文章编号:1001-5078(2008)07-0662-04 ·红外技术·

用光谱主成分分析选择烟草近红外模型校正集样品

段焰青[1],者 为[1],李青青[2],杨 涛[3]
(1.红云烟草(集团)有限责任公司技术中心,云南 昆明 650202;2.云南师范大学生命科学学院,
云南 昆明 650092;3.云南瑞升烟草技术(集团)有限公司,云南 昆明 650106)

摘 要:为研究光谱主成分分析方法在选择烟草近红外模型校正集样品过程中的可行性,本文采用常规的人为挑选和光谱主成分分析两种方法选择校正集样品,并利用偏最小二乘法建立烟叶总糖、总氮和烟碱含量的预测模型,同时比较了这两种校正集样品选择方法所建立模型的各项评价指标和预测结果的差异性。结果表明,上述两种方法建立的总糖、总氮和烟碱含量预测模型的各项评价指标非常接近,对20个烟叶样品的预测结果也无明显差异。因此,应用光谱主成分分析是烟草近红外模型构建过程中的一种非常有效,且节约常规化学分析成本的校正集样品选择方法。

关键词:近红外光谱;主成分分析;烟草;校正集
中图分类号:TN219 **文献标识码**:B

Application of Spectrum PCA in Selection of Calibration Sample Set for Tobacco Near-infrared Prediction Model

DUAN Yan-qing[1],ZHE Wei[1],LI Qing-qing[2],YANG Tao[3]
(1. Technology Center,Hongyun Tobacco (Group) Co.,Ltd.,Kunming
650202,China;2. Life Science College,Yunnan Normal University,Kunming 650092,China;
3. Yunnan Reascend Tobacco Technology (Group) Co.,Ltd.,Kunming 650106, China)

Abstract:In order to study the feasibility of spectrum PCA in the selection procedure of calibration sample set for tobacco near-infrared predicting model,calibration samples sets were selection by both methods of routine artificial choosing and spectrum PCA. The predicting modes of total sugar,total nitrogen and nicotine content were established by partial least square statistic method,and the differences of the evaluating indexes and predicting results of the modes, based upon the two methods were compared in this paper at the same time. Results showed that the evaluating indexes were very similar,and there were no significant differences in the predicting results of 20 tobacco leave samples,between the predicting modes of total sugar,total nitrogen and nicotine content established by the two methods. Therefore,application of spectrum PCA in the selection of calibration sample set was very efficient and cost effective in routine chemical analysis in the procedure of tobacco NIR mode establishment.

Key words:near infrared spectroscopy;PCA;tobacco;calibration set

1 引 言

目前,快速、稳定、无损的近红外光谱分析技术在烟草行业得到极大重视、发展和应用,并在烟草生产、加工和科研的相关分析中发挥出越来越大的优势[1-8]。由于近红外光谱分析技术是一种需对仪器

基金项目:云南省科技攻关及高新技术发展计划项目(No.2006GG22);云南省自然科学基金资助项目(No.2006C0027Q);云南中烟工业公司资助项目(No.2005JC04)。
作者简介:段焰青(1973-),男,白族,工程师,博士,主要从事烟草化学与品质研究。
收稿日期:2008-01-02

进行二次开发才可以应用的技术,其应用的前提条件是必须建立相关的数学模型。而数学模型的构建则需积累一定量同时具有光谱数据和化学(或其他定性特征)数据的校正集样品。同时,由于不同校正集样品的近红外预测模型对相同待测样品的预测结果可能会有较大差异,因此,校正集样品的选择是近红外光谱数据处理及分析过程中的关键环节。在校正集样品的选择中,要求所选样品在待测指标方面具有很好的代表性,样品的光谱特征及其性质范围应能涵盖以后待测的样品[9-10]。为保证校正模型的稳定性,烟草常规化学成分近红外预测模型校正集的样品数不应低于100个[11];样品的指标含量应在所测范围内;各样品在所测范围内必须分布均匀,避免共线性现象的发生[9-10]。

通过人工来选择确定近红外模型校正集样品是目前烟草模型构建过程中常使用的方法,即根据一定样品的积累和性质或组成等情况分布来选择校正集样品[12]。如要构建一个云南烤烟烟叶的近红外预测模型,就应该均匀地选择来自全省各烟叶产地的,涵盖各烟叶品种和等级的样品,同时,还要考虑样品的化学值分布范围,要求校正集样品在各化学值梯度都有一定的分布。但这种方法需对大量样品的化学数据进行准确测定后才可构建和优化模型,整个过程费时、耗财费力。因此,笔者对烟草近红外模型校正集的选择方法进行了大量的摸索和实验研究,认为通过采集样品的近红外光谱,并对其主成分空间分布情况进行研究后来选择校正集样品是一种非常有效且适用的方法。

2 材料和方法

2.1 仪器设备

Nicolet Antaris 型傅里叶变换近红外光谱仪,带有 ϕ5cm 石英样品杯的旋转样品采集台和漫反射积分球,配备 RESULT™样品光谱采集的集成软件和 TQ analyst 7.1 化学计算量学软件(美国 Thermo Nicolet 公司);FOSS CYCLOTEC 1093 型旋风磨(美国 FOSS 公司,带 40 目网筛);101A-6 型普通可调温烘箱(上海崇明实验仪器厂);BP-211D 型天平(感量 0.01g,德国赛多利斯公司);各种常规化学器皿。

2.2 样品与处理

本研究中的1200个烤烟烟叶样品均来自云南省各地烟区,其中2004年645个,2005年555个(如表1所示),包括K326、云87和红花大金元三个品种。样品在40~60℃下烘至含水率为8%左右,过40目筛,装入样品瓶,盖紧瓶盖,常温下避光保存。

表1 烟叶样品的基本情况

Tab.1 general situation of the tobacco leaves

年份	B2F	B2L	C2F	C2L	C3F	C3L	X2F	X2L	X3F
2004	120	18	154	32	175	21	89	22	14
2005	136	25	127	21	124	14	67	23	18
总计	256	43	281	53	299	35	156	45	32

2.3 光谱数据采集

取约6g烟叶粉末,置于5cm石英样品杯中,放到光谱仪采集光谱。采集时旋转样品杯,并使用积分球漫反射检测器,以积分球镀金内壁作背景,每采集一个样品扫一次背景。样品采集前,应用 RESULT™(Integration)集成软件编定样品光谱采集的工作流程,并使光谱仪开机预热1.5h。光谱采集条件为:光谱采集范围 10000~4000cm^{-1};8cm^{-1}分辨率;扫描72次。

2.4 化学值测定

按照行业标准《YC/T159-2002 烟草及烟草制品 水溶性糖的测定 连续流动法》、《YC/T161-2002 烟草及烟草制品 总氮的测定 连续流动法》和《YC/T160-2002 烟草及烟草制品 总植物碱的测定 连续流动法》规定的方法测定烟叶粉末样品的总糖、总氮和烟碱含量,以此作为样品的对应化学测定值。

2.5 光谱主成分分析

应用 TQ Analyst 7.1 化学计算量学软件的定量分析中的判别分析方法(Discriminant analysis),对1200个样品光谱数据进行主成分分析(取二阶导数),筛选并得到“选择样品”。

2.6 建 模

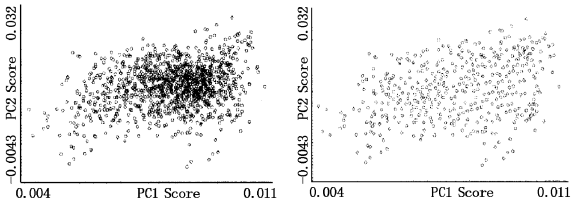
应用 TQ Analyst 7.1 化学计算量学软件分别处理1200个原始样品和选择样品光谱数据,再用软件中的偏最小二乘法(PLS)把各光谱数据与其对应的总糖、总氮和烟碱含量化学数据进行统计拟合,分别以1200个原始样品和选择样品为校正集,建立两套相应化学成分的预测模型,并对模型进行优化和检验,确保获得了数理指标最理想的数学模型。同时

对两套相应化学成分模型的各项评价指标，即相关系数(R^2)、内部交叉检验均方差(root mean square error of cross validation，RMSECV)、预测均方差(root mean square error of prediction，RMSEP)和运行指数(performance index，PI)进行考察。

3 结果与讨论

3.1 光谱主成分分析

通过对1200个样品光谱数据的主成分分析，根据其第一主成分和第二主成分(第一主成分方差贡献率为84.15%，第二主成分方差贡献率为12.86%，两种成分累计方差贡献率达97.01%)得分的空间分布情况去除相同主成分空间上的多个样品，最后筛选出350个均匀分布的样品作为“选择样品”。如图1所示，图1(a)为1200个原始样品的主成分分布图，图1(b)为剔除大部分样品后所选留的350个有代表性样品的主成分分布图。



(a) (b)

图1 云南烤烟样品第一和第二主成分分布图

(a)1200个样品；(b)去除重复样品后350个样品

Fig. 1 the first and second PCA plot of Yunnan flue-cured tobacco

(a) the 1200 samples;

(b) the 350 samples with removing of the repetitions

3.2 模型的建立

分别建立了1200个原始样品和350个选择样品为校正集的两套总糖、总氮和烟碱含量的预测模型，经对谱区范围、预处理方法和主因子数等建模参数的选择和优化后，各模型预测值与化学测定值之间都具较好的相关性，图2为1200个原始样品校正集总糖预测模型交叉验证预测值与化学测定值的相关性及误差分布情况，图3为350个选择样品校正集总糖预测模型的情况。总氮和烟碱模型的情况也与此类似。表2表明，各化学成分的两套模型的相关系数、内部交叉检验均方差、预测均方差和运行指数均没有明显差异，说明无论是用原始的1200个样品还是选择的350个样品作为模型校正集，所建模型在各项评价指标上都非常接近。

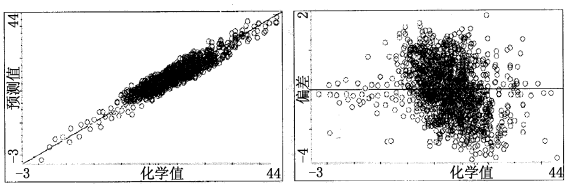


图2 1200个原始样品校正集总糖模型交叉验证预测值与化学测定值的相关性(左)及误差分布(右)

Fig. 2 pertinence (lift) and distribution of difference (right) between predictive value and chemical analyze values of cross validation for the sugar predictive model of calibration set with 1200 originalsamples

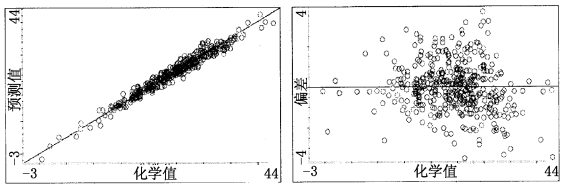


图3 350个选择样品校正集总糖模型交叉验证预测值与化学测定值的相关性(左)及误差分布(右)

Fig. 3 pertinence (lift) and distribution of difference (right) between predictive value and chemical analyze values of cross validation for the sugar predictive model of calibration set with 350 selective samples

表2 两套各化学成分预测模型的评价指标

Tab. 2 evaluative indexes of the chemical composition for the two set of predictive models

化学成分		R^2	RMSEP	RMSECV	PI
总糖	1200样品	0.9950	0.935	0.842	94.9
	350样品	0.9945	0.955	0.838	93.4
总氮	1200样品	0.9958	0.088	0.077	95.5
	350样品	0.9949	0.081	0.074	95.6
烟碱	1200样品	0.9939	0.125	0.118	94.2
	350样品	0.09941	0.127	0.114	94.6

3.2 模型的预测效果比较

为检验以上建立的两套烟叶总糖、总氮和烟碱含量近红外模型预测的准确性，分别对20个云南烟叶样品的总糖、总氮和烟碱含量作了实际预测分析，作为外部验证，结果如表3所示。

两套模型对20个烟叶样品的总糖、总氮和烟碱含量的预测结果基本一致。与化学测定值对比，1200个原始样品校正集模型和350个选择样品校正集模型对总糖实际预测的平均误差分别为0.60%和0.62%，对总氮实际预测的平均误差分别

为0.11%和0.13%,对烟碱实际预测的平均误差分别为0.14%和0.13%。以上结果表明两套模型实际预测的数据之间无明显差异。

表3 两套模型对20个烟叶样品总糖总氮和烟碱含量预测结果的比较

Tab.3 comparison of the predictive results of total sugar total nitrogen and nicotine from the two set of models

%

样品编号	总糖			总氮			烟碱		
	化学测定值	1200样品	350样品	化学测定值	1200样品	350样品	化学测定值	1200样品	350样品
1#	24.54	25.08	25.13	2.31	2.22	2.4	3.81	3.7	3.65
2#	31.51	31.74	32.06	1.69	1.67	1.88	2.77	2.9	2.88
3#	34.25	33.16	33.48	1.51	1.55	1.72	2.4	2.42	2.45
4#	33.96	32.69	32.93	1.45	1.5	1.71	2.31	2.24	2.35
5#	24.81	24.83	25.19	2.25	2.15	2.37	3.08	3	3.04
6#	29.11	27.19	27.74	1.96	1.81	2.05	2.45	2.38	2.43
7#	29.32	28.03	28.51	1.85	1.75	1.98	2.41	2.23	2.31
8#	29.11	28.5	28.9	1.64	1.74	2	2.2	2.13	2.21
9#	28.99	28.96	28.98	1.58	1.65	1.63	2.55	2.22	2.48
10#	29.88	30.36	30.54	1.71	1.62	1.74	2.08	1.99	2.15
11#	28.81	29.13	29.38	1.44	1.65	1.83	1.89	1.98	2.17
12#	29.22	28.81	29.01	1.67	1.59	1.81	1.37	1.52	1.77
13#	26.53	26.88	27.61	1.95	1.86	2.09	2.45	2.45	2.54
14#	30.44	30.98	30.91	1.76	1.8	1.92	2.25	2.38	2.46
15#	27.91	28.55	29.21	2.12	1.81	2.05	2.36	2.16	2.23
16#	30.18	29.07	29.86	1.74	1.87	2.17	1.85	1.62	1.75
17#	30.84	29.93	30.06	1.51	1.57	1.69	1.95	1.86	2
18#	30.7	30.71	30.73	1.55	1.49	1.64	1.47	1.62	1.76
19#	19.01	18.64	18.81	2.51	2.42	2.55	3.91	3.61	3.71
20#	21.88	22.51	22.74	2.44	2.04	2.19	2.95	2.64	2.87
平均误差		0.60	0.62		0.11	0.13		0.14	0.13

4 结 论

该研究利用云南烟叶样品,结合PLS法,分别利用1200个原始样品和350个通过光谱主成分分析方法筛选的样品作为校正集,构建得到两套总糖、总氮和烟碱的预测模型。通过对模型各项评价指标和实际预测效果的比较,表明两套模型在各项评价指标上非常接近,它们的实际预测数据之间也无明显差异。说明利用光谱主成分分析方法选择烟草近红外模型校正集样品是一种非常有效可行的方法,利用该方法可大幅度降低烟草近红外模型构建过程中人力、物力和时间的投入,切实节约近红外模型的构建成本。

参考文献:

[1] 段焰青,周红,王明锋,等.粒度对烟末总糖、总氮和烟碱含量NIR预测值的影响[J].烟草科技,2005,(7):22-23,40.

[2] 段焰青,周红,李青青,等.烟样水分质量分数对其常规化学成分近红外测定的影响[J].云南大学学报(自然科学版),2005,27(5):424-428.

[3] 王家俊,汪凡,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.

[4] 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.

[5] 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.

[6] 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.

[7] 段焰青,杨涛,孔祥勇,等.样品粒度和光谱分辨率对烟草NIR预测模型的影响[J].云南大学学报(自然科学版),2006,28(4):340-344.

[8] 束茹欣,张建平.应用近红外技术进行卷烟真伪鉴别的研究[J].上海烟业,2002(1):13-14.

[9] 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,2005:281.

[10] 严衍禄,景茂,张录达,等.傅里叶变换近红外漫反射光谱分析测量误差的研究[J].北京农业大学学报,1990,16:37-48.

[11] 李军会,秦西云,张文娟,等.局部偏最小二乘回归建模参数对近红外检测结果的影响研究[J].光谱学与光谱分析,2007,27(2):262-264.

[12] 于海燕,应义斌,刘燕德.农产品品质近红外光谱分析结果影响因素研究综述[J].农业工程学报,2005,21(11):160-163.