王林,郑明明,王翀,等.基于近红外光谱的卷烟配方模块香型预测[J].华中农业大学学报,2024,43(1):226-231.
DOI:10.13300/j.cnki.hnlkxb.2024.01.026

基于近红外光谱的卷烟配方模块香型预测

王林[1],郑明明[2],王翀[2],吴庆华[2],崔南方[2],李建斌[2]

1.湖北中烟工业有限责任公司,武汉 430040; 2.华中科技大学管理学院,武汉 430074

摘要 为提高卷烟配方模块的分类识别准确率,并为卷烟配方模块的科学评估提供技术支撑,提出了一种基于近红外光谱特征筛选的卷烟配方模块香型预测方法。选取2017—2019年238个卷烟配方模块样品的近红外光谱数据,结合特征工程中的递归特征消除法和BP神经网络、随机森林、XGBoost 3种机器学习技术,构建了基于特征变量的香型预测模型。与全光谱数据训练的分类效果对比,经过递归特征消除法筛选后的光谱特征变量能够有效提升卷烟配方模块香型的识别准确率,其中,XGBoost算法分类效果最佳,模型对测试集的识别准确率达到了90.41%。结果表明,基于近红外光谱特征筛选的香型预测方法对卷烟配方模块的快速定位、科学评价及卷烟配方设计等有一定的辅助决策作用。

关键词 烟叶; 香型; 近红外光谱; 递归特征消除; 随机森林; XGBoost

中图分类号 TS452$^+$.1 **文献标识码** A **文章编号** 1000-2421(2024)01-0226-06

卷烟配方模块的香型是影响卷烟香气风格的重要指标,也是卷烟配方设计与产品维护的重要评价依据[1]。卷烟配方模块的香型可分为清香型、中间香型和浓香型三大类[2]。在进行卷烟配方设计时,首先需要对配方中的原材料进行感官质量评定,以确定香型类别。传统技术下,香型的评定主要通过评吸专家的“五评三定”完成,这种评定方法主要依靠专家的经验,主观性较强,缺少科学技术支持。近年来,计算机技术发展迅速,烟草行业的技术人员开始探索评价香型类别的新技术。邱昌桂等[3]将烤烟中检测的68种致香成分作为输入变量,提出了一种基于致香成分结合GA-SVM算法的烤烟香型自动识别方法。周泽弘等[4]基于消除共线性的化学指标,利用RBF神经网络建立了库存烟叶香型的预测模型。郭东锋等[5]利用化学指标和感官评价指标建立了不同香型的机器学习模型并进行了性能比较。这些研究主要利用化学指标作为模型的输入变量,由于常规实验室检测流程繁琐,全面检测出配方模块中的数百种化学成分难度较大,能作为学习变量的化学指标较少,研究还有进一步提升的空间。

近红外光谱(NIRS)具有简单、快速、成本低、环保、信息量大等优点,已被广泛应用于烟草领域的定量检测和定性分析[6]。不同类型的烟草化学成分差异明显,这种差距可以通过近红外光谱中波峰和波谷的位置及强弱体现出来。许多研究人员试图通过基于近红外光谱的机器学习方法完成烟草的指标预测。鲁梦瑶等[7]应用卷积神经网络结合近红外光谱技术构建了烟叶产区分类模型,实现了较强的产地判别能力。栾丽丽等[8]应用近红外光谱技术和PPF-DPLS-SVM多算法融合方法,提高了通过客观数据判别烤烟香型的准确率。郝贤伟等[9]利用近红外光谱技术探索了全面评价烟叶质量的可行性。研究表明,由于高维、高频噪声和冗余信息的影响,运用近红外光谱的所有变量直接构建识别模型,不仅增加了建模复杂度,还会降低识别性能和泛化能力[10]。研究人员多数使用PCA方法对高维光谱数据进行降维,但PCA是一种线性算法,其目的是提取少量综合指标,最大化地反映原始数据的规律,不能解释特征之间复杂的多项式关系,也没有考虑数据的类别信息[11]。

笔者前期研究运用PCA算法对近红外光谱数据降维,可实现利用12个综合变量反映原数据96%信息的效果,但因为原数据中与香型相关度低的变量占多数,PCA算法未能消除这些冗余变量,导致模型

收稿日期:2022-06-22
基金项目:湖北中烟工业有限责任公司科技项目(2021JCYL3JS2B022)
王林,E-mail:wanglin@market.hbtobacco.cn
通信作者:吴庆华,E-mail:qinghuawu1005@gmail.com

的训练效果不理想(未发表)。特征递归消除法可以有效地消除冗余变量,提高预测模型的稳定性与鲁棒性,实现对高维大数据集特征的快速筛选,但运用该方法实现近红外光谱数据降维的研究还鲜见报道[12]。针对上述情况,本研究结合基于随机森林的特征递归消除法和BP神经网络、XGBoost、随机森林3种机器学习方法,构建卷烟配方模块的近红外光谱香型预测模型,通过对比全光谱数据和特征筛选后的光谱数据分别作为输入变量训练的3种模型的性能差异,探索出高效的模型训练方法,以期为卷烟配方的质量评价提供良好的辅助决策参考。

1　材料与方法

1.1　材料

试验材料为湖北中烟工业有限责任公司提供的2017—2019年份的卷烟配方模块,共计238份。其中,清香型配方模块样品76份,中间香型配方模块样品104份,浓香型配方模块58份,配方模块的生产年份主要集中在2017—2019年,具体分布如表1所示。

1.2　仪器设备

布鲁克MATRIX-I工业现场级傅立叶变换近红外光谱仪(德国布鲁克光谱仪器公司生产);宾德BD400标准培养箱(德国宾德公司生产);AU-ARI中药粉碎机(温岭市奥力中药机械有限公司生产)。

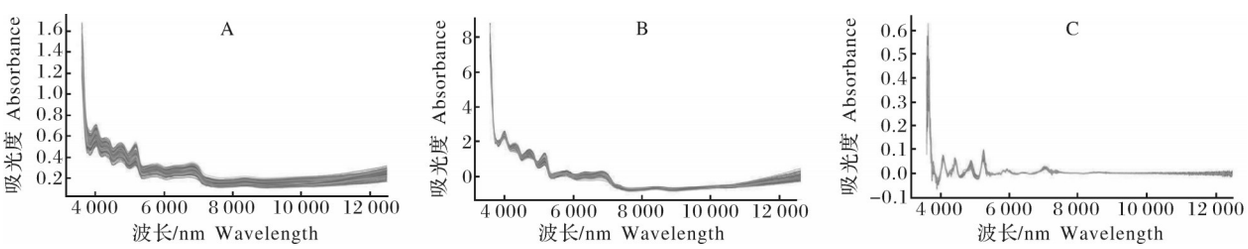
表1　2017—2019年试验数据统计

Table 1　Statistics of experimental data 2017-2019

香型 Aroma type	2017	2018	2019	总计 Total
清香型 Fragrance	25	11	40	76
中间香型 Intermediate fragrance	42	10	52	104
浓香型 Strong fragrance	22	7	29	58
样本量 Sample size	89	28	121	238

1.3　数据采集和预处理

在进行近红外光谱数据采集之前需对待测样品进行预处理。238份卷烟配方模块在取样后放置于标准培养箱,40 ℃烘干2 h,使用中药粉碎机将样品磨成粒径0.425 nm大小的粉末并用密封袋密封待测。取约50 g待测样品放入样品杯中,用压紧器压实后放到近红外光谱仪上进行数据采集。其中,数据采集的操作过程严格保持在温度22 ℃、相对湿度60%的环境条件下进行。近红外光谱仪采集的波段范围为3 600～12 500 cm^{-1},光谱的采集方式为漫反射,光谱分辨率设置为16 cm^{-1},每个待测样品均重复采集5条光谱数据,取平均值作为研究数据。由于原始近红外光谱中存在基线漂移、高频噪声、分量间相互干扰等问题,需要在应用光谱前进行相应的预处理,以提高信噪比,增强模型的预测性能[13]。本次研究采用常用的标准正态校正处理(SNV)和一阶导数平滑处理结合的方法来减小数据采集过程中基线漂移、高频噪声等问题带来的影响,处理后的数据如图1所示。



A:原始近红外光谱图 Initial NIR spectroscopy;B:标准正态校正处理后近红外光谱图 NIR spectroscopy after SNV;C:一阶导数处理后近红外光谱图 NIR spectroscopy after first derivative processing.

图1　近红外光谱预处理图

Fig.1　Pre-processed NIR spectroscopy

1.4　研究方法

随机森林(random forests)是Breiman[14]结合Bagging集成理论和随机子空间方法提出的机器学习算法,是一种集成多个决策树的集成学习分类模型。本研究中近红外光谱数据的变量较多,随机森林算法能够很好地解决多元数据带来的共线性和数据冗余问题,是目前分类效果较理想的算法之一[15]。XGBoost(extreme gradient boosting)是源于梯度提升框架的一种可扩展的树增强系统,被数据科学家广泛使用,并在许多问题上取得了良好的结果[16]。BP(back propagation)神经网络是一种利用误差反向传播进行训练的模型,具有学习性强、容错性高、实时更新等优点,能够进行复杂的信息运算,广泛应用于近红外光谱的数据分析[17]。递归特征消除法是

Guyon等[18]为了解决基因选择问题而提出的，与支持向量机等模型结合可以轻松处理特征变量庞大的数据集。基于随机森林的特征递归消除法（RF-RFE）可以有效地消除冗余变量，提高预测模型的稳定性与鲁棒性[12]，适用于本研究进行近红外光谱数据关键变量的筛选。本研究采用随机森林算法并利用Scikit-learn库中的随机森林包、XGBoost包的分类器和Scikit-learn库以及选择基于PyTorch库实现的BP神经网络算法作为模型的构建方法，并采用特征递归消除法进行近红外光谱数据关键变量的筛选。

1.5 评价指标

本研究通过训练准确率和预测准确率对模型进行性能评价，准确率（accuracy，A）的计算方法为：

$$A=\frac{N_c}{N_t}\times 100\%$$

式中，N_c为识别正确的样本数量，N_t为总的样本数量。

1.6 模型超参数设置

本研究为模型设置了提前终止训练的条件，以获得较好的BP神经网络模型。针对XGBoost和随机森林算法，使用网格搜索确定模型重要参数，并采用五折交叉验证减少数据集划分的随机性对结果的影响，超参数的取值如表2所示。

表2 超参数取值汇总

Table 2 Summary of super-parameter values

模型 Model	超参数 Super-parameter	取值 Value
XGBoost	learning_rate	0.01
	n_estimators	700
	max_depth	8
	min_child_weight	3
	gamma	0
	subsample	0.5
	colsample_btree	0.7
随机森林 Random forest	estimators	40
	max_depth	5
	max_features	9
	min_samples_leaf	7
	min_samples_split	17

2 结果与分析

2.1 数据集划分

将238条模块的近红外光谱数据按照4∶1的比例随机划分5次，形成5份数据集，每份数据集包含训练集190条数据，测试集48条数据，数据集中3种香型在训练集和测试集中分布比例相同，具体的数据分布情况如表3所示。

表3 数据集分布情况

Table 3 Distribution of data set

香型 Aroma type	训练集 Training set	测试集 Testing set
清香型 Fragrance	58	15
中间香型 Intermediate fragrance	83	21
浓香型 Srong fragrance	49	12

2.2 基于全光谱数据的香型分类

以全光谱作为输入数据、香型为分类标签，采用BP神经网络、XGBoost、随机森林模型对5份随机生成的训练集和测试集数据进行模型学习和预测，各指标取平均值作为模型的综合评价指标，训练结果见表4。基于BP神经网络、XGBoost、随机森林模型的香型预测模型在准确率上并非全都表现良好。训练集方面，基于XGBoost和随机森林模型的训练准确率均值达到了100%，基于BP神经网络模型的训练效果不理想，准确率为90.63%。对于测试集，基于XGBoost模型的测试集平均预测准确率达70%以上，其他2种方法的准确率较低，均在70%以下。通过分析全光谱数据训练模型的预测效果，可以发现上述3种模型普遍存在训练集准确率远高于测试集准确率的过拟合现象。原因是烟草公司2017—2019年间模块生产数量少且库存卷烟配方模块数量有限，本试验可用的训练数据较少，而大量的输入变量中与香型相关度低的干扰变量较多，模型训练过程中算法无法在少量的数据集中准确识别出与分类标签相关度较高的变量。为了解决这些问题，需要通过特征工程筛选出与香型相关的特征变量，提高模型的预测效果。

表4 基于全光谱数据构建的模型评价结果

Table 4 Results of the model trained by all data of NIR spectroscopy %

模型 Model	训练准确率均值 Average of training accuracy	预测准确率均值 Average of prediction accuracy
BP神经网络 BP neural network	90.63	58.75
XGBoost	100.00	76.25
随机森林 Random forest	100.00	65.42

2.3 特征筛选

为了减少无关变量对模型训练的干扰，采用基于随机森林的递归特征消除法和五折交叉验证法进行相关指标的特征变量筛选，实现对近红外光谱

1 153个特征变量的特征评价。经过试验，得到了各个特征变量的排名顺序，并筛选出与香型相关的39个特征波长，主要分布于区间[4 300，4 400]、[5 600，5 900]、[8 100，8 800]、[10 500，12 000]。由筛选结果可以看出，影响香型的特征大多为偏度和峰度，说明不同香型模块的偏度和峰度存在一定的差异，是进行影响香型分类的重要特征变量。

2.4　基于特征变量的香型分类

以筛选出的特征变量作为输入数据、香型作为分类标签，通过BP神经网络、XGBoost、随机森林3种模型对训练集和测试集数据进行模型学习和预测，训练结果见表5～7。由表5可知，基于特征变量训练出的3种模型相较于全光谱数据在预测效果上均有明显的提升。训练集方面，XGBoost和随机森林训练出的模型整体训练准确率仍然稳定地达到了100%，BP神经网络的整体分类准确率有所下滑，处在90%以下。对于测试集，随机森林模型的预测准确率均值为89.58%，而XGBoost在测试集上平均预测准确率达到了90.41%，且在数据集不同的情况下，模型的预测准确率较为稳定。相对于全光谱训练的模型，随机森林和XGBoost预测效果有明显的提升。而BP神经网络的整体正确率较低，但与全光谱数据训练出来的模型相比仍然有一定的提升。由表6可知，整体上而言，中香型和浓香型的模块预测效果较好，对于XGBoost和随机森林2种方法，中间型的预测正确率均达90%以上。而对于清香型来说，预测效果则相对较差，这种情况可能与数据集中各香型样本数量分布不均衡有关，如果要进一步提高预测效果，则需要在后续研究中进一步补充新的数据并平衡各个香型样本的数量占比。

表5　基于特征变量构建的香型模型评价结果

Table 5　Results of the model trained by characteristic wavelengths of aroma type　%

模型 Model	训练准确率均值 Average of training accuracy	预测准确率均值 Average of prediction accuracy
BP神经网络 BP neural network	87.57	75.83
XGBoost	100.00	90.41
随机森林 Random forest	100.00	89.58

表6　3种香型测试集预测准确率

Table 6　Results of mean prediction accuracy of three types of aroma type　%

香型 Aroma type	BP神经网络 BP neural network	XGBoost	随机森林 Random forest
清香型 Fragrance	72.00	85.53	84.00
中间香型 Intermediate fragrance	75.24	94.29	94.29
浓香型 Srong fragrance	81.67	90.00	88.33
总计 Total	75.83	90.41	89.58

表7　不同数据集下的模型评价结果

Table 7　Results of model under different datasets　%

数据集 Data set	随机森林 Random forest		XGBoost	
	训练准确率 Training accuracy	预测准确率 Prediction accuracy	训练准确率 Training accuracy	预测准确率 Prediction accuracy
数据集1 Data set 1	100.00	95.83	100.00	93.75
数据集2 Data set 2	100.00	91.67	100.00	89.58
数据集3 Data set 3	100.00	87.50	100.00	89.58
数据集4 Data set 4	100.00	87.50	100.00	89.58
数据集5 Data set 5	100.00	85.40	100.00	89.58
均值 Average	100.00	89.58	100.00	90.41

3　讨　论

为实现对卷烟配方模块香型风格的快速定位和科学评价，本研究提出了一种基于递归特征消除法和机器学习的分类模型。该模型利用配方模块的近红外光谱数据中与香型关联度较强的特征变量，通过递归特征消除法降低学习任务的难度并提升模型的泛化能力，实现对香型的识别。为了验证模型的有效性，本研究采用BP神经网络、随机森林、XGBoost 3种算法分别对全光谱数据和特征筛选后数据进行模型训练，并在训练集和测试集上进行了测试。结果表明，经过特征筛选的数据训练出的模型在预测效果上显著优于全光谱数据的模型，其中经过特征筛选的随机森林算法和XGBoost算法训练的模型平均预测准确率分别为89.58%和90.41%（表7）。

随机森林算法通过构建多个决策树并进行集成得出最终的分类结果，具有解决数据共线性和冗余问题以及测量变量重要性的优点，可以应用于高维数据集的特征筛选，并且具有很好的泛化能力和鲁棒性[15]。研究结果表明，经过特征筛选的随机森林算法表现出了较高的分类准确率。XGBoost是一种

提升框架，能够解决复杂的机器学习任务，特别适用于大规模且高维度的问题。XGBoost采用了梯度提升技术和正则化方法来防止过拟合，能够自动处理缺失数据和异常值，并且有效地应对样本不平衡和数据噪声问题[19]。在本研究中，XGBoost算法取得了较高的分类准确率。BP神经网络算法是一种反向传播算法，可通过输入层、隐藏层和输出层3种节点层构建模型，具有学习性强、容错性高、实时更新等特点，可以快速地处理大量数据，并且适用于非线性关系数据和特征提取[17]。但是，BP神经网络也存在着过拟合等问题，难以直接处理高维度的近红外光谱数据。

本研究中3种机器学习算法发挥了各自的优点，并结合递归特征消除法筛选特征变量进一步提升了算法的泛化能力和预测精度。试验结果表明，在利用递归特征消除法从高维度的近红外光谱数据中提取的特征信息结合随机森林算法、XGBoost算法可用于识别卷烟配方模块的香型风格特征，实现香型指标评价的客观化，为烟草行业人员提供辅助决策的科学依据。

参考文献 References

[1] 乔学义，申玉军，马宇平，等.不同香型烤烟烟叶香韵研究[J].烟草科技，2014，47(2)：5-7.QIAO X Y，SHEN Y J，MA Y P，et al.Study on characteristic aroma notes of flue-cured tobacco leaves of different flavor styles[J].Tobacco science & technology，2014，47(2)：5-7(in Chinese with English abstract).

[2] 李章海，王能如，王东胜，等.不同生态尺度烟区烤烟香型风格的初步研究[J].中国烟草科学，2009，30(5)：67-70.LI Z H，WANG N R，WANG D S，et al.Preliminary study of aroma type styles of flue-cured tobacco in different ecological scale regions[J].Chinese tobacco science，2009，30(5)：67-70(in Chinese with English abstract).

[3] 邱昌桂，孔兰芬，杨式华，等.基于GA-SVM算法的烤烟香型自动识别研究[J].烟草科技，2019，52(2)：101-108.QIU C G，KONG L F，YANG S H，et al.Automatic recognition of flavor types of flue-cured tobacco based on GA-SVM algorithm[J].Tobacco science & technology，2019，52(2)：101-108(in Chinese with English abstract).

[4] 周泽弘，曹淋海，王昌全，等.基于RBF神经网络建立库存烟叶香型的预测模型[J].中国烟草科学，2016，37(2)：65-70.ZHOU Z H，CAO L H，WANG C Q，et al.The establishment of prediction model of inventory tobacco flavor based on RBF neural network[J].Chinese tobacco science，2016，37(2)：65-70(in Chinese with English abstract).

[5] 郭东锋，闫宁，胡海洲，等.基于机器学习算法的烤烟香型分类研究[J].江西农业学报，2016，28(2)：43-48.GUO D F，YAN N，HU H Z，et al.Study on classification of flue-cured tobacco based on machine learning methods[J].Acta agriculturae Jiangxi，2016，28(2)：43-48(in Chinese with English abstract).

[6] 徐广通，袁洪福，陆婉珍.现代近红外光谱技术及应用进展[J].光谱学与光谱分析，2000，20(2)：134-142.XU G T，YUAN H F，LU W Z.Development of modern near infrared spectroscopic techniques and its applications[J].Spectroscopy and spectral analysis，2000，20(2)：134-142(in Chinese with English abstract).

[7] 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(in Chinese with English abstract).

[8] 栾丽丽，王宇恒，胡文雁，等.应用近红外光谱和多算法融合方法分析烤烟的香型风格特征[J].光谱学与光谱分析，2017，37(7)：2046-2049.LUAN L L，WANG Y H，HU W Y，et al.Analysis of flue-cured tobacco flavor style features using near-infrared spectroscopy and multiple algorithms fusion[J].Spectroscopy and spectral analysis，2017，37(7)：2046-2049(in Chinese with English abstract).

[9] 郝贤伟，黄文勇，徐志强，等.基于近红外光谱技术的云南片烟综合质量评价[J].中国烟草科学，2022，43(2)：58-63.HAO X W，HUANG W Y，XU Z Q，et al.Comprehensive quality evaluation of Yunnan tobacco strips based on near infrared spectroscopy[J].Chinese tobacco science，2022，43(2)：58-63(in Chinese with English abstract).

[10] ZHANG L，DING X Q，HOU R C.Classification modeling method for near-infrared spectroscopy of tobacco based on multimodal convolution neural networks[J/OL].Journal of analytical methods in chemistry，2020：9652470[2022-06-22].https://doi.org/10.1155/2020/9652470.

[11] 李武，胡冰，王明伟.基于主成分分析和支持向量机的太赫兹光谱冰片鉴别[J].光谱学与光谱分析，2014，34(12)：3235-3240.LI W，HU B，WANG M W.Discrimination of varieties of borneol using terahertz spectra based on principal component analysis and support vector machine[J].Spectroscopy and spectral analysis，2014，34(12)：3235-3240(in Chinese with English abstract).

[12] 冯晓荣，瞿国庆.基于深度学习与随机森林的高维数据特征选择[J].计算机工程与设计，2019，40(9)：2494-2501.FENG X R，QU G Q.Feature selection for high dimensional data based on deep learning and random forest[J].Computer engineering and design，2019，40(9)：2494-2501(in Chinese with English abstract).

[13] 王玲，李定明，钱红娟，等．近红外分析中的基线漂移及校正方法[J]．分析试验室，2016，35(10)：1203-1208．WANG L，LI D M，QIAN H J，et al．Baseline drift and calibration methods in NIR analysis[J]．Chinese journal of analysis laboratory，2016，35(10)：1203-1208(in Chinese with English abstract)．

[14] BREIMAN L．Random forests[J]．Machine learning，2001，45(1)：5-32．

[15] 丁子予，岳学军，曾凡国，等．基于机器学习和深度学习的玉米种子活力光谱检测[J]．华中农业大学学报，2023，42(3)：230-240．DING Z Y，YUE X J，ZENG F G，et al．Spectral detection of maize seed vigor based on machine learning and deep learning [J]．Journal of Huazhong Agricultural University，2023，42(3)：230-240(in Chinese with English abstract)．

[16] CHEN T Q，GUESTRIN C．XGBoost：a scalable tree boosting system[DB/OL] ．arXrv，2016：1603.02754[2022-06-22]．https：//doi.org/10.48550/arXiv.1603.02754．

[17] 刘秀英，余俊茹，王世华．光谱特征变量和BP神经网络构建油用牡丹种子含水率估算模型[J]．农业工程学报，2020，36(22)：308-315．LIU X Y，YU J R，WANG S H．Estimation of moisture content in peony seed oil using spectral characteristic variables and BP neural network [J]．Transactions of the CSAE，2020，36(22)：308-315(in Chinese with English abstract)．

[18] GUYON I M，WESTON J，BARNHILL S，et al．Gene selection for cancer classification using support vector machines[J]．Machine learning，2002，46：389-422．

[19] WANG Z Y，POON J，WANG S Z，et al．A novel method for clinical risk prediction with low-quality data[J/OL]．Artificial intelligence in medicine，2021，114：102052 [2022-06-22]．https：//doi.org/10.1016/J.ARTMED.2021.102052．

Predicting aroma type of cigarette recipe module based on near infrared spectroscopy

WANG Lin[1]，ZHENG Mingming[2]，WANG Chong[2]，WU Qinghua[2]，CUI Nanfang[2]，LI Jianbin[2]

1.*China Tobacco Industrial Co. Ltd. at Hubei Province*，*Wuhan 430040*，*China*；
2.*College of Management*，*Huazhong University of Science and Technology*，*Wuhan 430074*，*China*

Abstract　A method for predicting the aroma type of cigarette recipe module based on near-infrared spectral feature dimensionality reduction was proposed to classify and identify the aroma type of cigarette recipe modules with near-infrared spectroscopy. The near-infrared spectral data of 238 cigarette recipe module samples from 2017 to 2019 were selected to construct an aroma prediction model based on feature variables through combining the recursive feature elimination method in feature engineering and three machine learning techniques including BP neural network，random forest and XGBoost. Compared with the classification effect of full spectrum data training，the spectral feature variables filtered by recursive feature elimination method effectively improved the recognition accuracy of aroma type of cigarette recipe module. Among them，the algorithm of XGBoost had the best classification performance，with a model recognition accuracy of 90.41% for the test set. It is indicated that the prediction method of aroma type based on the recursive feature elimination of near-infrared spectral features has a certain role in assisting decision-making in the rapid positioning，scientific evaluation and cigarette formulation design of cigarette recipe modules.

Keywords　tobacco；aroma type；near infrared spectroscopy；recursive feature elimination；random forest；XGBoost

（责任编辑：陆文昌）