工　　业

近红外信息用于烟叶风格识别及卷烟配方研究的初步探索

张建平[1]，陈江华[2]，束茹欣[1]，刘建利[2]，杨　凯[1]

1 上海烟草(集团)公司技术中心，上海长阳路 717 号 D 楼 403　200081；

2 中国烟草总公司烟叶公司，北京　100055

摘　要：以烟草特征风格的数字化鉴别为最终研究目标，使用近红外分析技术，通过烟叶产地识别的可行性研究，可知近红外光谱中具有与烟叶产地相关的信息，应用近红外进行产地识别具有技术可行性。通过烟叶近红外光谱的压缩，基于烟叶风格特征的量化分析，结合专家经验，设定特定的目标烟叶，以国内外烟叶作为可选资源，选择部分烟叶进行配伍模拟目标烟叶，通过三点法感官评吸和叶组配方试验，验证表明模拟样品具有目标烟叶的风格特征。研究结果初步表明应用近红外信息可望进行烟叶产地特征的数字化研究，并为卷烟配方提供技术辅助。

关键词：近红外；烟叶风格；模式识别；感官评吸

中图分类号：TS452.1　　**文献标识码**：A　　**文章编号**：1004－5708(2007)05－0001－05

Tobacco characteristics identification and blending formula study by using NIRs

ZHANG Jian-ping[1], CHEN Jiang-hua[2], SHU Ru-xin[1], LIU Jian-li[2], YANG Kai[1]

1 Technology Center of Shanghai Tobacco (Group) Corporation, Shanghai 200081, China;

2 China National Leaf Tobacco Corporation, Beijing 100055, China

Abstract: Characteristics of tobacco leaves were studied by near infrared spectra (NIRs) information and pattern recognition technology. Results showed that NIRs indicated information about growing area of the leaves. Based on NIRs data and quantified analysis of tobacco flavor style, tobacco leaves from home and abroad were selected to simulate Zimbabwe tobacco leaf. According to three points sensory evaluation the mixed samples were similar to Zimbabwe tobacco leaf. It was proved that NIRs can be used to identify tobacco growing area and improve blending formulation.

Key words: near infrared spectra; tobacco style; pattern recognition; sensory evaluation

近红外检测技术已比较普遍地应用于烟叶中的烟碱、总糖、还原糖、总氮、钾、氯含量等常规化学成分的检测。其原理是由于近红外光对C－H、N－H、O－H等含氢基团有吸收，据理论推断，烟叶中多达 80% ~ 90%以上的化学成分是可能应用近红外技术进行研究和检测的，近红外光谱所包含的烟草化学成分的关联信息非常丰富，基于近红外信息进行烟叶聚类分析和模式识别具有可靠的物质基础，应用近红外信息进行烟叶质量的定性定量研究更具有广阔的应用前景。

不同地区烟叶具有不同的香气特征，这是烤烟型卷烟设计及烟叶选择使用的重要依据，但由于烟叶和烟气成分的复杂性，烟草香气类型的判断还只能依靠感官评吸这种经验方法。如果应用近红外光谱信息可识别不同产地的烟叶，那就说明有望在烟草近红外光谱信息中捕捉到与地区风格特征具有紧密关系的数据信息用于描述不同烟叶的风格特征，实现烟叶风格特征的客观描述和识别，从而在技术上支持和辅助卷烟

作者简介：张建平(1965—)，男，研究员，上海烟草(集团)公司技术中心。Tel：021-61669404，E-mail：zjp21@sohu.com

基金项目：国家烟草专卖局基金资助项目“部分替代进口烟叶生产示范及工业验证”中的部分内容。

收稿日期：2006-10-09

的配方设计。根据上述思路,本文顺着烟叶产地的识别、特征数据的抽提、特定目标烟叶的数字化模拟、目标和模拟烟叶的香气风格评吸比较这条技术路线,对其可行性进行了探索研究。

1 材料和方法

1.1 样品材料

2004 年烤烟样品共 545 个,分别来自巴西、津巴布韦和湖南、湖北、云南、贵州、四川、重庆、广东、河南、安徽、福建、陕西、黑龙江、山东、吉林、辽宁等 15 个地区。其中 480 个建立成为建模集,65 个建立成为预测集(每个集合中的样品均为随机抽取)。

1.2 样品的测试

将烤烟烟叶磨成粉,在低于 50 ℃下烘 0.5 h,使含水率达到 10% ~ 12%。过 60 目筛,取适量烟末放入样品杯中,取一固定重量砝码放置在样品上方,使其自然压实后进行近红外光谱扫描,近红外光谱仪分辨率设定为 8 cm^{-1},扫描次数 64。

1.3 光谱处理与模式识别

对烟叶样本的近红外光谱曲线(NIRs)求一阶导数后进行光滑处理,采用主成分分析(PCA)法进行特征抽提,根据主成分空间下的马氏距离建模并判别样本的产地归属 根据主成分得分向量 Ti 描述的 2 个样本 i、j 间的马氏距离[1],用被测样本距各个类中心(该类所有建模样本的主成分得分的平均值)的马氏距离进行判断。

2 研究与结果

2.1 产地模式识别试验与结果

2.1.1 NIRs 的特征抽提与特征个数的确定

采用 PCA 法对 2004 年建模集烟叶样品的 NIRs 进行特征抽提(图 1)。

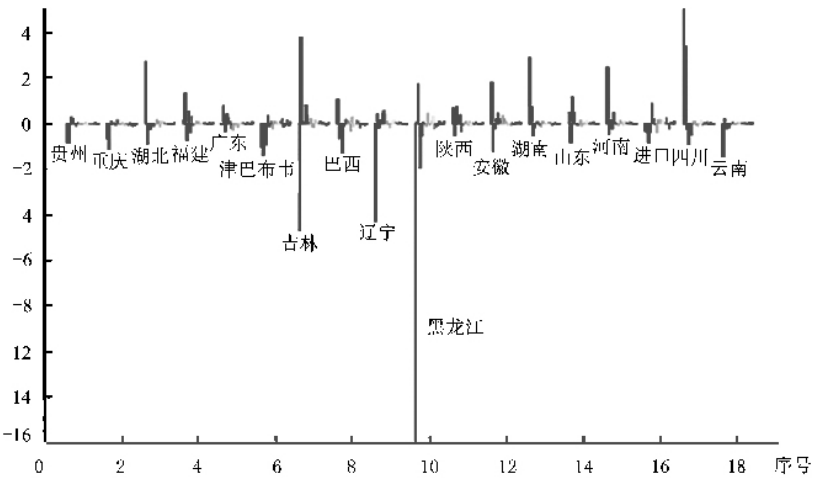

图 1 2004 年烤烟一阶导数 NIRs 的前 10 个主成分得分柱状图

结果表明 不同产地烟叶的主成分有较明显的差异,可用烟叶 NIRs 的主成分来区分烟叶的产地。然而,取多少个主成分合适?商品化软件中一般不能如定量分析那样根据交叉验证结果提供最佳主成分数。采用留一交叉验证法考察预测准确率与主成分个数的关系时发现,预测准确率随主成分个数的增大而增大,到后期呈锯齿状跳动[2]。主成分数过多会引入仪器噪声和误差,但如果太少则信息量不足,难以完全体现不同产地的风格特征[3]。因此需要根据近红外仪的噪声、误差水平来合理选择主成分个数。本研究采用前 m 个主成分复原光谱的残差矩阵来辅助选择。

由误差分析原理可知,主成分个数与仪器噪声、误差水平相关[4],因此,若测量方法的误差大小可以估计,则可将 RSD(复原残差的标准差)[5] 的计算值与测量误差(RE)比较,二者接近时的 m 值就是主成分数。RE 与样品关系很大,在实际中较难预先确定,一般都是根据经验或多次测量结果的评价分析来判断测量误差。本研究采取如下方法来确定 RE :多次重复测定某一烟草标样的光谱,对于 r 个重复测量的光谱,按照相同的方法进行预处理得矩阵 $X_{n\times r}$,用这些重复性检测

光谱的标准差来代替测量误差 RE。

一般选取 RSD 为 RE 的 1 ~ 2 倍。这样可确定主成分个数的上限 m_{max}，在 $m < m_{max}$ 的前提下，采用不同个数主成分对建模样本进行模式识别，取识别正确率最高的主成分个数建模和预测[6]。如此，既能取得足够的有效信息，获得较好的预测效果，而且能够防止引入过多主成分导致噪声的引入及过拟和现象（即建模结果很好但预测结果很差）的产生。

2.1.2　烟叶产地判别分析

2 种马氏距离判别法给出的平均准确率见表 1 所示。由表 1 可知，本研究提出的方法可有效识别烟叶产地，采用二阶导数光谱的主成分得分进行烟叶产地识别可获得最高准确率。在相同光谱类型下，MDTS（在采用建模样本确定的载荷轴下进行的马氏距离判别）法的平均预测准确率及整体平均准确率均高于 MDAS（在采用 PCA 法确定的所有样本的载荷轴下进行的马氏距离判别）[7]。而采用 MDTS 法对 Fi > 1（类间方差大于类内方差）的光谱区间进行烟叶产地识别时，无论是平均预测准确率还是整体平均准确率均低于全光谱分析，说明类间方差低于类内方差的光谱信息对于烟叶产地模式识别也是有用的，不能只考虑 Fi > 1 的光谱区间。

表 1　不同光谱预处理方法对烟叶产地识别的结果

判别方法	光谱处理								
	原始光谱			SNV 光谱			二阶导数光谱		
	AAP①	WAA②	$L_{max}/\bar{L}$③	AAP	WAA	$L_{max}/\bar{L}$	AAP	WAA	$L_{max}/\bar{L}$
MDTS	88.80%	84.00%	5/5	93.85%	95.84%	15/11	97.69%	97.12%	15/12.6
MDAS	88.80%	80.16%	5/5	92.31%	93.92%	15/12.6	91.54%	94.40%	15/14

注：①AAP 表示 5 个随机分组样本集下对预测集样本识别的准确率的平均值，即平均预测准确率；②WAA 表示对建模及预测集样本（所有样本）识别的准确率的平均值，即整体平均准确率；③为 5 组样本集下得到的最佳主成分个数的平均值。

对 480 个烤烟样品进行建模，用二阶导数作为光谱预处理方法。模型内部的整体预测正确率达到 94.7%，还有 5.3% 的样品被误判，判断错误的样品列于表 2。

预测是根据马氏距离的大小来进行判别的，分析错判的烟叶，发现以下现象：(1) 错判往往发生在相邻省之间，即将一个省的烟叶判入相邻的另外一个省；(2) 按最近距离判入邻省，但按次近距离又判入本省，说明相邻省之间的烟叶往往相似性比较高，容易发生互为误判的情况。从烟叶的行政区划看这些是错判的结果，分析这种现象：第一种可能是由于距离相近，邻省间相邻地区的烟叶风格比较接近，这是由于相邻地区间差异较小而导致的识别误差，另外一种可能是虽然行政上为某个省的烟叶、但其风格可能确实与相邻的另一个省更为接近，从而出现看起来是错误的、实际上是正确的判断。这从另一个角度给我们一个重要启示：这种方法识别所依据的信息与烟叶质量风格特征有关系，而且能识别更为细微的差别，这正是所希望得到的结果。

表 2　480 个建模样品中误判情况分析

实际类别	预测最近结果	距离	预测次近结果	距离
广东	湖南	0.62	广东	0.67
贵州	重庆	0.68	贵州	0.89
贵州	云南	2.2	贵州	2.28
贵州	广东	1.29	贵州	1.50
河南	广东	1.05	河南	1.23
河南	陕西	1.78	河南	2.08
湖北	贵州	0.91	云南	1.14
湖北	河南	1.30	贵州	1.32
湖南	湖北	0.78	湖南	0.79
湖南	广东	1.31	湖南	1.39
湖南	云南	0.77	湖南	0.85
湖南	云南	0.68	湖南	0.69
津巴布韦	巴西	3.97	津巴布韦	4.13
辽宁	吉林	2.05	辽宁	2.16
四川	湖北	1.31	四川	1.44
四川	湖北	1.57	湖南	1.70
云南	贵州	1.12	云南	1.21
云南	贵州	1.61	云南	1.69
云南	贵州	1.09	云南	1.18
云南	四川	1.27	云南	1.32
云南	陕西	1.10	云南	1.11
云南	贵州	0.76	云南	0.78
重庆	贵州	1.10	重庆	1.16
重庆	陕西	0.85	贵州	0.99
重庆	贵州	0.93	重庆	0.97

2.2 烟叶模拟的探索性研究

综合分析这些正确判别和误判的情况,可以非常清晰地看出,在近红外光谱信息中,使用现代数据分析方法,可以挖掘出与产地风格相关联的有关信息。

在烟叶产地成功识别的基础上,我们继续进行烟叶的辅助配方技术研究和试验,研究思路是:(1)根据烟叶产地识别所依据的重要信息,进行不同烟叶之间的差异性观测和分析;(2)使用这些数据信息对模拟目标和模拟结果之间进行一致性分析和控制;(3)根据比较满意的数据模拟结果,配制出相关模拟样品,再采用经验方法进行验证,观察数据模拟的有效性。

2.2.1 单等级烟叶的模拟研究

2.2.1.1 评吸验证

以 2004 年津巴布韦烟叶 L10T 为目标,以 2004 年国内 627 个烟叶样本及国外 2004 年(来自巴西、南非及加拿大)的 9 个进口烟叶样本为可选原料,通过数字化分析计算得到与目标样品数字化质量特征比较接近的配方比例,然后配制模拟的混合样品,以三点法感官评吸,由评吸专家检验模拟样品与目标样品的特征风格一致性,结果见表 3。

表 3 以三点法对津巴布韦及模拟烟叶的评吸结果

评吸专家	津巴布韦 L10T(模拟)		
	样品 1(模拟)	样品 2 (L10T)	样品 3 (L10T)
1 号			
2 号		√	√
3 号		√	√
4 号		√	√
5 号			
6 号	√		√
7 号			

注:将认为风格相同的样品打√,认为 3 个样品很接近或没有区别则不打√

总体评吸评价:7 名专家中 3 名认为无区别,1 名识别错误(统计为无区别),3 名可识别,评吸专家总体认为模拟样品有一定的津巴布韦烟叶的特征,在香气、刺激和余味方面与目标样品均较为接近,评吸意见表明目标样品和模拟样品的整体风格具有一定的相似性。但是在烟气的成团性和劲头方面与目标样品还略有差异,需要作进一步的调整。

2.2.1.2 评吸结果的经验分析

津巴布韦烟叶具有比较明显香气风格,一般有经验的配方人员或评吸专家都能比较敏感地体会或识别这种差异。

对评吸数据的理解:一般对于我们所试验的对象,如果全部人员认为不相似而且都能轻易分辨出来,那就非常清晰地说明模拟是失败的,但如果逐渐开始有人认为相似了,说明相似性在提高。高水平专家的辨别能力非常高,很小的差异均能识别出来,这时就要看他的观点:是否具有类似的风格?

在上述试验中,完全用津巴布韦以外的烟叶,所模拟制作的混合模拟样品,7 名专家中 3 名认为无区别,1 名识别错误,3 名虽能识别但认为模拟样品已经具有一定的津巴布韦烟叶的风格特征,在没有任何香精香料而纯粹依靠烟叶的配伍,达到这样的效果,这说明在风格方面的模拟已具有可喜的效果,从专业工作经验的角度初步判断,分析所使用的数字化信息确实抓住了一些能一定程度反映烟叶香气风格方面的重要信息,是一个有价值的技术现象。

2.2.1.3 评吸结果的统计检验

上述是依靠经验的分析结果,那么从数据统计分析的角度分析上述数据,又能说明什么问题呢?

分析评吸专家对目标和模拟样品的评吸数据(评吸数据略),对这些数据进行 I 类风险检验所得 α 的临界值如下。

表 4 拒绝 H_0 的 α 临界值表

	等级	香气质	香气量	杂气	劲头	刺激性	余味	总体
α 临界值	L10T	0.92	0.96	0.96	0.26	0.92	0.92	0.64

由上表可知,津巴布韦烟叶与模拟样本的 6 个评吸指标的统计临界 α 值均远大于 0.05,即要拒绝模拟烟叶与目标烟叶是同一类所冒的总体风险概率是 64%,表明模拟组与实际津巴布韦烟叶组之间无显著性差异,即从统计意义上讲,可认为模拟烟叶与目标烟叶总体上较为一致。

从表 4 进一步可看出,L10T 的模拟烟叶与目标烟叶总体较为接近:其总体 α 值为 0.64,用通俗的话解释意味着 100 个评吸专家评吸该组烟叶时,有 64 人不能区分两者的差异;其中劲头的临界值为 0.26,意味着

100 个评吸人员中只有 26 人不能区分两者的差异，这与评吸时大部分人感觉目标烟叶与模拟烟叶在劲头方面的差异较明显的结果一致。

通过经验分析和统计分析，均表明模拟烟叶和目标烟叶的风格比较接近。

2.2.2　叶组配方的模拟研究

上述仅仅是目标样品和模拟样品一对一的比较分析，由于目标样品是单一烟叶，模拟样品是混合烟叶，特别敏感的评吸专家仅仅通过这一特征就能分辨其差别。为了避免这种一对一比较的局限性，同时观察模拟样品加入到正常配方以后是否还具有我们所期望的效果？我们又进行了叶组配方的模拟试验，将上述模拟样品用于某产品配方，用来替代该配方中所使用的津巴布韦 L10T 烟叶，所有制作工艺与参数不变。

2.2.2.1　模拟配方的小样试验

组织 7 位评吸专家对目标和模拟小样样品进行评吸，结果如下：

表 5　某产品中津巴布韦模拟小样评吸得分统计

样品编号	光泽	香气	谐调	杂气	刺激性	余味	合计
模拟样品	5.0	32.6	5.0	14.3	14.3	17.0	88.2
正常样品	5.0	32.6	5.0	14.2	14.2	17.0	88.0

总体评吸评价：模拟后的卷烟样品与原配方卷烟样品风格接近，劲头近似，烟气更为柔和，认为可以进行中样试验。

2.2.2.2　配方模拟的中样试验

组织 7 位评吸专家对目标和模拟中样样品进行评吸，结果如下：

表 6　某产品中津巴布韦模拟中样评吸得分统计

样品编号	光泽	香气	谐调	杂气	刺激性	余味	合计
模拟样品	5.0	33.0	5.0	14.2	14.2	17.0	88.4
正常样品	5.0	32.8	5.0	14.3	14.1	17.0	88.1

总体评吸评价：模拟后的卷烟样品与原配方卷烟样品风格接近，烟气细腻柔和。

上述两方面的试验表明，模拟烟叶无论一对一的比较，还是配方试验中，初步体现了模拟烟叶与目标烟叶之间具有一定的风格相似性。

3　结果讨论

研究表明，采用近红外数字化信息作为配方基础，利用津巴布韦以外的国内外烟叶，可以配置出初步具有津巴布韦风格特征的烟叶；将模拟烟叶用到有关的卷烟配方中，小样和中样试验表明，与直接使用津巴布韦烟叶的产品比较接近，专家评吸认为风格比较相似。证实了采用近红外光谱信息有效模拟出了初步具有津巴布韦烟叶风格特征的烟叶，说明了这种方法辅助烟叶配方设计具有理论可行性。采用该方法与感官评吸相结合，有望为传统经验方法提供一种有力的技术支持手段，从而使烟叶配方更加科学客观。

目前为止所得到的只是理论上可行的初步结论。在研究过程中我们发现，由于不同烟叶之间的差异较小，而同一烟叶不同仪器扫描的结果、甚至同一烟叶同一仪器不同时间扫描的差异往往大于烟叶之间的差异；另外如要进一步研究和使用，同一地区、同一等级烟叶的数字化信息也会在一定范围内变化，烟叶的同一性和差异性又如何表示？等等，这些问题还有待进一步探索。这些问题的关键和核心是如何有效获得真正的规律性特征、去除随机性和非本质的差异，并在此基础上构筑基于近红外信息的数字化特征描述及这些特征与感官质量之间的关系和模型，这是本研究下一步所要攻克的重点和难点。

参考文献

[1] Maha Hana，Mcclure W F. Applying artificial neural networks. II. Using near infrared data to classify tobacco types and identify native grown tobacco [J]. J Near Infrared Spectra，1997(5)：19-25.

[2] 梁逸曾，俞汝勤. 分析化学手册. 第十分册[M]. 2 版. 北京：化学工业出版社，2000：328-329.

[3] 王芳，陈达，邵学广. 近红外谱与卷烟样品常规成分的关系模型研究[J]. 烟草科技，2002(5)：23-26.

[4] Candolfi A，Maesschalck R De，Jouan-Rimbaud D. The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra[J]. J Pharma & Biom Anal，1999，21：115-132.

[5] 王国东，束茹欣，张建平，等. 不同产地国产烤烟近红外光谱的特征分析及其模式识别[J]. 烟草科技，2006(5)：36-40.

[6] Savitzky A，Golay M J E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures[J]. Anal Chem 1994，36：1627-1639.

[7] 束茹欣，王国东，张建平，等. 国产烤烟烟叶的 NIRS 模式识别[J]. 烟草科技，2006(8)：12-15.