟叶产地识别模型,对于河南、云南、福建和吉林 4 个产地可以建立准确率高的识别模型.原因是这些不同省份的地理位置距离比较远,气候差异大,烟叶本身的差异也比较大,因此模型识别准确率高.但对于河南省内部的漯河、南阳、平顶山、许昌和驻马店 5 个地级市产地,由于地理位置比较近,气候差异小,烟叶本身的差异也相应地比较小,因此模型识别准确率低,而且模型出现了过拟合现象.本文对近红外光谱数据补充了不同维度的 Heracles 电子鼻数据,两类数据融合后,增加了更多的数据信息,以此建立了河南省内 5 个地级市的产地识别准确率高的模型.

利用 PLS 对烟叶近红外光谱和 Heracles 电子鼻融合数据进行降维,选取前 14 个 PLS 因子,选择线性核函数,惩罚因子取 30,建立了河南省内部漯河、南阳、平顶山、许昌和驻马店的 5 个地级市产地识别模型(表 4),其模型建模准确率为 100%.模型留一法准确率为 98.26%,其中漯河的准确率为96.30%(表 5),南阳的准确率为 100.00%,平顶山的准确率为 100.00%,许昌的准确率为 96.30%,驻马店的准确率为 100.00%.可以看出基于融合数据建立的河南省内 5 个地级市的产地识别模型的准确率明显高于仅基于单一近红外光谱数据建立的模型,同样也高于基于单一 Heracles 电子鼻数据建立的模型,特别是南阳和许昌的识别率明显提高.

表 4　基于融合数据 5 个地级市产地识别模型建模结果

Tab. 4　Confusion matrix for the training set of SVM model based on merged data to identify 5 cities

地区	漯河	南阳	平顶山	许昌	驻马店	准确率/%
漯河	27	0	0	0	0	100.00
南阳	0	15	0	0	0	100.00
平顶山	0	0	25	0	0	100.00
许昌	0	0	0	27	0	100.00
驻马店	0	0	0	0	21	100.00

表 5　基于融合数据 5 个地级市产地识别模型留一法内部交叉验证结果

Tab. 5　Confusion matrix for the leaving-one-out cross-validation of SVM model based on merged data to identify 5 cities

地区	漯河	南阳	平顶山	许昌	驻马店	准确率/%
漯河	26	1	0	0	0	96.30
南阳	0	15	0	0	0	100.00
平顶山	0	0	25	0	0	100.00
许昌	0	0	0	26	1	96.30
驻马店	0	0	0	0	21	100.00

需要说明的是，本研究受到烟叶样品收集时间和地点的影响，收集样本比较困难，收集到的样品数比较少，特别是河南省内部 5 个地级市的样品更少，因此没有对数据进行建模集、验证集和测试集的划分，只考察了模型的建模准确率和留一法内部交叉验证准确率，这些结果初步验证了基于融合数据建立的产地识别有着更高的准确率.

3　结　论

综合以上实验结果可知，仅基于近红外光谱数据或 Heracles 电子鼻数据可有效识别地理位置较远的烟叶产地，但对地理位置较近的产地其准确率都相对较低.Heracles 电子鼻数据作为烟叶的另一种重要的信息源，可以辅助近红外光谱数据进行烟叶产地的识别.将近红外光谱数据和 Heracles 电子鼻数据进行融合，可显著提高地理位置较近的烟叶产地识别的准确率，也消除了模型过拟合问题，可能的原因是不同信息源的数据融合后，有效信息明显增加导致模型准确率提升.本文探讨了多数据源综合利用的策略，用以获取更多烟叶信息，进而建立更准确的产地识别模型.这些研究成果在烟叶品种溯源、质量监测和市场监管等方面具有重要意义，可为烟草行业的进一步发展和创新提供借鉴.

参　考　文　献

[1]　耿莹蕊，沈欢超，倪鸿飞，等.近红外光谱结合灰狼算法优化支持向量机实现烟叶产地快速鉴别[J].光谱学与光谱分析，2022，42(9)：2830-2835.
GENG Y R，SHEN H C，NI H F，et al.Support vector machine optimized by near-infrared spectroscopic technique combined with grey wolf optimizer algorithm to realize rapid identification of tobacco origin[J].Spectroscopy and Spectral Analysis，2022，42(9)：2830-2835.

[2]　鲁梦瑶，杨凯，宋鹏飞，等.基于卷积神经网络的烟叶近红外光谱分类建模方法研究[J].光谱学与光谱分析，2018，38(12)：3724-3728.
LU M Y，YANG K，SONG P F，et al.The study of classification modeling method for near infrared spectroscopy of tobacco leaves based on convolution neural network[J].Spectroscopy and Spectral Analysis，2018，38(12)：3724-3728.

[3]　束茹欣，孙平，杨凯，等.基于 NIR-PCA-SVM 联用技术的烤烟烟叶产地模式识别[J].烟草科技，2011，44(11)：50-52.
SHU R X，SUN P，YANG K，et al.NIR-PCA-SVM based pattern recognition of growing area of flue-cured tobacco[J].Tobacco Science & Technology，2011，44(11)：50-52.

[4]　张浩，刘振，王玲，等.基于近红外光谱结合机器学习算法检测食用明胶品种溯源的研究[J].河南农业大学学报，2021，55(3)：460-467.
ZHANG H，LIU Z，WANG L，et al.Determination of edible gelatin origins based on near-infrared spectroscopy coupled with machine learning methods[J].Journal of Henan Agricultural University，2021，55(3)：460-467.

[5]　王文俊，沙云菲，汪阳忠，等.近红外和电子鼻数据融合识别不同香型风格[J].光谱学与光谱分析，2023，43(1)：133-137.

WANG W J,SHA Y F,WANG Y Z,et al.Discriminating flavor styles via data fusion of NIR and EN[J].Spectroscopy and Spectral Analysis,2023,43(1):133-137.

[6] ZHANG H B,LIU T A,SHU R X,et al.Using EN-NIR with support vector machine for classification of producing year of tobacco[J].Spectroscopy and Spectral Analysis,2018,38(5):1620-1625.

[7] 张玖捌,张伟,费程浩,等.基于 Heracles NEO 超快速气相电子鼻的硫熏白芍快速鉴别研究[J].中国中药杂志,2022,47(14):3781-3787.
ZHANG J B,ZHANG W,FEI C H,et al.Rapid identification of raw and sulfur-fumigated Paeoniae Radix Alba based on Heracles NEO ultra-fast gas phase electronic nose[J].China Journal of Chinese Materia Medica,2022,47(14):3781-3787.

[8] 鄢悦,张红光,卢建刚,等.基于光谱信息散度的近红外光谱局部偏最小二乘建模方法[J].计算机与应用化学,2017,34(5):351-355.
YAN Y,ZHANG H G,LU J G,et al.Spectral-information-divergence based local pls modeling algorithm in near infrared spectroscopy[J].Computers and Applied Chemistry,2017,34(5):351-355.

[9] 赵娟娟,叶顺,徐可,等.基于提取不同中红外光谱特征信息的烟叶部位判别研究[J].河南师范大学学报(自然科学版),2021,49(1):45-49.
ZHAO J J,YE S,XU K,et al.Research on discrimination of tobacco leaf parts based on extracting different information of MIR[J].Journal of Henan Normal University(Natural Science Edition),2021,49(1):45-49.

[10] ZENG Z D,ZHANG B H,ZHAN Y F,et al.Method comparison of sample pretreatment and discovery of differential compositions of natural flavors and fragrances for quality analysis by using chemometric tools[J].Journal of Chromatography B,2023,1222:123690.

[11] WU R X,TIAN Z Z,ZHANG C T,et al.Uniformity evaluation of stem distribution in cut tobacco and single cigarette by near infrared spectroscopy[J].Vibrational Spectroscopy,2022,121:103401.

[12] VAPNIK V N.Statistical learning theory[M].New York:Wiley,1998.

[13] QIN Y H,LIU X P,ZHANG F M,et al.Improved deep residual shrinkage network on near infrared spectroscopy for tobacco qualitative analysis[J].Infrared Physics & Technology,2023,129:104575.

[14] ARIANTI N D,SAPUTRA E,SITORUS A.An automatic generation of pre-processing strategy combined with machine learning multivariate analysis for NIR spectral data[J].Journal of Agriculture and Food Research,2023,13:100625.

Classification of tobacco leave parts based on the fusion of near-infrared spectroscopy and Heracles electronic nose data

Wang Yangzhong[1], Zhang Xin[1], Cai Zhenbo[1], Huang Wen[1], Fei Ting[1], Wu Da[1], Zhang Xufeng[2], Meng Xiangzhou[2], Shu Ruxin[1]

(1. Technology Center, Shanghai Tobacco Group Co., Ltd., Shanghai 200082, China;
2. College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China)

Abstract: In this study, the tobacco leaf origin identification models were established in four provinces in China (Yunnan, Henan, Fujian, and Jilin) and five prefecture-level cities within Henan Province (Luohe, Nanyang, Pingdingshan, Xuchang, and Zhumadian) by utilizing near-infrared spectroscopy data, Heracles electronic nose data, and a fusion of both datasets. In geographically distant provinces, accurate origin identification models with relatively high precision were successfully constructed by using a single data source. However, in the five closely located cities in Henan Province, where geographical proximity, minimal climate variations, and high tobacco leaf similarity were evident, the accuracy of the origin identification model based on a single information source was comparatively lower. To enhance the accuracy of tobacco origin identification in the five prefecture-level cities in Henan Province, a fusion of near-infrared spectroscopy data and Heracles electronic nose data is performed. The increased information content in the fused dataset significantly improved the identification accuracy in these five origin regions. The Leave-One-Out cross-validation accuracy in these regions was measured at 98.26%, surpassing the 86.96% accuracy of the single-data-source discrimination model before data fusion. This study demonstrates the capability of Heracles electronic nose data to complement near-infrared spectroscopy data across different data dimensions, providing new perspectives for tobacco variety tracing, quality monitoring, and market supervision.

Keywords: near-infrared spectroscopy; Heracles electronic nose; data fusion; support vector machine

[责任编校 赵晓华 陈留院]