第26卷, 第10期
2006年10月

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

Vol. 26, No. 10, pp1858-1862
October, 2006

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

王家俊[1], 汪 帆[2], 马 玲[1]

1. 红河卷烟总厂产品中心, 云南 弥勒 652300
2. 曲靖师范学院化学系, 云南 曲靖 655000

摘 要 应用 SIMCA 分类法与 PLS 算法结合卷烟纸的傅里叶变换近红外光谱(FT-NIR)建立了卷烟纸的分类模型, 用于卷烟纸的判别分类, 效果良好; 同时, 建立了测定卷烟纸定量、厚度、透气度、水分和灰分等性质的校正模型, 其相应的相关系数分别为 0.976 8, 0.966 4, 0.947 0, 0.956 3 和 0.975 9; 全交互校验均方残差分别为 0.561 4, 0.096 0, 1.274 1, 0.096 7 和 0.260 3。校正模型应用于样品实测, 结果准确, 令人满意。

主题词 SIMCA 分类法; PLS 算法; 近红外光谱; 卷烟纸; 质量控制
中图分类号: O657.3 文献标识码: A 文章编号: 1000-0593(2006)10-1858-05

引 言

长度为 85 mm 的普通卷烟, 尽管卷烟纸占烟支总重量很小的比例(5%左右), 但它对静态燃烧率、抽吸口数、通风度、烟气递送量, 乃至卷烟的外观特征都有较大的影响[1], 检测与控制卷烟纸的整体质量, 予达到卷烟设计的预期效果有着重要的作用和意义。通常, 是按国标 GB/T 12655—1998 对表征卷烟纸质量的定量、水分、透气度和灰分等性质进行质检。而检测这些指标, 整个分析过程涉及较多的仪器, 操作繁琐, 耗工耗时速度慢, 难于满足大批量质量检测的需要。

物质结构理论认为, 物质的性质与结构和组成有着密切的内在关系, 若结构和组成的变异在光谱上得于表征, 那么, 这些性质的测定就可应用光谱分析技术来实现。傅里叶变换近红外光谱(FT-NIR)分析技术有机融合了光谱量测技术、化学计量学和计算机技术为一体, 对复杂多组分体系的分析, 通过多元校正和模式识别方法, 建立相应的校正模型和分类模型, 不但可预测未知样品的多个组分或性质, 同时还可对样品品质归属进行分类, 又因其具有样品前处理简单、分析速度快、精度高等优点, 被广泛应用于农业、食品、石化、医疗等领域[2, 3]。近些年来, 在国内烟草行业, 该项分析技术也被应用于原料质量控制、卷烟品质检测[4-7], 但将该项技术应用于卷烟纸的质量控制尚未见报道。

本文选择具有一定代表性的不同厂商生产的卷烟纸作为样本, 应用 SIMCA 分类法[8], 对 4 种卷烟纸的近红外光谱进行分类, 建立了相应的类模型, 并对未进入校正集的样品进行判别分析, 获得了正确的分类结果。此外, 在进行定量分析时, 分别以 GB/T451.2—1989, GB/T451.3—1989, GB/T458—1989, GB/T462—1989 和 GB/T463—1989 等标准方法测定的卷烟纸的定量、厚度、透气度、水分和灰分等性质数据为基础, 应用偏最小二乘算法(PLS)[9] 与相应的近红外光谱数据进行拟合, 分别建立了测定卷烟纸定量、厚度、透气度、水分和灰分等性质的校正模型, 模型采用独立校验集实测验证, 配对 t-检验表明, 在显著水平为 0.05 的条件下, 模型预测值与标准方法测定结果无显著性差异。本法应用于卷烟纸的类别归属判别分析及其相应品质指标的测定, 操作简捷、准确, 取得了令人满意的结果, 可为卷烟纸质量控制提供一种可行的快速分析手段。

1 实验部分

1.1 主要仪器及参数设置

Spectrum One NTS 近红外光谱仪, 包括带 InGaAs 检测器的漫反射积分球附件(美国 PE 公司); TQ Analyst 6.2(美国 Nicolet 公司)和 Pirouette 3.11 数据处理软件(美国 Infometrix 公司)。

仪器的主要工作参数设置为光谱扫描范围: 10 000~4 000 cm^{-1}; 分辨率: 8 cm^{-1}; 扫描次数: 64。

1.2 实验标样取制及基础性质数据测定

收稿日期: 2005-06-16, 修订日期: 2005-09-08
基金项目: 云南省高教厅教学科研学术带头人项目资助
作者简介: 王家俊, 1962 年生, 红河卷烟总厂产品中心工程师

参照 GB/T450－1989 标准，取制不同批次不同厂家的纸样 800 个，其中包括 200 个不同批次名优卷烟的纸样。将制好的纸样装入密封袋保存，即为实验标样。在恒温(23±1) ℃、恒湿(50±2)% RH 的条件下，分别以 GB/T451.2—1989，GB/T451.3—1989，GB/T458—1989，GB/T462—1989 和 GB/T463—1989 等标准方法测定卷烟纸的定量、厚度、透气度、水分和灰分等性质数据。

1.3 采集标样光谱数据

开机预热 2 h 后，在与测定基础数据相同的实验条件下，将纸样置于积分球上轻压至平，即可采集纸样的近红外漫反射光谱数据。

1.4 建立模型

应用 Pirouette 软件，设风险水平为 5%，对 250 个不同批次的 A，B，C 和 D 等 4 个质量标准纸样的光谱进行 SIMCA 分类，建立相应的类模型；应用 TQ Analyst 软件中的 PLS 算法把采集到的 800 个纸样的光谱数据与标准方法测定的相应基础性质数据，结合“剔一”(Leave-one-out) 交互效验方法确定最佳主因子，分别建立校正模型。最后，将建立的分类模型与定量校正模型结合，即为卷烟纸的定性定量质量控制模型。

2 结果与讨论

2.1 训练集样品的选择

在应用 SIMCA 建模时，本实验主要选取进厂的 A，B，C 和 D 等 4 个质量标准纸样共 250 个作为训练集，即可满足质量控制的需求。而在应用 PLS 算法建立校正模型时，除了选择和增加 A，B，C 和 D 等 4 个质量标准的纸样(共 350 个)进入训练集外，还考虑选入了 200 个不同批次名优卷烟的纸样进入训练集，这样，训练集既具有较好的代表性，同时又拓宽模型的预测范围，增强了模型的适应能力，使建立的校正模型除了与 SIMCA 分类模型结合应用于进厂卷烟纸的质量控制外，还可满足更为宽泛的定量分析需要。

2.2 最佳光谱区的选择与模型的建立

采用全谱区建模，虽然保留了全部信息，但也引入了不必要的噪声和无用的信息而使模型变差。依据光谱与预测性质表现出来的统计特征，可以方便地删除冗余信息，准确地确定有效信息率较高的最佳谱区。

应用 SIMCA 分类法建模，类与类的分离效果可用距离来衡量[10]，即

$$D_{pq} = \left(\frac{s_{pq}^2 + s_{qp}^2}{s_{pp}^2 + s_{qq}^2}\right)^{1/2} - 1$$

最优谱区的选取，依据各谱区光谱吸收强度见图 1(a)，谱区与 SIMCA 识别能力[10]的相关性见图 1(b)，选择 8 000~4 000 cm^{-1}谱区分类，效果理想，分类结果见图 2。

实验表明，若选用全谱分类，则引入了吸收弱、包含较多高频噪声的谱区(10 000~ 8 000cm^{-1})，导致分类效果变差。但值得注意的是，选择识别能力与相关性较高的过窄的谱区，虽可获得较好的分类效果，但会造成类模型的代表性损失。采用不同谱区分类，类与类之间的距离比较见表 1。

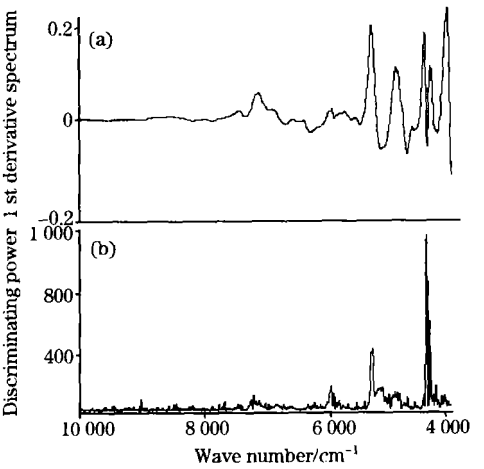

Fig. 1　NIR diffusion reflectance spectrum(a) and their regions *versus* discriminating power from SIMCA analysis(b)

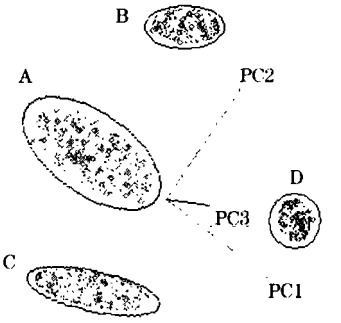

Fig. 2　Class projections plot from SIMCA analysis of A, B, C and D

Table 1　Interclass distances with different spectral regions

Spectral regions	Class	A@	B@	C@	D@
10 000-4 000 cm^{-1}	A	0.000 0	3.613 2	1.712 2	2.855 3
	B	3.613 1	0.000 0	5.307 0	6.724 4
	C	1.712 2	5.307 0	0.000 0	3.197 6
	D	2.855 3	6.724 4	3.197 7	0.000 0
9 000-4 000 cm^{-1}	A	0.000 0	5.099 6	2.132 5	3.050 2
	B	5.099 6	0.000 0	7.370 4	7.706 7
	C	2.132 5	7.370 4	0.000 0	3.817 9
	D	3.050 2	7.706 7	3.817 9	0.000 0
8 000-4 000 cm^{-1}	A	0.000 0	5.792 0	2.571 5	3.461 8
	B	5.792 0	0.000 0	8.739 2	8.804 2
	C	2.571 5	8.739 2	0.000 0	4.241 2
	D	3.461 8	8.804 2	4.241 2	0.000 0
7 500-4 000 cm^{-1}	A	0.000 0	6.181 6	2.577 4	3.669 2
	B	6.181 6	0.000 0	9.159 2	9.279 4
	C	2.577 4	9.159 2	0.000 0	4.562 4
	D	3.669 2	9.279 4	4.562 4	0.000 0

同样，在建立校正模型时(以灰分建模为例说明)，虽然全谱区(10 000~ 4 000 cm^{-1})与灰分这一性质表现出良好的相关性，见图 3(a)，但有效信息率较高的，也就是对建模贡献率大的是其中大方差的谱区(7 500~ 4 050 cm^{-1})，见图 3

(b)，它表征了光度的变化与纸样灰分含量变异的相关性，应用高相关性、大方差的谱区建模，效果理想，表 2 为采用不同谱区建立的灰分校正模型的统计结果，其中，建模效果最佳的谱区为 7 500~ 4 050 cm^{-1}。

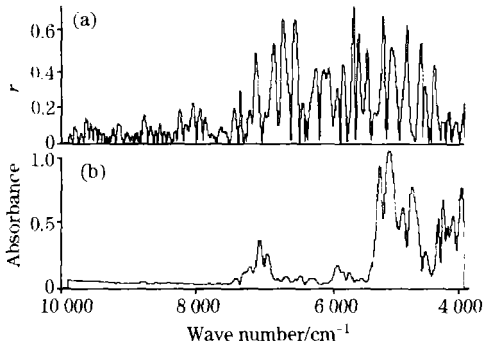

Fig 3 Spectral region *versus* correlation coefficient(r) for ash (a) and their variance spectrum(b)

Table 2 Statistics results for calibration models of ash with different spectral regions

Spectral regions/ cm^{-1}	Correlation coefficient	Factor	RMSECV
10 000-4000	0. 966 7	17	0 325 6
7 500-4 050	0. 975 9	16	0 260 3

应用 PLS 算法建模是一个协同过程，在对光谱进行预处理的同时，为了建立较为稳健的模型，避免“欠拟合或过拟合”未充分利用信息或引入过多的噪声，本实验采用了结合“剔一”的交互效验法来确定最佳主因子，即当交互效验均方残差(RMSECV) 达到最小时的主因子。图 4 为灰分的标准值(Actual) 与模型预测值(Calculated) 的散点图和相应的主因子(Factor) 与全交互效验均方残差的变化图，其中最佳主因子为 16。其余的定量、厚度、水分和透气度等性质建模的最佳谱区选择、最佳主因子确定亦然，采用最佳谱区、最佳主因子建立的校正模型见表 3。

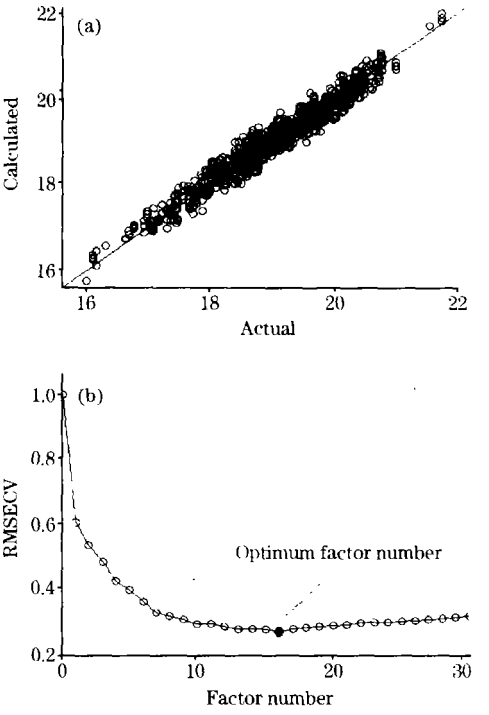

Fig 4 Actual value *versus* calculated value of ash (a) and changes of RMSECV with factor number(b)

Table 3 Statistics results gor calibration models of thickness, grammage, permeability, moisture and ash

Model name	Number of sample in calibration set	Correlation coefficient	Optimum factor	RMSECV	Predicted range
Grammage	700	0 976 8	7	0 561 4	24 00~ 34 00/(g • m^{-2})
Thickness①	500	0 966 4	11	0 096 0	4 20~ 5 40/ × 10^{-2} mm
Permeability	400	0 947 0	15	1. 274 1	51. 00~ 66 00/ CU
Moisture	450	0 956 3	5	0 096 7	(3 90~ 5. 10) / %
Ash	550	0 975 9	16	0 260 3	(16 60~ 21. 00) / %

Mark: ① × 10^{-2} (mm)

3 模型的可靠性验证

3. 1 SIMCA 分类模型的验证

采集未进入类模型的纸样 46 个作为测试集，其中 A，B，C 和 D 等 4 个质量标准的纸样各 10 个，红河甲级、精品假烟纸样(Bogus) 各 3 个。应用以上建好的 SIMCA 分类模型进行判别，均获得了正确的分类结果，特别是对假冒制品，具有良好的判别能力。分类效果见图 5。

3. 2 校正模型的准确性验证

通常，从校正模型的统计参数(相关系数，交互校验均方残差和残差分布等) 基本可判断模型的预测能力。但在建模过程中，常因删除一些样本，从而影响到模型校正集的代

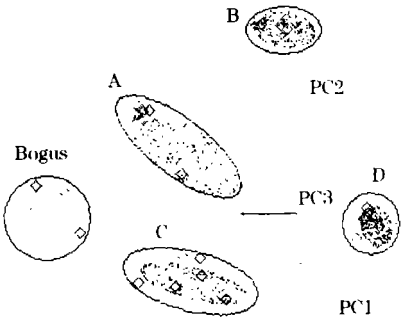

Fig 5 Class projections plot from SIMCA analysis of A, B, C, D and bogus

表性及可靠性。本实验另行随机采集 20 个纸样作为独立校验集，进一步实测验证模型，即用以上建好的模型和标准方法在相同的实验条件下进行测定，测定结果见表 4。通过配对 t 检验验证，在显著性水平 0.05 时，t 分布值均小于临界值($t_{(0.05,\ 20)}$ = 2.086)，说明两种测定方法不存在显著性差异，即表明两种方法的测定结果吻合。

Table 4　Comparison results of PLS-FT-NIR method and standard methods for grammage, thickness, permeability, moisture and ash (n= 20)

No	Grammage/(g·m^{-2})			Thickness/×10^{-2} mm			Permeability/CU			Moisture/%			Ash/%		
	GB/T 451.2	PLS-FT-NIR	Deviation	GB/T 451.3	PLS-FT-NIR	Deviation	GB/T 458	PLS-FT-NIR	Deviation	GB/T 462	PLS-FT-NIR	Deviation	GB/T 463	PLS-FT-NIR	Deviation
1	29.23	29.37	0.14	5.02	5.12	0.10	63.78	61.08	−2.70	3.81	3.63	−0.18	18.24	18.17	−0.07
2	30.26	30.22	−0.04	5.02	5.17	0.15	65.58	61.89	−3.69	3.75	3.62	−0.13	19.23	19.54	0.32
3	30.23	30.21	−0.02	5.43	5.38	−0.05	60.75	62.03	1.28	3.84	3.73	−0.11	18.74	18.94	0.20
4	29.61	29.83	0.22	5.45	5.39	−0.06	60.75	60.15	−0.60	3.76	3.62	−0.14	18.73	18.74	0.01
5	29.76	29.91	0.15	5.05	5.12	0.07	57.48	59.39	1.91	3.83	3.73	−0.10	18.76	18.72	−0.04
6	29.67	29.73	0.06	5.40	5.15	−0.25	58.13	59.94	1.81	3.66	3.68	0.02	18.85	19.11	0.26
7	29.59	29.95	0.36	5.46	5.26	−0.20	61.33	61.70	0.37	3.83	3.77	−0.06	19.27	19.05	−0.22
8	32.19	32.43	0.24	5.54	5.47	−0.07	61.33	61.35	0.02	3.72	3.65	−0.06	19.11	19.10	−0.01
9	31.79	31.92	0.13	5.70	5.39	−0.31	60.05	60.01	−0.04	3.72	3.64	−0.07	20.88	20.93	0.06
10	32.45	32.81	0.36	5.07	5.17	0.10	61.14	60.14	−1.00	3.75	3.66	−0.09	20.59	20.85	0.26
11	32.45	32.64	0.19	5.04	5.07	0.03	59.56	59.17	−0.39	3.29	3.58	0.29	21.29	21.12	−0.17
12	31.68	31.92	0.24	4.99	5.03	0.04	61.27	61.12	−0.15	3.59	3.74	0.15	21.49	21.35	−0.14
13	29.93	29.64	−0.29	4.99	5.05	0.06	61.28	60.71	−0.57	3.58	3.70	0.12	21.08	21.28	0.21
14	29.84	29.74	−0.10	5.04	5.13	0.09	62.92	63.47	0.55	3.50	3.58	0.08	19.13	19.29	0.16
15	29.98	29.79	−0.19	5.03	5.03	0.00	65.28	65.31	0.03	3.65	3.69	0.04	19.10	19.10	0.00
16	30.52	30.31	−0.21	5.45	5.33	−0.12	63.41	63.54	0.13	3.72	3.65	−0.07	18.43	18.41	−0.01
17	29.94	30.02	0.08	5.48	5.34	−0.14	61.74	60.43	−1.31	3.45	3.61	0.16	18.28	18.39	0.12
18	29.58	29.74	0.16	5.17	5.19	0.02	63.50	61.17	−2.33	3.49	3.57	0.08	18.94	18.96	0.02
19	29.75	29.91	0.16	4.94	5.07	0.13	63.64	62.90	−0.74	3.45	3.64	0.19	18.46	18.30	−0.16
20	31.25	31.25	0.00	5.01	5.17	0.16	63.15	61.43	−1.72	3.87	3.20	−0.67	19.43	19.36	−0.07
t-value		2.029			0.415			1.440			0.642			0.974	

3.3　校正模型的精密度试验

在与上述相同的实验条件下，对同一样品进行 10 次测定，结果见表 5。比照相关标准要求，单层厚度测定的重复性在 ISO 534: 1988(与 GB/T 451.3—1989 等效)中，测定间的误差不应大于 1.3 μm；透气度测定的重复性在 ISO 5636/2: 1984(与 GB/T 458—1989 等效)中，变异系数不应大于 6.4%；灰分测定的重复性在 GB/T 463—1989 中，以 2 次测定的算术平均值作为结果，测定间的误差不应大于平均值的 5%。与之比较，该法对厚度、透气度和灰分具有良好的量测精度。

GB/T 451.2—1989，GB/T 462—1989 及相关标准对卷烟纸的定量和水分测定未明确精确度的数值。本文的研究成果对人的健康有一定的关系，类似的工作亦可参阅文献[11]。

4　结　论

实验结果表明，在严格的实验条件下，本方法应用于卷烟纸的定性分析，特别是质量保证，真伪甄别，具有良好的判别能力；用于定量分析，与相应的标准方法相比，两者测定结果比较吻合。该法具有操作简便、速度快和精度高等优点。将所建立的分类模型与校正模型相结合，应用于卷烟纸的定性定量质量控制是可行的。

Table 5　The precision test(n= 10)

No.	Grammage /(g·m^{-2})	Thickness /(×10^{-2} mm)	Permeability/CU	Moisture /%	Ash/%
1	32.11	5.08	61.78	4.17	21.39
2	32.13	5.10	61.25	4.20	21.34
3	32.15	5.08	61.26	4.18	21.25
4	32.05	5.09	61.78	4.18	21.55
5	32.32	5.10	61.76	4.19	21.50
6	32.31	5.11	61.82	4.19	21.43
7	32.32	5.16	61.65	4.19	21.55
8	32.42	5.11	62.24	4.18	21.34
9	31.93	5.11	60.86	4.17	21.39
10	32.15	5.15	60.82	4.16	21.36
Mean	32.19	5.11	61.52	4.18	21.41
S.D.	0.1496	0.0269	0.4575	0.0120	0.0980
RSD/%	0.46	0.53	0.74	0.29	0.46

参 考 文 献

[1] HU Qun, MA Jing, XIAO Yan, et al(胡 群, 马 静, 肖 燕, 等编译). Study of Auxiliary Material for Cigarette(卷烟辅料研究). Kunming: Yunnan Science and Technology Press(昆明: 云南科技出版社), 2001. 25.

[2] Jerome J, Workman Jr. Applied Spectroscopy Reviews, 1999, 34(1&2): 1.

[3] XU Guang-tong, YUAN Hong-fu, LU Wan-zhen(徐广通, 袁洪福, 陆婉珍). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2000, 20(2): 134.

[4] WANG Jia-jun(王家俊). Chinese Journal of Spectroscopy Laboratory(光谱实验室), 2003, 20(2): 181.

[5] WANG Jia-jun, LUO Li-ping, LI Hui, et al(王家俊, 罗丽萍, 李 辉, 等). Tobacco Science & Technology(烟草科技), 2004, (12): 24.

[6] WANG Jia-jun, LIANG Yi-zeng, WANG Fan(王家俊, 梁逸曾, 汪 帆). Chinese J. Anal. Chem.(分析化学), 2005, 33(6): 793.

[7] WANG Jia-jun, LIANG Yi-zeng, WANG Fan(王家俊, 梁逸曾, 汪 帆). Computers and Applied Chemistry(计算机与应用化学), In press.

[8] Wold S. Pattern Recognition, 1976, 9: 127.

[9] LIANG Yi-zeng(梁逸曾). White, Grey and Black Multicomponent System and Their Chemometric Algorithms(白灰黑复杂多组分分析体系及其化学计量学算法). Changsha: Hunan Publishing House of Science and Technology(长沙: 湖南科技出版社), 1996. 32.

[10] Edited by Infometrix Inc. Multivariate Data Analysis, Woodinville: Infometrix Inc., 2003. 6, 14.

[11] SHI Wei-wei, GAN Wu-er, SU Qing-de(石玮玮, 淦五二, 苏庆德). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(7): 1135.

The Quality Assessment of Cigarette Paper by SIMCA and PLS Combined with Near Infrared Spectrum

WANG Jia-jun[1], WANG Fan[2], MA Ling[1]

1. Production Research Center, Honghe Cigarette General Factory, Mile 652300, China
2. Department of Chemistry, Qujing Teacher's College, Qujing 655000, China

Abstract By using algorithm of SIMCA and partial least squares(PLS) combined with Fourier transform near infrared spectra (FT-NIR), the classification methods were established for the discrimination of cigarette paper. Meanwhile, the calibration models were established for the determination of the grammage, thickness, permeability, moisture and ash of cigarette paper. Correlation coefficients of the models were 0.976 8, 0.966 4, 0.947 0, 0.956 3 and 0.975 9, and the root mean square errors of cross validation (RMSECV) were 0.561 4, 0.096 0, 1.274 1, 0.096 7 and 0.260 3 respectively. The methods has been applied to the determination and discrimination of unknown samples with satisfactory results.

Keywords SIMCA; PLS; FT-NIR; Cigarette paper; Quality assessment

(Received Jun. 16, 2005; accepted Sep. 8, 2005)