

基于 FT-NIR 分析技术的 SIMCA 建模及其在卷烟配方过程质量监测中的应用

王家俊, 李娟

红河烟草(集团)有限责任公司, 云南省弥勒县桃园路 50 号 652300

关键词: FT-NIR 分析技术; SIMCA 建模; MSPC; 卷烟配方; 过程监测

摘要: 针对传统卷烟配方过程质量控制主要以物理质量监测的不足, 基于 FT-NIR 分析技术, 结合 SIMCA 统计建模方法和多元统计过程控制(MSPC)的基本原理, 对卷烟制丝配方等过程质量监测进行了探讨。结果表明, 通过建立 SIMCA 过程类模型对卷烟配方过程质量进行多元统计过程控制, 进而实现对配方模块宏观特性的均匀性和稳定性监测, 方法简单, 操作容易, 具有一定的实用性。

中图分类号: TS452.1 **文献标识码:** B **文章编号:** 1002-0861(2008)03-0005-05

SIMCA Modeling Based on FT-NIR Analysis Technology and its Application in Quality Monitoring in Cigarette Blending Process

WANG JIA-JUN and LI JUAN

Honghe Tobacco (Group) Co., Ltd., Mile 652300, Yunnan, China

Keywords: FT-NIR analysis technology; SIMCA modeling; MSPC; Cigarette blending; Process monitoring

Abstract: Aiming at the shortcomings of traditional quality control in cigarette blending process focused mainly on physical quality monitoring, the process monitoring of cigarette blending based on FT-NIR analysis technology, the method of SIMCA statistical modeling and the basic principle of multivariate statistical process control (MSPC) was discussed. The results showed that: cigarette blending process control and further, monitoring the uniformity and stability of general characteristics of blending blocks implemented with multivariate statistical process through establishing SIMCA process model was simple, practical and ease of operation

卷烟制造是在一定的时间内,按预先设计、优化好的工序,将不同品质特性的配方模块组成预期质量要求的产品。因此,它是一个多品种、多批次配方模块组成的间歇过程(Batch processes),配方模块的稳定性和均匀性对产品质量有着重要的影响。目前国内卷烟企业对卷烟配方过程质量的监测主要以物理质量(如填充值、含水率和烟丝宽度等)为主,如欲较为全面地监测理化综合信息表征的配方模块特征质量变异,则存在一定的局限性。

FT-NIR 光谱分析是一种环境友好的绿色快速分析技术,一张近红外光谱包含着样品丰富的理化信息,应用现代化学计量学中的多元校正、模式识别与近红外光谱分析技术相结合等方法,不仅可预测未知样品的多个组分,而且可以对其品质属性进行分类,目前在石化、农业、烟草、食品和医药等领域^[1-2]已得到了广泛应用。近年来,该技术也逐渐被国内烟草行业应用于原料、辅料的质量控制和卷烟品质检测等方面^[3-14],并且已将近红外光谱分析技术与化学计量学中的多变量分析方法(偏最小二乘法、主成分分析-马氏距离法)相结合,应用于过程质量的表征与诊断^[15]。本文针对传统卷烟配方过程质量监测的不足和卷烟制造过程间歇性的特点,应用 FT-NIR 分析技术结合 SIMCA 建模方法^[16-17]和 MSPC (Multivariate statistical process control)的基本原理^[18-19],对卷烟制丝配方等过程质量监

作者简介:王家俊(1962-),学士,红河烟草(集团)有限责任公司技术中心工程师,主要从事化学计量学方法与光谱分析技术的应用研究。E-mail: honghe@vip.163.com

收稿日期: 2007-09-17

责任编辑: 石永新 E-mail: syx@tobaccoinfo.com.cn

电话: 0371-67672659

测进行探讨,同时对 SIMCA 建模的最佳谱区(变量)选择进行讨论。

1 SIMCA 建模的基本原理和步骤

SIMCA (Soft independent modeling class analogy) 是由瑞典化学家 Wold 于 1976 年提出的一种基于主成分分析(PCA, Principal component analysis)的分类方法,即对每一类建立一个类模型,然后应用这些类模型来判别未知样品的归属。本文基于 FT-NIR 光谱数据,应用 SIMCA 建模,其基本原理和步骤为:在过程正常的状态(即只存在偶因没有异因的工况)下,按一定采样周期,应用 FT-NIR 光谱分析仪采集不同批次、不同品质特征配方模块的近红外光谱数据,即可构成样本的光谱数据集。对任意一个样本可用矢量表示为:

$$x_i = [x_{i1} \ x_{i2} \ \dots \ x_{in}]$$

那么,对 m 个样本构成的样本集可表达为 $m \times n$ 阶的光谱数据矩阵:

$$X_{m \times n} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \dots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

通过主成分分析,建立各个类(如第 q 类)的主成分分析模型为:

$$X_q = \bar{X}_q + T_q P_q^T + E_q$$

式中: \bar{X}_q ——第 q 类经中心化处理后所得的矩阵; T_q ——得分矩阵; P_q ——载荷矩阵; E_q ——残差矩阵。

对矩阵 X_q 中的每一个测量值 x_{ij} ,主成分分析可表达为:

$$x_{ij}^q = \bar{x}_j^q + \sum_{a=1}^{A_q} t_{ia}^q p_{aj}^q + e_{ij}^q$$

式中: \bar{x}_j^q ——第 q 类变量 j 的平均值; A_q ——第 q 类中的显著主成分数; t_{ia}^q ——第 q 类中样本 i 在第 a 个主成分上的得分值; p_{aj}^q 为第 q 类中变量 j 在第 a 个主成分上的载荷值; e_{ij}^q ——第 q 类中样本 i 的变量 j 的残差值。

在 SIMCA 分类中,第 q 类样本到主成分模型的距离,常用其残余标准偏差 s_i 表示,即残余方差为:

$$s_i^2 = \sum_{j=1}^n \frac{(e_{ij}^q)^2}{n - A_q}$$

对于整个 q 类,其总残余方差为:

$$s^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(e_{ij}^q)^2}{(m - A_q - 1)(n - A_q)}$$

以上两式中: s_i^2 、 s^2 ——分别为样本 i 的残余方差和整个 q 类的总残余方差; m ——样本数量; n ——变量,即维数; $(n - A_q)$ 、 $(m - A_q - 1)(n - A_q)$ 分别为第 q 类引入主成分数 A_q 后 s_i^2 和 s^2 所具有的自由度。

于是,就可应用 F 检验确定数据点是否为异常值,若 s_i^2 与 s^2 没有显著差异,则样本 i 归属该类;反之,样本 i 远离模型,就是异常值。应用 SIMCA 类模型预测未知样本 g 的归属,也就是判别未知样本 g 与某类主成分模型的相似程度,此时,引入 F 检验统计量,记为 F_{crit} ,用于获得一个判定某样本是否属于某个类(如类 q)的残余方差的上限值: $s_{maxq}^2 = s_q^2 F_{crit}$ 。 F 检验的显著性水平(α)一般设为 5% 或 1%,若 s_g^2 比 s_{maxq}^2 小,则未知样本 g 归属第 q 类,否则未知样本 g 属于其它类。

将 MSPC 技术应用于过程质量监测,主要是对过程正常状态下采集的大量过程变量进行主成分分析(PCA),将大量过程变量映射到由少量具有代表性的特征隐变量定义的低维空间,使数据降维,消除变量间多重相关性造成的信息重叠,从而建立多元统计模型(如 PCA、PLS 模型),并采用 Hotelling T^2 、 Q 等统计量建立 MSPC 控制图,从而对过程质量进行监测^[18-19]。本文结合 MSPC 的基本原理及其思路,针对卷烟制造过程具有间歇性的特点,通过采集正常状态下不同批次不同品质特征配方模块的近红外光谱数据,应用 SIMCA 方法构造各个配方模块的主成分模型。然后,通过设置合适的置信度,即可用类模型监测出未知过程样本的距离,进而判别出样本的质量特性归属。同时,通过置信度和样本到类模型的距离等统计量建立 MSPC 控制图,以监测在加工或配方过程中配方模块总体质量特性的稳定性和均匀性等变化趋势。

2 材料与方 法

2.1 主要仪器与实验样本的准备

Spectrum One NTS FT-NIR 光谱仪(包括带 InGaAs 检测器的漫反射积分球,9 cm 石英采样杯、旋转台等附件,美国 PE 公司);Pirouette v3.11 和 InStep v2.11 等数据处理软件(美国 Infometrix 公司)。

实验样本包括训练集和验证集,是在红河卷烟总厂三级配方工艺过程(二级配方环节)正常状态下,按时序采集的 3 个配方模块(B、P 和 X)烟丝,其组成情况见表 1。

表 1 实验样本的采集与组成

样本集	采样时间	B 模块	P 模块	X 模块
训练集	2005.11 ~ 2006.03	132	80	106
验证集	2006.04 ~ 2006.07	50	54	53

注:实验样本的采集从 2005 年 11 月开始到 2006 年 7 月底结束,其中 B 模块在 2006 年 5 月后作过配方调整。在 B 模块验证集的 50 个样本中有 20 个样本是在配方调整后采集的,因此这 20 个样本的总体评吸结果与前 30 个样本存在明显差异。

2.2 方法

2.2.1 采集实验样本的光谱数据

设定仪器的主要工作参数,光谱扫描范围:10000 ~ 4000 cm^{-1} ;分辨率:8 cm^{-1} ;扫描次数:64。为减少仪器的开关频次且保证仪器的稳定运行,整个实验从开始到结束仪器保持免关机状态。在二级配方工艺过程正常的条件下,现场采集约 1 kg 同一批次和同一配方模块的烟丝,混合均匀后取 30 g 直接装入采样杯轻压至平,用于采集样本的光谱数据。

2.2.2 建立类模型

应用 Pirouette v3.11 软件中的 SIMCA 统计建模方法,结合 B、X 和 P 3 个模块烟丝的近红外光谱数据,构造相应的类模型。

2.2.3 验证类模型

通过 InStep v2.11 软件,调用建好的 B、X 和 P 3 个配方模块的 SIMCA 类模型,结合 MSPC 基本原理,应用该软件所具有的自动生成控制图的功能实测验证集样本,以验证类模型的可靠性。

3 结果与讨论

3.1 实验样本光谱数据的采集与处理

由于本实验直接采集烟丝的光谱,丝状样品的均匀性较差,原始光谱存在明显的噪声和基线漂移,见图 1(a)。因此在建模前,采用多元散射校正(MSC)处理^[20]光谱,以扣除样品粒径和散射造成的影响;同时应用 Savitzky-Golay 平滑滤波^[20-21]并结合二阶微分,以过滤噪声,消除基线漂移干扰。由于原始光谱的信号质量与样品光谱特性、仪器参数设置和仪器硬件水平等因素有关,因此平滑窗口宽度的选择具有一定的经验性,较大的窗口宽度会造成信号失真,较小的窗口滤波效果不佳。在本实验中,根据所使用的仪器和测量对象,平滑窗口宽度选择 11 点,二阶微分选择 21 点,获得了较好的建模效果。经处理后的光谱见图 1(b)。

3.2 最佳光谱区(变量)的选择与类模型的建立

采用全谱区(10000 ~ 4000 cm^{-1})建模,虽然保留了全部信息,但也引入了吸收弱、包含有较多高频噪声的谱区(如 10000 ~ 7700 cm^{-1})而使建模效果变差,见图 1(b)。为此,依据变量与模型化能力(TMP, Total modeling power)和类别识别能力^[22](DP, Discriminating power)的相关性,剔除一些不必要的冗余变量,选择最佳谱区,即:

$$TMP = 1 - \frac{s_{j(\text{total})}}{s_{0j(\text{total})}}$$

式中:TMP——变量的总模型化能力; $s_{j(\text{total})}$ ——每个类残

差矩阵中第 j 个变量的残余标准偏差的总和; $s_{0j(\text{total})}$ ——每个类第 j 个变量的标准偏差的总和。

$$DP_{BP}(j) = \left(\frac{s_{BP}^2(j) + s_{PB}^2(j)}{s_{(PP)}^2(j) + s_{BB}^2(j)} \right)^{1/2} - 1$$

式中: $DP_{BP}(j)$ ——变量的类别识别能力; $s_{BP}^2(j)$ 、 $s_{PB}^2(j)$ ——分别为 P 类和 B 类第 j 个变量的残余方差; $s_{(PP)}^2(j)$ ——B 类第 j 个变量拟合于 P 类产生的残余方差; $s_{(BB)}^2(j)$ ——P 类第 j 个变量拟合于 B 类产生的残余方差。

当 TMP 值(0 ~ 1 之间)越大,变量对建模的贡献率就较高,即在主成分模型中的作用越大。当 $DP_{BP}(j) \gg 1$,就意味着变量 j 具有越强的类别(如 B 类和 P 类)识别能力。

通过初步建模比较, TMP 值和 $DP_{BP}(j)$ 值与谱区(变量)对应的相关性见图 2(a)和(b)。结合两者选择谱区(变量)构建类模型,其效果可通过类与类之间的分离程度(如 B 类和 P 类),用距离^[22]来衡量,即:

$$D_{BP} = \left(\frac{s_{BP}^2 + s_{PB}^2}{s_{(PP)}^2 + s_{BB}^2} \right)^{1/2} - 1$$

式中: D_{BP} ——B 类与 P 类之间的距离; s_{BP}^2 、 s_{PB}^2 ——分别为 P 类样本和 B 类样本的光谱残余方差; $s_{(PP)}^2$ ——B 类样本拟合于 P 类产生的光谱残余方差; $s_{(BB)}^2$ ——P 类样本拟合于 B 类产生的光谱残余方差。

当 D_{BP} 越大,就意味着 B 类与 P 类之间的分离效果越好。采用不同的谱区建模,类与类之间的距离比较见表 2。由表 2 可以看出,建模效果理想的最佳谱区为 7500 ~ 6700 cm^{-1} 和 6120 ~ 4050 cm^{-1} 。但值得注意的是,选择识别能力较高的过窄的谱区(如 5590 ~ 4320 cm^{-1})建模,虽然可提高分类效果,但会造成类模型代表性的损失^[22]。

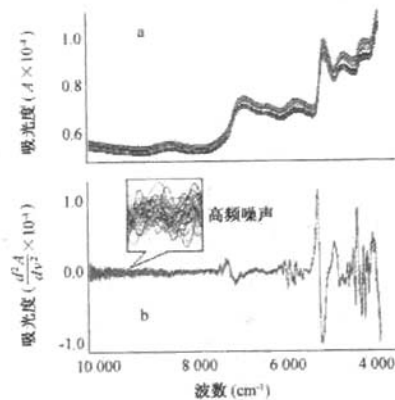


图 1 试样的 FT-NIR 光谱图(a)和相应的二阶导数光谱图(b)

在建模过程中,同时结合“剔一”(Leave-one-out)的交互校验方法^[17]剔除异常样本(P模块2个,B和X模块各3个),并确定最优主成分数。采用以上方法构建的B、P和X配方模块的类模型统计结果见表3;相应的SIMCA分类图(PC1、PC2和PC3等3个主成分的投影图)见图3。

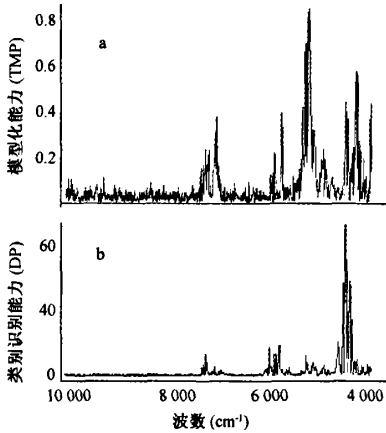


图2 谱区(变量)与模型化能力(TMP)和类别识别能力(DP)的相关图

表2 采用不同谱区建模类与类之间的距离比较

谱区 (cm ⁻¹)	类别	B@	P@	X@
10000 ~ 4000	B	0.0000	0.9987	0.7750
	P	0.9987	0.0000	0.7343
	X	0.7750	0.7343	0.0000
7500 ~ 6700 +	B	0.0000	1.883	1.5026
	P	1.883	0.0000	2.0143
6120 ~ 4050	X	1.5026	2.0143	0.0000
6120 ~ 4050	B	0.0000	1.8691	1.5140
	P	1.8691	0.0000	1.9533
	X	1.5140	1.9533	0.0000
5590 ~ 4320	B	0.0000	1.9132	1.6789
	P	1.9132	0.0000	2.0731
	X	1.6789	2.0731	0.0000

表3 B、P和X配方模块SIMCA统计建模的相关结果

项目	B模块	P模块	X模块
模型样本数量	129	78	103
数据预处理	多元散射校正(MSC)		
信号降噪处理	二阶微分(21点) + Savitzky-Golay平滑滤波(11点)		
最优谱区(变量)	7500 - 6700 cm ⁻¹ + 6120 - 4050 cm ⁻¹		
置信度(P)	0.95		
最优主成分数	12	5	8

3.3 类模型的验证——MSPC图的建立与应用

通过InStep v2.11软件将建好的SIMCA类模型打包,应用该软件自动生成控制图的功能,选择置信度和距离等统计量来建立MSPC控制图,并设定0.95作为置信度的控制线,对应的距离值由该软件自动生成。这样,应用类模型便可实现对过程样本的多元统计过程控制。在对验证集样本的实测中,为便于讨论,调入InStep v2.11软件检测的样本,按B模块50个样本、P模块54个样本和X模块53个样本的顺序进行。通过实测表明,在B模块的50个样本(序号1~50)中,配方调整前的30个批次的样本置信度在0.95以下,稳定性和均匀性较好,说明与B模块类模型中的样本无显著性差异,预测正确。后采集的20个样本因配方进行过调整[其中15个样本超出了预先设定的置信度控制限(>0.95),表现出显著性差异;另5个样本接近0.95],其相应的距离总体上要比配方调整前的30个样本远离模型,即这20个样本与B模块类模型中的样本相似性弱于配方调整前的30个样本,见图4(a)和(b)。这种因配方调整表现出来的质量差异总体上也获得了正确的预测,实测结果与呼吸结果基本一致。在P模块的54个样本(序号51~104)和X模块的53个样本(序号105~157)中,分别有2个和1个样本接近0.95,其余均未见异常,说明过程质量表现稳定,实测结果与实际吻合,分别见图4(c)、(d)、(e)和(f)。由此可以看出,监测配方过程中模块总体质量特性的变化趋势,采用“距离”和“置信度”这两张MSPC控制图基本可满足需求,且简单直观,容易把握应用。但值得注意的是,由于卷烟制造是一个多批次配方模块组配的间歇过程,时常会遇到配方模块调整和优化的情况,此时模块的整体特性已随之改变,因此必须新的正常过程状态下重建与之相应的类模型后才能进行MSPC监测。

采用类模型建立MSPC控制图,其目的不是为了

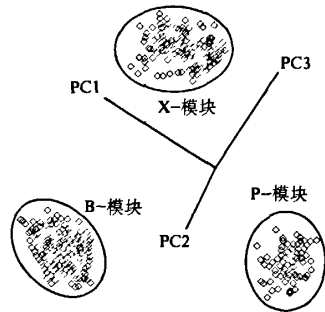


图3 B、P和X配方模块的SIMCA分类图

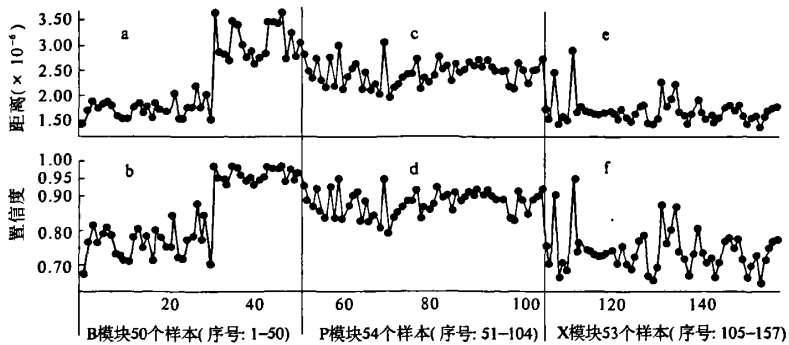


图4 B、P和X 3个配方模块的MSPC控制图

观测模块各组分的质量特性,而是为了对模块的整体宏观特性进行监测。同时,若要定量监测模块中的一个或多个特定质量参数的变化,就必须捆绑预先建立的校正模型^[15]。

参考文献

[1] Jerome J, Workman Jr. Review of process and non-invasive near-infrared and infrared spectroscopy: 1993-1999 [J]. *Applied Spectroscopy Reviews*, 1999, 34 (1&2): 1-89.

[2] 陆婉珍, 袁洪福, 徐广通, 等. 现代近红外光谱分析技术 [M]. 北京: 中国石化出版社, 2000.

[3] 刘国珍, 陈祖刚, 李丹, 等. 近红外光谱分析技术进展及其在烟草行业中的应用 [J]. *烟草科技*, 2001 (11): 5-17.

[4] 陶晓秋, 杜顺莲, 李维娜. 近红外光谱法测定烟草中硫含量的研究 [J]. *烟草科技*, 2004 (8): 33-35.

[5] 王家俊. FT-NIR 光谱分析技术测定烟草中总氮、总糖和烟碱 [J]. *光谱实验室*, 2003, 20 (2): 181-185.

[6] 王家俊, 罗丽萍, 李辉, 等. FT-NIR 光谱法同时测定烟草根、茎、叶中的氮、磷、氯和钾 [J]. *烟草科技*, 2004 (12): 24-27.

[7] 王家俊, 梁逸曾, 汪帆. SIMCA 分类法与偏最小二乘算法结合近红外光谱检测卷烟的内在品质 [J]. *计算机与应用化学*, 2006, 23 (11): 1133-1136.

[8] 王家俊, 汪帆, 马玲. SIMCA 分类法与 PLS 算法结合近红外光谱应用于卷烟纸的质量控制 [J]. *光谱学与光谱分析*, 2006, 26 (10): 1858-1862.

[9] 王家俊, 梁逸曾, 汪帆. 偏最小二乘法结合傅里叶变换近红外光谱同时测定卷烟焦油、烟碱和一氧化碳的释放量 [J]. *分析化学*, 2005, 33 (6): 793-797.

[10] 周汉平, 王信民, 宋纪真, 等. 烟叶结构和油份的近红外光

谱预测 [J]. *烟草科技*, 2006 (1): 10-14.

[11] 邱军, 王允白, 付秋娟, 等. 近红外光谱法快速测定烟草中的钙 [J]. *烟草科技*, 2006 (2): 31-32, 36.

[12] 蒋锦锋, 赵明月. 近红外光谱法快速测定烟草中的总挥发酸和总挥发碱 [J]. *烟草科技*, 2006 (3): 33-37.

[13] 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.

[14] 段焰青, 孔祥勇, 李青青, 等. 近红外光谱法预测烟草中的纤维素含量 [J]. *烟草科技*, 2006 (8): 16-20.

[15] 王家俊, 袁洪福, 陈剑明. 多变量分析方法结合近红外光谱表征卷烟配方的过程质量 [J]. *烟草科技*, 2006 (10): 5-9.

[16] Wold S. Pattern Recognition by Means of Disjoint Principal Components Models [J]. *Pattern Recognition*, 1976 (8): 127-139.

[17] 许禄. 化学计量学: 一些重要方法的原理及应用 [M]. 北京: 科学出版社, 2004.

[18] Wise B M, Gallagher N B. The process chemometrics approach to process monitoring and fault detection [J]. *Process Control*, 1996, 6 (6): 329-348.

[19] MacGregor J F, Kourti T. Statistical process control of multivariate processes [J]. *Control Engineering Practice*, 1995, 3 (3): 403-414.

[20] 严衍禄. 近红外光谱分析基础与应用 [M]. 北京: 中国轻工业出版社, 2005.

[21] Savitzky A, Golay M J E. Smoothing and differentiation of data by simplified least squares procedures [J]. *Analytical Chemistry*, 1964 (36): 1627-1639.

[22] Infometrix Inc. *Multivariate Data Analysis* [M]. Woodville: Infometrix Inc., 2003.