中红外和近红外数据融合的香型风格判别

沙云菲[1]，黄 雯[1]，王 亮[1]，刘太昂[2]，岳宝华[2]，李敏杰[2]，尤静林[2]，葛 炯[1*]，谢雯燕[1*]

1. 上海烟草集团有限责任公司技术中心，上海 200082
2. 上海大学化学系，上海 200444

摘 要 烤烟香型的判别一直是烟草行业的关注焦点。利用中红外和近红外光谱对189份不同香型的烟叶进行分析。分别从中红外谱图数据中提取21个特征波数处以及近红外谱图数据中13个特征波数处的吸光值作为影响因素。通过主成分分析方法分别对选取的中红外、近红外数据进行烟叶清香型、中间香型和浓香型三种香型风格的定性分析。结果表明基于中红外和近红外数据PCA投影图中三种香型混淆严重，区分界面不清晰。随后，将中红外、近红外数据进行融合，将提取的34个特征波数处的吸光值同时代入主成分分析，得到基于中红外和近红外融合数据的PCA投影图。该投影图可以将不同香型的烟叶明显地区分出来。随后利用后退法和遗传算法对中红外和近红外融合后的34个吸光度值进行变量选择，后退法选择出了24个变量，遗传算法选择出了19个变量。对比34，24和19个变量的烟叶三种香型风格的主成分投影图，遗传算法虽然选择了比较少的变量，但其仍然可以将烟叶进行准确的分类。利用遗传算法对中红外和近红外融合后数据进行变量选择，剔除对烟叶香型分类影响小的因素。最后，利用支持向量机建立烟叶清香型、中间香型和浓香型分类判别模型。该模型的建模结果准确率为92.72%，其中清香型、中间香型和浓香型的准确率分别为93.75%，92.11%和91.84%。内部交叉验证留一法结果准确率为88.74%，其中清香型、中间香型和浓香型的准确率分别为90.63%，86.84%和87.76%。对未知样本预报结果的准确率为86.84%，其中清香型、中间香型和浓香型的准确率分别为88.24%，85.71%和85.71%。无论是建模结果、留一法结果和预报结果其准确率都大于85%。研究结果表明中红外和近红外数据融合可以提供更多的特征信息，利用这些信息可以建立烟叶香型风格的分类判别模型，为烟叶香型风格快速鉴别提供帮助。

关键词 中红外光谱；近红外光谱；烤烟；数据融合

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)02-0473-05

引 言

中式卷烟风格的重要构成因素之一是烤烟香型，这也一直是烟草行业的研究热点。烤烟香型通常分为清香型、中间香型和浓香型3大类[1]。近年来，随着对烤烟香型研究和认识的进一步加深进而细分成清香型、清偏中型、中偏清型、中间型、浓香型、浓偏中型和中偏浓型七大类。早期对于烤烟香型分类一般都是通过评吸人员进行感官评价，后来不少研究希望通过烟叶化学成分或近红外光谱数据建立烤烟香型的判别模型。邱昌桂[2]等利用烟叶中的68种致香成分结合数据分析和模式识别技术，提出了一种基于烟草致香成分和遗传算法-支持向量机算法的烤烟香型自动识别方法；郭东锋[3]等利用烟叶中常规化学成分结合人工神经网络算法建立烤烟香型评价模型。宋楠[4]提出了一种改进局部线性嵌入非线性降维算法首先对烟草近红外数据进行降维，然后建立了香型风格投影模型和判别模型。在前期研究中，无论是利用烟叶化学成分或者是近红外光谱数据，可能是包含的信息量还不够多，因此模型还有进一步优化的空间。在文献调研中发现中红外在快速检测中得到了应用[5-6]。刘岩[7]等运用三级红外宏观指纹图谱法对三种不同香型的白酒进行了鉴定；中红外光谱携带有大量信息，可以用食品类香型的快速识别。本研究尝试将中红外和近红外光谱进行数据融合，用来建立烤烟香型风格快速识别模型。并与仅仅利用中红外、近红外光谱数据建立烟叶香型风格模型的准确率进行对比。

收稿日期：2020-02-17，**修订日期**：2020-06-23

基金项目：中国烟草总公司科技重大专项(中烟办[2016]259)，国家自然科学基金青年基金项目(21706156)资助

作者简介：沙云菲，女，1980年生，上海烟草集团有限责任公司技术中心高级工程师 e-mail：shayf@sh.tobacco.com.cn

*通讯作者 e-mail：gej@sh.tobacco.com.cn；xiewy@sh.tobacco.com.cn

1 实验部分

1.1 材料

选取 2018 年清香型、中间香、型浓香型的烟叶样本共 189 个，其中清香型 81 个，中间香型 45 个，浓香型 63 个。

1.2 烟叶中红外光谱

称取 1 g 烘干后的烟叶粉末于试管中，加入 10 mL 正己烷，超声混匀静置一段时间，抽取 5 mL 经滤膜过滤至小试管中，静置挥发三天，利用 ThermoFisher 公司的 Nicolet iS50 傅里叶变换红外光谱仪扫描得到中红外光谱，扫描范围 4 000～650 cm^{-1}，分辨率为 4 cm^{-1}，扫描次数 16 次。烟叶中红外光谱如图 1(a)所示。

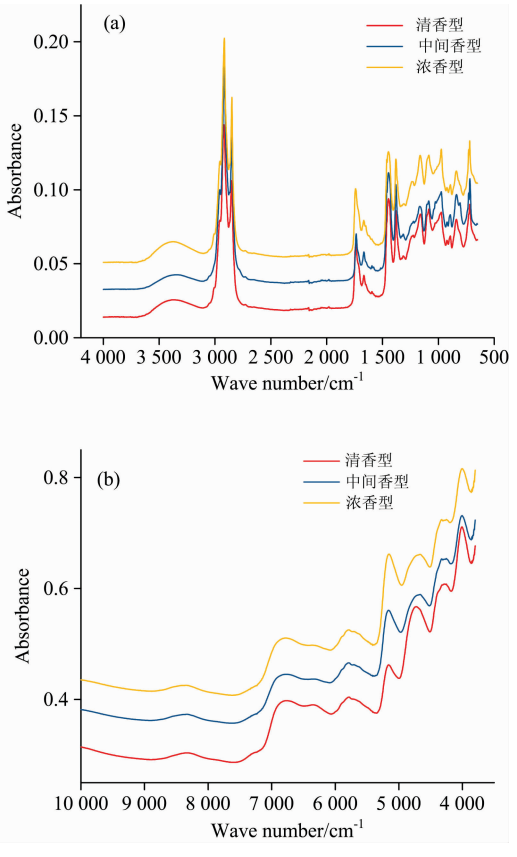

图 1 烟叶中红外光谱(a)和近红外光谱(b)

Fig. 1 MIRs (a) and NIRs (b) of tobacco samples

1.3 烟叶近红外光谱

将 15 g 60 目的烟叶粉末，放置在内径为 5 cm 样品杯中，压实后，利用 ThermoFisher 公司的 Antaris FT-NIR 分析仪扫描得到近红外光谱，扫描范围 3 800～10 000 cm^{-1}，分辨率为 4 cm^{-1}，扫描次数 16 次。烟叶近红外光谱如图 1(b)所示。

1.4 化学计量学方法

主成分分析法[8](principal component analysis，PCA)通过线性变换将烟叶中红外光谱数据或者近红外光谱数据投射到一些新的主成分变量(principal components，PCs)，这些主成分变量之间依次正交，每一个主成分都是由中红外光谱数或者近红外光谱数据线性组合而成，利用 PCA 可以考察样本在空间分布情况。

遗传算法[9](genetic algorithm，GA)是一种模仿生物界的进化规律(适者生存，优胜劣汰)演化而来的自适应全局优化搜索方法。与其他变量选择算法相比，GA 直接对研究对象操作，不需要求导和连续函数，具有全局寻优、自适应调整寻优方向等特点。

后退法[10]则是首先将所有变量都用在建模方程中，然后删除偏相关系数最小的变量，随后重复这一选择过程直到不再删除变量为止。

支持向量机分类算法[11-12](support vector classification，SVC)的核心内容是在进行建模分类过程中，构建出一个最优分类面，此最优分类面可以将样本正确分开，而且要使两类的分类空隙最大。对于构建最优分类面过程即为求函数全局最优解的过程。在利用支持向量机分类算法建立分类模型的过程中惩罚参数 c 是一个重要的影响参数，对于建立的分类模型的准确率和预报能力影响显著。

2 结果与讨论

2.1 预处理

为了提高信噪比，对中红外和近红外谱图数据进行一阶导数和 Savizky-Golay 平滑。选取烟叶中红外光谱数据 21 个

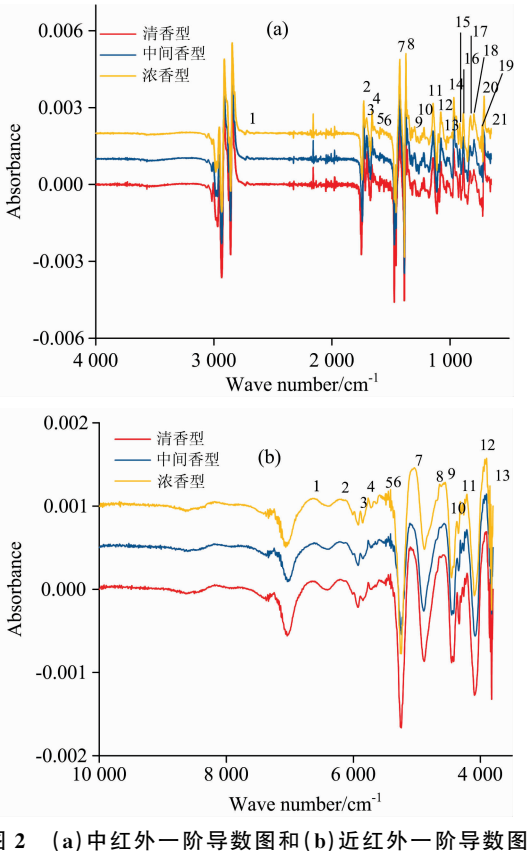

图 2 (a)中红外一阶导数图和(b)近红外一阶导数图

Fig. 2 (a)The first derivative MIR spectra and (b)The first derivative NIR spectra

和近红外光谱数据 13 个特征波数对应的吸光度值作为影响因素。如图 2 所示。

2.2　香型风格特征投影分析模型结果

图 3 分别是基于中红外数据(21 个影响因素)、近红外数据(13 个影响因素)及中红外和近红外融合数据(34 个影响因素)的烟叶清香型、中间香型、浓香型三种香型的 PCA 投影图。

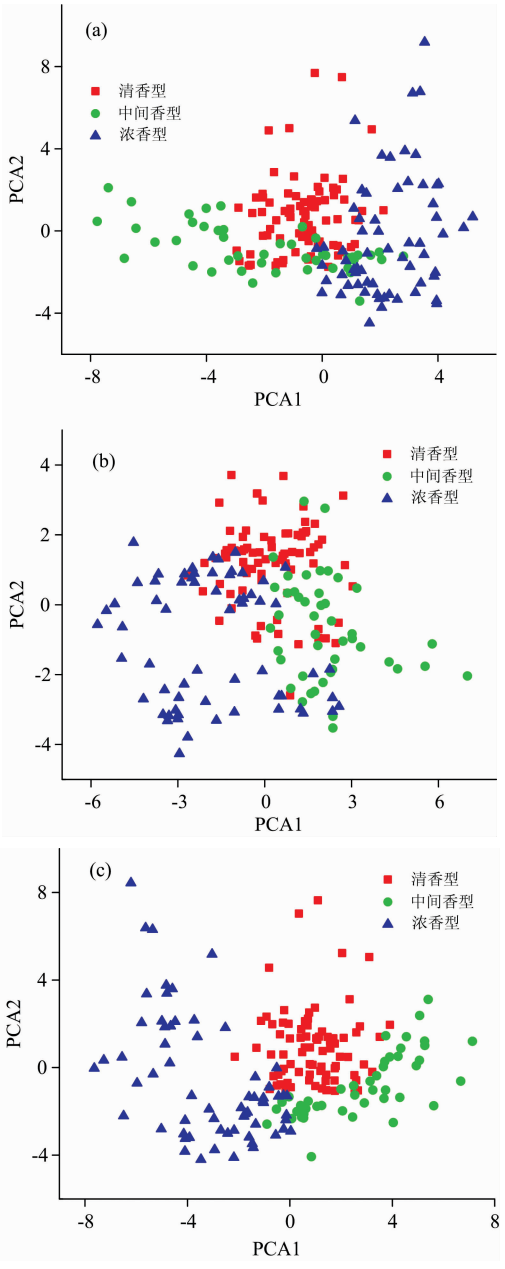

图 3　(a)基于中红外数据的 PCA 投影图;(b)基于近红外数据的 PCA 投影图和(c)基于中红外和近红外融合数据的 PCA 投影图

Fig. 3　(a) PCA projection plot based on MIR; (b) PCA projection plot based on NIR and (c) PCA projection plot based on MIR and NIR

由图 3 可见，基于中红外和近红外数据 PCA 投影图中三种香型混淆严重，区分界面不清晰。基于中红外和近红外融合数据的 PCA 投影图清香型、中间香型和浓香型数据分类清晰，有比较明显的区分界面。

2.3　中红外和近红外融合数据的变量选择

中红外和近红外融合数据共有 34 个影响因素，分别用后退法和 GA 进行变量选择。图 4 是基于 34 个全部影响因素、后退法选择的 24 个影响因素(中红外 14 个，近红外 10 个)，GA 选择的 19 个影响因素(中红外 11 个，近红外 8 个)的清香型、中间香型、浓香型三种香型风格的 PCA 投影图。

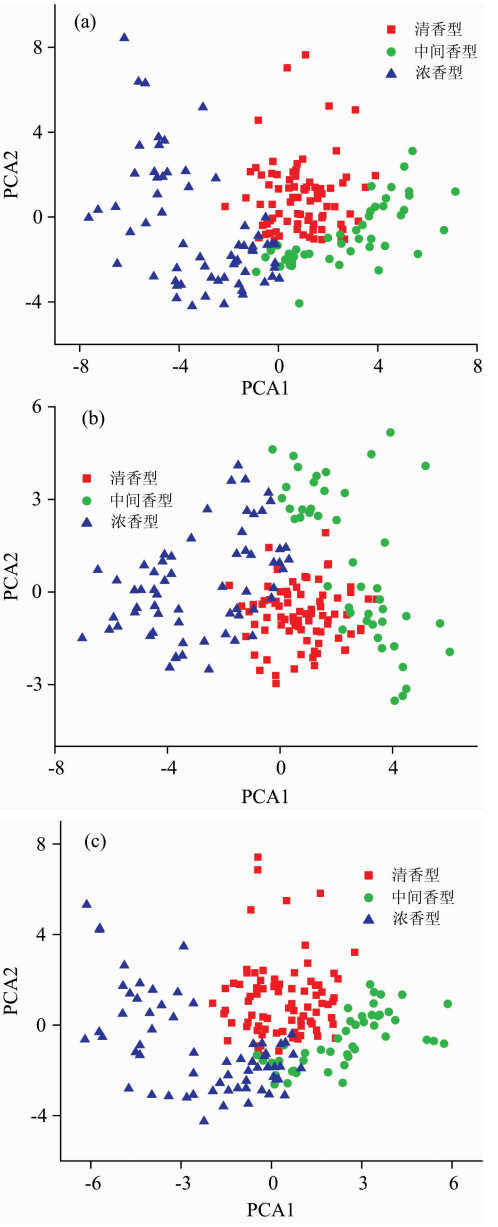

图 4　(a)基于 34 个变量的 PCA 投影图；(b)基于 24 个变量的 PCA 投影图和(c)基于 19 个变量的 PCA 投影图

Fig. 4　(a) PCA projection plot based on 34 varieties; (b) PCA projection plot based on 24 varieties and (c) PCA projection plot based on 19 varieties

由图 4 可以看出：对比全部 34 个变量、后退法选择的 24 个变量和 GA 选择的 19 变量的 PCA 算法投影图，可以看

出 GA 即使选择了比较少的变量，但三种香型风格的烟叶分类效果还比较好。因此利用 GA 对中红外和近红外融合后数据进行变量选择，可以剔除对烟叶香型分类影响小的因素。

2.4 烟叶香型分格分类判别的 SVC 模型

选取上述 189 个三种香型烟叶样本，随机提取 20%共 38 个样本作为预报集，剩余 80%共 151 个样本作为建模集，以 GA 选择的 19 个变量输入变量，建立烟叶香型风格判别的 SVC 模型，在 SVC 建模过程中选择线性核函数，惩罚因子 C 取 10。该模型的建模结果、留一法结果和预报结果如表 1 所示。

由表 1 可以看出：基于 GA 选择的中红外和近红外融合数据的 19 个变量输入变量，建立的烟叶香型风格判别的 SVC 模型，其建模结果、留一法结果和预报结果都有着比较高的准确率，整体准确率都高于 85%。

3 结 论

由于中红外和近红外融合数据提取了更多的特征信息，对于烟叶香型风格的分类效果更佳。利用 GA 算法对融合后的数据进行变量选择，删除了中红外和近红外融合数据的冗余信息，虽然选择比较少的变量，但烟叶香型风格的分类效果还较好。进一步利用以 GA 选择的变量，对 189 个三种香型烟叶样本建立烟叶香型风格判别的 SVC 模型，模型的建模结果、留一法结果和预报结果的准确率都大于 85%。以上结果表明中红外和近红外数据融合可以提取更多特征信息，利用这些信息可以建立烟叶香型风格的分类判别模型，为烟叶香型风格快速鉴别提供帮助，减少专业人员的感官评吸工作量。

表 1 SVC 模型准确率

Table 1 The accuracies of the SVC

	建模结果				留一法结果				预报结果			
	清香型	中间香型	浓香型	准确率/%	清香型	中间香型	浓香型	准确率/%	清香型	中间香型	浓香型	准确率/%
清香型	60	3	1	93.75	58	4	2	90.63	15	2	0	88.24
中间香型	2	35	1	92.11	4	33	1	86.84	1	6	0	85.71
浓香型	2	2	45	91.84	2	4	43	87.76	1	1	12	85.71
整体准确率/%			92.72				88.74				86.84	

References

[1] DING Rui-kang, WANG Cheng-han, ZHU Zun-quan(丁瑞康，王承瀚，朱尊权). Cigarette Technology(卷烟工艺学). Beijing: Food Industry Press(北京：食品工业出版社), 1958.

[2] QIU Chang-gui, KONG Lan-fen, YANG Shi-hua, et al(邱昌桂，孔兰芬，杨式华，等). Tobacco Science & Technology(烟草科技), 2019, 52(2): 101.

[3] GUO Dong-feng, YAN Ning, HU Hai-zhou, et al(郭东锋，闫 宁，胡海洲，等). Acta Agriculturae Jiangxi(江西农业学报), 2016, 28(2): 43.

[4] SONG Nan(宋 楠). Acta Tabacaria Sinica(中国烟草学报), 2015, 21(5): 16.

[5] Catauro M, Daniele N, Monica G, et al. Journal of Essential Oil Research, 2019, 31(2): 138.

[6] Vermeulen P, Fernández Pierna J A, Abbas O, et al. Food Chemistry, 2015, 189: 19.

[7] LIU Yan, LI Chang-wen, WEI Ji-ping, et al(刘 岩，李长文，魏纪平，等). Liquor-Making Science & Technology(酿酒科学), 2007, 3: 48.

[8] Dong W, Ni Y, Kokot S. Journal of Agricultural and Food Chemistry, 2013, 61(3): 540.

[9] Arman M G, Seyed H T, Emmanue M C. Journal of Geochemical Exploration, 2015, 157: 81.

[10] ZHANG Wen-jun, XU Lu(章文军，许 禄). Chinese Journal of Applied Chemistry(应用化学). 2001, 18(3): 188.

[11] Vapnik V N. The Nature of Statistical Learning Theory (Second Edition), New York: Springer-Verlag, 1999.

[12] WU Sheng-chao, LIU Tai-ang, GE Jiong, et al(吴圣超，刘太昂，葛 炯，等). Journal of Henan Normal University · Natural Science Edition(河南师范大学学报 · 自然科学版), 2018, 46(1): 77.

Merging MIR and NIR Spectral Data for Flavor Style Determination

SHA Yun-fei[1], HUANG Wen[1], WANG Liang[1], LIU Tai-ang[2], YUE Bao-hua[2], LI Min-jie[2], YOU Jing-lin[2], GE Jiong[1*], XIE Wen-yan[1*]

1. Technology Center of Shanghai Tobacco Group Co., Ltd., Shanghai 200082, China
2. Department of Chemistry, Shanghai University, Shanghai 200444, China

Abstract Tobaccos flavor type's determination is an important field tobacco industry. In this work, 189 tobacco samples with different flavor were tested by middle infrared (MIR) spectrum and near-infrared (NIR) spectrum. After the test, 21 characteristic absorption value from a certain wavelength in the MIR spectrum and 13 characteristic absorption value from a certain wavelengthin the IR spectrum were selected as main variants. Then the characteristic data extracted from MIR and IR spectrum were submitted to the principal component analysis (PCA), respectively. The PCA pattern showed a poor classification result by using MIR and IR data solely. After that, the MIR and IR variants were submitted to PCA analysis as merged data. The PCA pattern calculated from merged data showed a good classification result. Through the data analysis, there different flavor Style (fen-flavor Style, medium flavor Style and robust flavor Style) can be classified clearly into their category. After PCA analysis, different mathematical algorithms as step-back algorithm and genetic algorithm were applied to select 34 variants that used in PCA model. 24 variants and 19 variants were selected by step-back algorithms and genetic algorithms, respectively. Compared to the projection pattern by using different variant selected by a different algorithm, we found that though the genetic algorithms used the least variants, the classification result is as good as PCA algorithms and step-back algorithms. After that, genetic algorithms were chosen to make projection drawing that separated three different flavors into different planes by using least variants chosen from MIR and IR merged data. Finally, a support vector classification(SVC)model was built to determine different tobacco flavor by using the variants selected by the genetic algorithm. The accuracy of the model was 92.72%, the accuracy in discriminating fen-flavorstyle, medium flavorstyle and robust flavorstyle were 93.75%, 92.11% and 91.84%. The accuracy of predicted outputs was tested by the leave-one-out cross validation (LOOCV). And the accuracy of LOOCV was 88.24%, the accuracy in discriminating fen-flavorstyle, medium flavorstyle and robust flavorstyle were 90.63%, 86.84%, and 87.76%. The accuracy in prediction of the unknown sample was 86.84% and the accuracy in discriminating fen-flavorstyle, medium flavorstyle and robust flavorstyle were 88.24%, 85.71% and 85.71%. The results of accuracy are above 85% in model test, LOOCV teat and the prediction of unknown sample. The result shows that the mixing data from the MIR spectrum and NIR spectrum can provide more information in the mathematical model building and provide an efficient way in fast tobacco flavor discrimination.

Keywords Middle infrared spectrum; Near infrared spectrum; Tobacco flavor; Data fusion

(Received Feb. 17, 2020; accepted Jun. 23, 2020)

* Corresponding authors