第31卷, 第5期
2011年5月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 31, No. 5, pp1254-1257
May, 2011

近红外光谱的主成分分析-马氏距离聚类判别用于卷烟的真伪鉴别

张灵帅[1], 王 卫东[1], 谷运红[1], 邢 军[2*]

1. 郑州大学离子束生物工程省重点实验室, 河南 郑州 450052
2. 国家烟草质量监督检验中心, 河南 郑州 450001

摘 要 为了快速准确的鉴别卷烟真伪, 以A牌和假冒A牌卷烟为实验材料, 采用近红外光谱法结合主成分分析-马氏距离判别分析方法建立了鉴别模型。首先对经过预处理的光谱数据进行主成分分析, 分析结果表明, 前4个主成分的累积贡献率已达98.46%, 说明这4个变量能够代表原始光谱的主要信息。从120个样品中随机抽取100个用于建立4个主成分变量的定性判别模型, 模型的相关系数达到了0.95, 对20个未知样品的预测结果准确率为100%。说明近红外光谱结合模式识别方法进行卷烟真伪定性鉴别在技术上是可行的, 可以作为卷烟真伪鉴别的一种辅助手段。

关键词 近红外光谱; 卷烟; 主成分分析-马氏距离; 真伪; 鉴别
中图分类号: TS474 文献标识码: A **DOI**: 10.3964/j.issn.1000-0593(2011)05-1254-04

引 言

目前, 卷烟真伪鉴别主要应用感观鉴别法, 也就是通过观察卷烟包装、颜色、印刷以及评吸风味等来进行鉴别。由于是凭借检测人员的个人感官检验, 这需要有经验的评吸人员, 而且鉴别结果容易受评吸人员个人生理、心理各种主观因素以及检测的环境条件等的影响。如果用常规的仪器鉴别法, 则存在着检测周期长、费用高、结果重复性差等缺点, 难以满足实际工作的需要。近红外光谱(NIR)分析技术具有速度快、效率高、成本低、测试重现性好、无污染等特点, 已被广泛应用在农业[1,2]、食品[3-5]、石化[6]、纺织[7,8]、制药[9]等众多领域。在烟草行业里, 近红外光谱技术被广泛应用于烟草水分和常规化学成分的定量检测[10,11]。除此之外, 在烤烟烟叶的产地、部位、等级的模式识别以及卷烟配方研究等方面也有相关的报道[12]。但将NIR技术用于卷烟品种识别和真伪鉴别方面的研究则比较少[13]。因此, 本研究将近红外光谱技术结合主成分分析-马氏距离聚类判别方法用于国产品牌卷烟的真伪定性鉴别, 以探索近红外光谱技术在卷烟真伪鉴别方面的应用, 以期为感观鉴别法提供一种有效的技术支撑。

1 实验部分

1.1 仪器与样品

FOSS公司NIRSystems 5000型(FOSS NIRSystems Inc., MD, USA)漫反射近红外光谱仪, 石英卤灯, PbS检测器。120个A牌烟丝样品(由国家烟草质量监督检验中心提供, 为一段时间内各地送检的A牌卷烟), 其中假烟样品用常规化学方法结合感观鉴别法鉴定, 真假样品各60个。

1.2 光谱采集

在样品扫描前, 光谱仪开机预热一个小时, 进行IT(Instrument Test)测验, 测验通过后开始扫描样品, 采集光谱时, 将样品烟丝放入旋转样品槽, 使用仪器内置聚苯乙烯陶瓷片作为参比, 扫描区间1 100~2 500 nm, 扫描步长(Interval)2 nm, 采集反射强度R。仪器将对每个样品自动进行全区间扫描32次, 取其平均光谱作为该次扫描的光谱, 为了减少装样造成的误差, 每个样品做三次重复, 每次测量后将样品倒出重新装样, 以保证样品的代表性。扫描完样品后采集的反射强度由软件自动转化为$\log(1/R)$值存储为ASCII文件。取三次测量得到的光谱平均值作为该样品的光谱。

1.3 数据处理软件

主成分分析和聚类分析采用多元变量统计分析软件包Unscrambler 9.7(CAMO PROCESS AS, Oslo, Norway)完

收稿日期: 2010-12-09, 修订日期: 2011-03-20
基金项目: 国家自然科学基金项目(10505018)资助
作者简介: 张灵帅, 1981年生, 郑州大学物理工程学院博士研究生 e-mail: zhanglsl120@yahoo.cn
* 通讯联系人 e-mail: xingj@ztri.com.cn

成，主成分空间下的马氏距离判别则用光谱仪随机软件 WINISⅡV1.50 计算。

1.4　主成分分析结合马氏距离法

主成分分析常常被用来进行数据压缩，以进行数据降维[14]。在本研究中，主成分分析被用来对光谱数据矩阵进行压缩，也就是将光谱数据矩阵分解得到主成分载荷矩阵和得分矩阵，然后使用样本集的主成分对应其相应的得分作图。这样的得分图通常比较直观，可以观察到样品的聚类效果。

这样作图分析直观但步骤较繁琐，所以进一步的采用主成分分析结合马氏距离法建立判别模型，即采用光谱矩阵的得分向量来计算马氏距离，这样可以避免选择过多波长造成过度拟合，同时保留了光谱的主要信息。马氏距离的引入可以使校正集样品数据的内部变化表示出来，得到更好的分类结果。

2　结果与讨论

2.1　光谱数据预处理

120 个烟丝样品的漫反射近红外光谱如图 1 所示。在收集样品的近红外光谱过程中，许多高频随机噪音、基线漂移、信号本底、样品不均匀和光散射等噪音信息被带入光谱中，因此，在光谱分析时，对光谱数据进行预处理以滤除噪音。从图 1 中可以看出，光谱的基线都有所倾斜，因此采用一阶导数处理来提高光谱的分辨率并减小基线的漂移，光谱图经一阶导数处理后如图 2 所示。从图 2 中可以看到经过导数处理后，谱线集中较好的解决了基线漂移的问题。而为了去掉高频噪音对信号的干扰，又进行了平滑处理。此外，鉴于烟丝样品的不均一性，以及在测量过程中出现的诸如粒度不均匀、光散射等问题，在研究中，用标准正态变量转换(transformation of standard normal variate, SNV)对谱图进行散射校正处理，从而尽可能最大限度的消除各种误差。

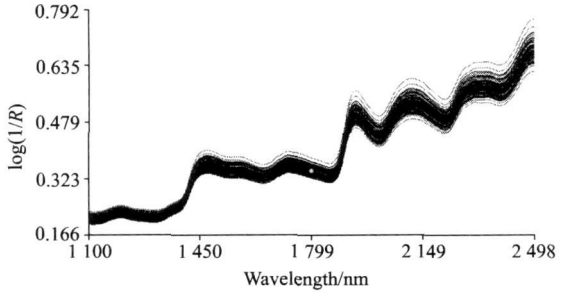

Fig. 1　Original NIR spectra of cigarette samples

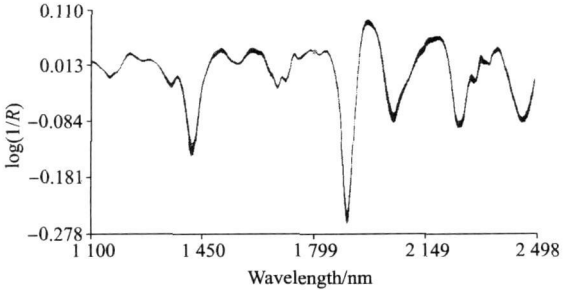

Fig. 2　First derivative NIR spectra of cigarette samples

2.2　聚类分析结果

从 120 个烟丝样品中随机抽取 100 个样品作为定标集，100 个样品中包含 50 个 A 牌烟丝和 50 个假冒 A 牌烟丝，其余 20 个作为检验集。将样品光谱数据从 ASCⅡ文件中导入到 EXCEL 中，再从 EXCEL 中导入到 Unscrambler 中。光谱数据经预处理后，进行主成分分析，聚类结果如图 3 所示。X 轴表示每个样品的第一主成分得分，Y 轴表示每个样品的第二主成分得分。从图中可以看出，前两个主成分对 A 牌卷烟烟丝和假冒 A 牌烟丝有较好的聚类效果，虽然两者没有聚合成完全分离的两团，但是界限明显，只有个别样品分散在一起。为了得到更好的分类效果，在研究中进一步将主成分分析和马氏距离结合起来建立判别模型。由于前 4 个主成分的累积贡献率达到 98.46%，也就是说前 4 个主成分能够解释原始波长变量的 98.46%，代表了样品光谱的主要信息。所以在计算样品的马氏距离时，就只选择前 4 个主成分进行计算。

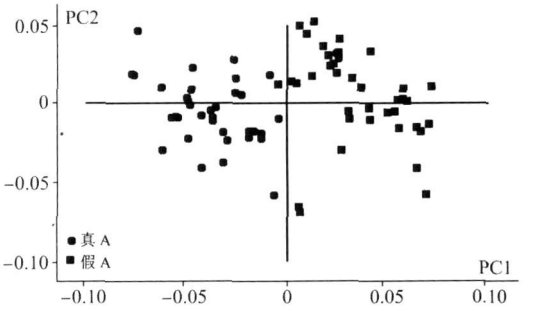

Fig 3　Clustering plot (PC1 × PC2) for 100 cigarette samples

2.3　判别模型的建立

100 个定标集样品经过光谱预处理后，选择前 4 个主成分计算样品的马氏距离，主成分空间下的马氏距离聚类散点图如图 4 所示。从图中可以看出，与主成分分析聚类结果相比，将主成分分析与马氏距离结合后聚类得到了更好的分类结果，100 个定标集样品分成明显的分成两类，两类之间的区域没有重叠，得到的定标方程相关系数为 0.95，说明所建立的判别模型是可靠的。

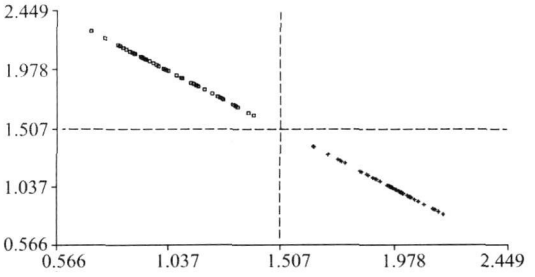

Fig 4　Clustering plot of PCA-Mahalanobis distance for 100 cigarette samples

为了检验所建立判别模型的可靠程度，用 20 个未参与定标的检验集样品对模型进行预测验证。20 个未知样品的

预测结果显示在图 5 中。从图 5 中可以看出，20 个未知样品被准确的分成了两类，10 个 A 牌烟丝和 10 个假冒 A 牌烟丝。这表明模型对未知样品的预测准确率达到了 100%。

这种建立定性判别模型的方法除了在 A 牌烟丝上应用外，在其他多个国产品牌烟丝上进行真伪定性鉴别也都取得了良好的分类效果。

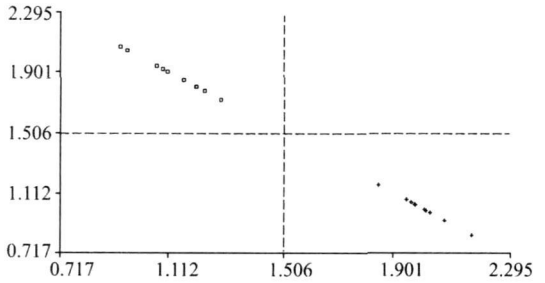

Fig 5 Prediction results of 20 unknown cigarette samples

3 结 语

与常规的理化仪器分析和感官鉴别法相比，近红外光谱技术进行卷烟真伪定性具有速度快、操作简便、样品预处理简单、结果重复性强等特点，其能进行定性鉴别的核心是基于烟草内在品质的差异。在本研究中，获取了 60 个 A 牌烟丝和 60 个假冒 A 牌烟丝样品的近红外光谱，并通过样品的光谱特征结合主成分分析和马氏距离的模式识别方法建立了定性判别模型，模型的相关系数为 0 95，对未知样品的鉴别率达到了 100%。实验结果表明利用近红外光谱技术进行卷烟真伪定性鉴别是可行的，有较好的应用前景，可作为判断卷烟真伪的一种辅助手段，对于提高鉴别的准确性和效率都有一定的帮助。

References

[1] HAN Liang-liang, MAO Pei-sheng, WANG Xin-guo, et al(韩亮亮, 毛培胜, 王新国, 等). Journal of Infrared Millimeter Waves(红外与毫米波学报), 2008, 27(2): 86.
[2] Mahesh S, Manickavasagan A, Jayas D S, et al. Biosystems Engineering, 2008, 101: 50.
[3] Del Moral F G, Guillén A, Del Moral L G, et al. Journal of Food Engineering, 2009, 90(4): 540.
[4] Alessandrini L, Romani S, Pinnavaia G, et al. Analytica Chimica Acta, 2008, 625(1): 95.
[5] LIU De-yan, LUO Ji, CHEN Xing-miao(刘德燕, 罗 吉, 陈兴苗). Journal of Infrared Millimeter Waves(红外与毫米波学报), 2008, 27(2): 119.
[6] WANG Li, HE Ying, WANG Yan-ping, et al(王 丽, 何 鹰, 王颜萍, 等). Marine Environmental Science(海洋环境科学), 2004, 23(2): 58.
[7] Blttner A, Marbach R, Heise H M. Journal of Molecular Strueture, 1995, 349.
[8] Andreev G N, Schrader B, Schulz H, et al. Analytical Chemistry, 2001, 371: 1009.
[9] Chen Yi, Xie Mingyong, Yan Yan, et al. Analytica Chimica Acta, 2008, 618(2): 121.
[10] Dane A D, Rea G J, Walmsley A D, et al. Analytica Chimica Acta, 2001, 429(2): 185.
[11] Zhang Yong, Cong Qian, Xie Yunfei, et al. Spectrochimica Acta Part A, 2008, 71(4): 1408.
[12] ZHANG Jian-ping, CHEN Jiang-hua, SHU Ru-xin, et al(张建平, 陈江华, 束茹欣, 等). China Tobacco Acta(中国烟草学报), 2007, 13(5): 1.
[13] Shao Yongni, He Yong, Wang Yanyan. European Food Research and Technology, 2007, 224: 591.
[14] Pizarro C. Analytica Chimica Acta, 2004, 509(1): 217.

Identification of Authentic and Fake Cigarettes Using Near Infrared Spectroscopy Combined with Principal Component Analysis-Mahalanobis Distance

ZHANG Ling-shuai[1], WANG Wei-dong[1], GU Yun-hong[1], XING Jun[2*]
1. Henan Province Key Laboratory of Ion Beam Bio-engineering, Zhengzhou University, Zhengzhou 450052, China
2. China National Tobacco Quality Supervision & Test Centre, Zhengzhou 450001, China

Abstract In order to discriminate fake and genuine cigarettes correctly and rapidly, cigarettes of brand A and fake cigarettes of brand A were scanned by the NIR spectrometer, and an identifying model was developed by near infrared spectroscopy combined with principal component-Mahalanobis distance pattern recognition method. The pretreated spectra data of cigarette samples

were analyzed through principal component analysis (PCA), and the result of the analysis suggested that the accumulation of first 4 principal components was more than 97.46%. One hundred samples from total 120 cigarette samples were selected randomly. Then they were used to build qualitative discriminating model and the correlation coefficient was 0.95. Twenty unknown samples were validated by this model. The recognition rate is 100%. The model is reliable and practicable, and could be used as an assistant means for identifying fake and genuine cigarettes.

Keywords Near infrared spectroscopy; Cigarette; Principal component-mahalanobis distance; Identification; Authentic and fake

(Received Dec. 9, 2010; accepted Mar. 20, 2011)

* Corresponding author

《光谱学与光谱分析》期刊社决定采用 ScholarOne Manuscripts 在线投稿审稿系统

《光谱学与光谱分析》期刊社与汤森路透集团签约，自 2010 年 12 月 1 日起《光谱学与光谱分析》决定采用 Thomson Reuters旗下的 ScholarOne Manuscripts 在线投稿审稿系统！

• ScholarOne Manuscripts，该系统不仅能轻松处理稿件，而且能提速科技交流。

• 全球已有 360 多家学会和出版社的 3 800 多种期刊选用了 ScholarOne Manuscripts 系统作为在线投稿、审稿平台，全球拥有超过 1 350 万的注册用户，代表着全球学术期刊在线投审稿的一流水平。

• ScholarOne Manuscripts 与 EndNote，Web of Science 无缝链接和整合；使科研探索、论文评阅和信息传播效率大为提高。

• ScholarOne Manuscripts 是汤森路透科技集团的一个业务部门，拥有丰富的学术期刊业务经验，为学术期刊提供综合管理工作流程系统，使期刊更有效管理投稿、同行评审、加工和发表过程，提高作者心中的专业形象，缩短论文发表时间，削减管理成本，帮助期刊提高科研绩效和实现学术创新。

《光谱学与光谱分析》采用“全球学术期刊首选的在线投稿审稿系统—ScholarOne Manuscripts”，势必对 2010 年 11 月 30 日以前向本刊投稿的作者在查阅稿件信息时，会带来某些不便，在此深表歉意！为了推进本刊的网络化、数字化、国际化进程，以实现与国际先进出版系统对接；为了不断提高期刊质量，加快网络化、数字化建设，加快与国际接轨的进程，希望能得到广大作者、读者们的支持与理解，对您的理解和配合深表感激。这是一件新事物，肯定有不周全、不完善的地方，让我们共同努力，不断改进和完善起来。

《光谱学与光谱分析》期刊社

2010 年 12 月 1 日