第28卷,第5期 光 谱 学 与 光 谱 分 析 Vol. 28,No. 5,pp1057-1061
2008年5月 Spectroscopy and Spectral Analysis May, 2008

潜变量聚类分析法在近红外光谱波长范围选择中的应用研究

鲍峰伟[1],彭黔荣[2*],刘景艳[3],蔡元青[2],毛寒冰[2],唐 珂[2],吕燕文[2]

1. 贵州大学化工学院,贵州 贵阳 550003
2. 贵州中烟工业公司技术中心,贵州 贵阳 550003
3. 河北科技大学生物科学与工程学院,河北 石家庄 050018

摘 要 介绍了潜变量聚类分析方法的基本原理,并将该方法应用于近红外光谱定量模型的谱区选择。以烟草样品为例,对107个样品的光谱进行处理,将光谱分为5簇,从化学角度分别解释了这5簇各自反映的信息。在此基础上,选择相应的波长范围用PLS方法建立了总糖、还原糖和尼古丁的定量分析模型。与全谱模型相比,3个模型的交互验证相关系数(Rtraining)分别由0.977 1,0.917 2,0.987 4提高到0.995 5,0.975 1,0.994 4;验证样品相关系数(Rtest)由0.977 8,0.941 2,0.993 2提高到0.992 7,0.967 9,0.994 0;交互验证均方差(RMSECV)由1.09,1.43,0.14降为1.05,1.05,0.13;预测残差均方差(RMSEP)由0.92,1.17,0.16降为0.39,0.63,0.11;预测样品间平均标准误差(*D*)由1.274%,1.972%,0.829%降为0.711%,0.843%,0.768%,表明用该方法建立模型的预测准确度和精密度均有所提高,对实际应用有一定的指导作用。

关键词 近红外光谱;潜变量聚类分析;波长选择
中图分类号:O657.3 **文献标识码**:A **文章编号**:1000-0593(2008)05-1057-05

引 言

近红外谱区的吸收强度比较弱,一般有机化合物在该处没有明显的特征吸收峰或吸收带,因此不能用特征波长的方法进行定量分析,必须借助化学计量学方法建立近红外光谱与样品化学数据之间的多元校正模型[1,2]。在建模过程中,若采用全谱数据,不仅计算量大,而且在某些光谱区域样品的光谱信息很弱,或与样品的组成和性质间缺乏相关关系,会造成模型的预测精度和稳定性降低,因此选择合适的谱区范围对建立一个好的模型是相当重要的[3-5]。常见的波长选择算法有相关系数法、逐步回归法、循环预测权重法(IPW)以及遗传算法等[6-8],这些算法对消除不相关因素的影响、提高模型预测能力起到很了大的作用,只是在选择波长的过程中会与校正样品集的化学值相关联,一定程度上受到化学值准确度的影响。本文在前人工作的基础上,引入潜变量聚类分析法用于波长选择,旨在消除化学值准确度对选择过程的影响。

1 潜变量聚类分析方法的基本原理

潜变量聚类分析是由Vigneau[9]提出的一种新型聚类算法,其基本原理描述如下。

记 $X_{(n\times p)}$ 为一组经过标准化处理后得到的光谱数据矩阵,n为样品数目,p为光谱维数。x_1,x_2,x_3,…,x_p为光谱矩阵的 p 个列向量。潜变量聚类分析方法的目的是将光谱向量 x_1,x_2,x_3,…,x_p 聚成 K 簇,分别记为 G_1,G_2,G_3,…,G_k。聚类的标准是每个簇内的向量更多的反映相同的信息。

计算方法如下。

1.1 采用层次聚类算法确定聚类数和初始聚类状态[10]

假定光谱矩阵共有 p 维,首先将这 p 维各自视为一簇,层次聚类逐级将 p 簇聚为一簇,标准是使每级聚类均有最大的 T 值,$T=\sum_{i}^{k}\lambda_{\max}^{g_i}$,其中 k 为聚类数,$\lambda_{\max}^{g_i}$ 为 g_i 的最大特征值。采用自下而上的算法,计算步聚如下。

(1)将 x_1,x_2,x_3,…,x_p 分别视为一簇,记为 g_1,g_2,g_3,…,g_k。

(2)计算 $\Delta T=\lambda_{\max}^{g_i}+\lambda_{\max}^{g_k}-\lambda_{\max}^{(g_i\cup g_k)}$,其中,$\lambda_{\max}^{g_i}+\lambda_{\max}^{g_k}-$

收稿日期:2006-05-10,**修订日期**:2006-08-20
基金项目:国家自然科学基金项目(20472057)资助
作者简介:鲍峰伟,1981年生,贵州大学化工学院在读硕士研究生 *通讯联系人 e-mail:pengqr@public.gz.cn

$\lambda_{max}^{g_j \cup g_k}$，分别是指 g_j，g_k，和 $g_j \cup g_k$ 的最大特征值，且可以证明 $\Delta T > 0$。

(3) 计算任意两簇 ΔT 记为 $\Delta T_{j,k}$，将 ΔT 最小的两簇合为一簇，并计算合并之后的 T 值。

(4) 重复(2)和(3)步骤，直到将所有变量聚为一簇。

(5) 选择聚类数 K，并记录当聚类数为 K 时的聚类结果，作为下一步计算的初使状态，各簇分别记为 G_1，G_2，G_3，…，G_k。聚类数的选择要根据聚类的目的，并且如果某次聚类产生相对较大的 ΔT，则保留此次聚类之前的聚类数。

1.2 对矩阵进行潜变量聚类处理

以 TC 值作为聚类分析的选择标准，TC 反映的是各向量与潜变量相关性之和，因此 TC 值越大说明各簇内部相关的信息越集中。

$$TC = n \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \mathrm{cov}^2(x_j, c_k)$$

当 x_j 属于第 k 簇时 $\delta = 1$，反之 $\delta = 0$，其中 c_k 为潜变量，这里取 g_k 的最大特征值对应的特征向量为 c_k，且 $c_k \times c_k' = 1$。

计算步骤如下。

(1) 将 x_1 分别置于 G_1，G_2，G_3，…，G_k 中，计算其 T 值，最后将 x_1 置于移动后能使 T 值最大的簇中。从 x_1 开始，x_1，x_2，x_3，…，x_p 分别选择使 T 值最大的簇。若 $x_i(i = 1, 2, \cdots, p)$ 的归属改变，则 $n_i = 1$，否则 $n_i = 0$。

(2) 重复步骤(1)，直到 $n_i(i = 1, 2, \cdots, p)$。

2 试验仪器与材料

2.1 试验仪器

采用美国 Thermo 公司生产的 Antaris 型傅里叶变换近红外光谱分析仪，配有积分球漫反射采样系统，RESULT 集成软件和 TQ Analyst6.2 定量分析软件；法国 Alliance 8 通道连续流动分析仪。

2.2 光谱采集条件

将一定量的样品装入样品杯，并用专用的砝码压紧，以空气作为光谱背景，采用积分球和旋转台测定样品 NIR 漫反射光谱。光谱的扫描区间为 4 000～10 000 cm^{-1}，设置扫描分辨率为 4 cm^{-1}，扫描 40 次取平均光谱。

2.3 试验材料及处理方法

387 个烤烟烟叶样品由贵州黄果树烟草集团公司技术中心提供，样品用旋风磨粉碎，过 40 目筛。全部样品采集完光谱后用逐步投影算法[11](Successive projections algorithm)对样品光谱进行筛选，最终获得 107 个具有代表性的样品用于建模。另外，从余下的样品中任选 15 个作为外部验证样品。对选出的定标样品和验证样品按照烟草行业标准方法由连续流动分析仪测定其总糖、还原糖、尼古丁的含量(称为化学值)。

3 结果与分析

3.1 潜变量聚类分析结果

光谱先经 4 阶 9 点平滑、多元散射校正、一阶微分处理和中心化处理[12]。对 100 个光谱进行层次聚类分析，层次聚类分析结果如图 1 所示。从图中可以看出，当聚类数从 5 变到 4 时，T 值有较大变化，因此选定 5 为聚类数，进行潜变量聚类分析的初始状态也一并产生。对层次聚类产生的 5 簇初始数据进行主成分分析，这 5 簇的第一主成分包含了原矩阵 86.38 %的方差，表明层次聚类分析已经使光谱信息有了较高的集中度。

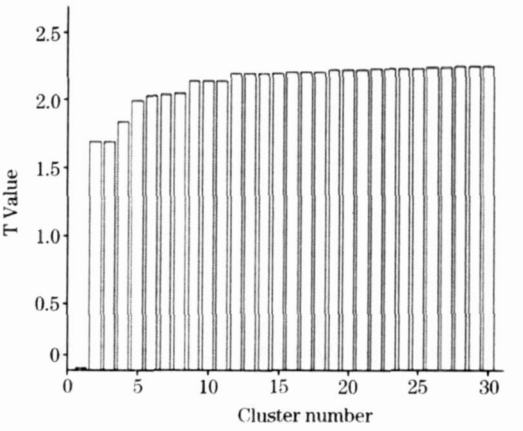

Fig. 1 T value under different cluster numbers

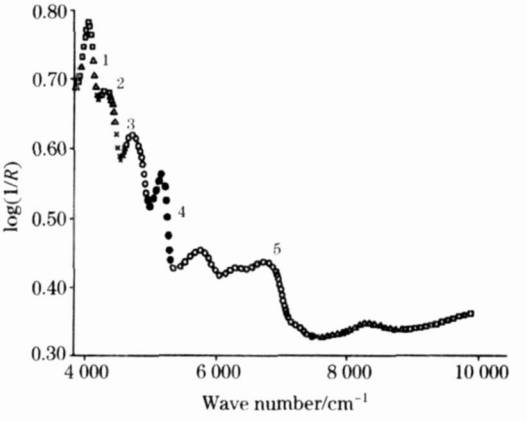

Fig. 2 Identification of the partition of the spectral variables

1: First cluster; 2: Second cluster; 3: Third cluster; 4: Forth cluster; 5: Fifth cluster

对光谱进行潜变量聚类分析处理，其结果如图 2 所示(横坐标为波数，纵坐标为吸光度值)。其中，第一簇区域范围比较复杂，分为 5 个独立的区域，最大的区域在 7 520～8 920 cm^{-1} 范围内，这一区域包含了键的二级倍频信息，其他几个区域则落在了 C—H 键与其他化学键的合频区，所以认为第一簇主要反映了 C—H 的二级倍频与合频信息。第二簇也分为 5 个区域，主要反映了芳香化合物中 C—H 键二级倍频、键的二级倍频及与其他化学键的合频信息。第三簇有两个区域，分别在 4 609～4 918 cm^{-1} 和 5 342～7 425 cm^{-1} 范围内，该簇所在的谱区是近红外光谱检测烟草化学指标最重要的谱区，它包含了 C—H，N—H，S—H，O—H 的一级倍频以及 O—H 与 N—H 的合频信息。第四簇分为两个区域，一个在 7 450 cm^{-1} 附近的，另一区域波数范围在 4 937

～5 323 cm^{-1}之间，主要分布在 C═O 键二级倍频的特征谱区。第五簇由四个区域组成，这四个区域大部分在芳香化合物的特征谱区，因此认为第五簇着重反映了芳香化合物的信息[13]。

3.2　建立定量分析模型

根据谱图各簇主要反映的化学信息，对不同的化学指标选取相应的簇建模。这里以总糖、还原糖和尼古丁为例，总糖、还原糖含有大量的 C—H，C═O，O—H，因此选择第三簇和第四簇为建模谱区；尼古丁含氮量很高，故选择第三簇作为建模谱区。

用 107 张样品集光谱对此三种指标分别建立全谱和选定谱区的模型。所有模型均用 PLS 方法建立，光谱经 4 阶 9 点平滑、一阶导数预处理。

3.3　模型评价

分别对模型从预测精密度和预测准确度两个方面进行评价。

模型预测的精密度是衡量模型抗干扰能力的重要指标，用预测样品间的平均相对误差（$D = \frac{1}{m}\sum_{i}^{m} \in_i$，其中 m 为验证集样品数，$\in_i$ 为各样品平行样间的平均相对误差）来衡量。本文对 15 个外部验证样品重复采集 3 次光谱，因此上式中 m 的值为 15，$\in_i$ 为每个验证样品的 3 张平行谱图预测值的平均相对误差。验证样品各指标的化学值、全谱模型和所选谱区模型的预测值分列于表 1 和表 2，各模型的值列于表 3。

模型预测的准确度从内部交互验证和外部样品验证两个角度进行，分别用 107 个校正样品的交互验证相关系数（$R_{training}$）、交互验证均方差（RMSECV）和 15 个外部验证样品的验证样品相关系数（R_{test}）和验证样品均方差（RMSEP）四个指标来评价，结果列于表 3。各指标计算公式如下

$$R = \sqrt{1 - \frac{\sum(\hat{C}_i - C_i)^2}{\sum(C_i - \bar{C})^2)}},$$

$$\mathrm{RMSECV} = \sqrt{1 - \frac{\sum(\hat{C}_i - C_i)^2}{n - 1}},$$

$$\mathrm{RMSEP} = \sqrt{1 - \frac{\sum(\hat{C}_i - C_i)^2}{m}}$$

其中 C_i 是标准化学法测定值，$\hat{C}_i - C_i$ 是模型预测值，$\bar{C}$ 是化学值的均值，n 是校正集样品数，m 为验证集样品数。

从表中数据可以看出，经过优选谱区建立的 3 个模型的 $R_{training}$ 和 R_{test} 值较全谱模型都高，RMSECV，RMSEP 和 D 值 3 个指标较全谱模型都低，说明其线性拟合能力、预测能力和稳定性均优于全谱模型。

4　结　论

本实验采用潜变量聚类分析方法，对烤烟样品的近红外光谱进行分析，将光谱分成若干个簇，每簇都能反映相对独立的信息，使得光谱区域可以在化学角度上进行解释，从而为建模时的光谱区域选择提供了参考依据。以总糖、还原糖和尼古丁为例，分别用全谱和优化后的谱区建立了定量分析模型，通过对比几个模型的参数，发现谱区经优化选择后的模型准确度和预测精度均高于全谱模型。由此可知，通过该算法可以优化选择模型的光谱范围，在实际应用或进行成分分析中有一定的参考价值。

Table 1　Prediction result with whole NIR spectrum region models

样品编号	总糖/ %				还原糖/ %				尼古丁/ %			
	化学值	预测值			化学值	预测值			化学值	预测值		
		光谱一	光谱二	光谱三		光谱一	光谱二	光谱三		光谱一	光谱二	光谱三
1	16.74	16.35	16.56	16.56	15.02	14.01	13.87	14.37	4.71	4.65	4.52	4.61
2	21.37	20.06	20.70	20.89	17.01	16.52	16.93	17.01	4.57	4.46	4.54	4.53
3	21.45	21.73	22.20	22.09	15.72	17.94	17.25	17.39	4.94	4.90	4.92	4.87
4	23.54	25.35	24.04	23.83	18.57	21.44	20.54	20.42	3.41	3.23	3.27	3.25
5	12.86	13.03	13.34	13.14	11.37	11.72	11.75	11.74	5.12	5.03	5.11	5.03
6	23.50	22.98	22.77	23.27	20.52	18.60	18.54	19.03	4.45	4.79	4.74	4.70
7	17.09	18.87	18.87	18.75	14.96	15.75	16.16	15.43	3.89	4.02	3.96	4.01
8	27.75	26.37	27.25	26.56	21.82	20.67	20.97	21.12	4.99	4.78	4.68	4.61
9	15.31	16.76	17.26	16.99	13.30	14.60	14.35	13.97	3.66	3.62	3.63	3.60
10	23.84	24.21	22.91	23.74	19.55	19.23	18.50	19.65	5.03	5.11	5.08	5.07
11	20.92	21.96	21.56	21.32	16.32	17.34	17.23	16.79	2.20	2.18	2.16	2.08
12	27.24	26.80	26.20	26.64	20.56	22.03	20.47	21.84	3.09	2.99	2.95	3.01
13	18.60	19.30	18.52	18.78	14.96	16.03	14.28	14.46	2.27	2.40	2.43	2.39
14	18.93	18.63	17.80	18.24	15.74	15.85	14.33	16.40	3.31	3.29	3.22	3.21
15	21.12	20.78	21.78	22.01	16.88	18.25	17.45	17.72	3.78	3.44	3.55	3.60

Table 2 Prediction result with selected NIR spectrum region models

样品编号	总糖/ %				还原糖/ %				尼古丁/ %			
	化学值	预测值			化学值	预测值			化学值	预测值		
		光谱一	光谱二	光谱三		光谱一	光谱二	光谱三		光谱一	光谱二	光谱三
1	16.74	16.72	16.83	16.80	15.02	14.57	14.44	14.86	4.71	4.65	4.72	4.66
2	21.37	21.10	21.07	21.17	17.01	16.90	16.94	16.98	4.57	4.63	4.68	4.65
3	21.45	21.51	21.47	21.43	15.72	16.66	16.95	16.75	4.94	4.97	4.99	4.95
4	23.54	24.67	24.14	23.95	18.57	18.93	18.49	18.61	3.41	3.35	3.44	3.37
5	12.86	12.48	12.88	12.57	11.37	11.20	11.24	11.24	5.12	4.97	5.00	4.97
6	23.50	23.12	23.98	23.44	20.52	19.23	19.04	19.33	4.45	4.65	4.56	4.65
7	17.09	16.98	16.53	16.84	14.96	15.69	15.73	15.49	3.89	3.87	3.89	3.96
8	27.75	27.20	27.25	27.57	21.82	21.53	21.46	21.47	4.99	5.02	4.94	4.91
9	15.31	15.59	16.21	15.76	13.30	14.12	13.82	13.60	3.66	3.71	3.71	3.68
10	23.84	23.53	24.01	23.87	19.55	19.21	18.82	19.57	5.03	5.30	5.37	5.29
11	20.92	20.92	20.72	20.71	16.32	16.90	16.85	16.48	2.20	2.24	2.16	2.16
12	27.24	27.31	27.09	27.21	20.56	20.98	20.76	20.89	3.09	2.97	3.00	3.09
13	18.60	18.77	18.50	18.53	14.96	15.45	14.82	14.90	2.27	2.26	2.26	2.22
14	18.93	18.06	18.33	18.31	15.74	15.41	14.98	15.84	3.31	3.24	3.29	3.23
15	21.12	20.80	20.64	21.00	16.88	18.09	17.73	17.87	3.78	3.68	3.70	3.74

Table 3 Parameters of models

评价指标	总糖		还原糖		尼古丁	
	全谱模型	选谱模型	全谱模型	选谱模型	全谱模型	选谱模型
$R_{training}$	0.977 1	0.995 5	0.917 2	0.975 1	0.987 4	0.994 4
RMSECV	1.09	1.05	1.43	1.05	0.14	0.13
R_{test}	0.977 8	0.992 7	0.941 2	0.967 9	0.993 2	0.994 0
RMSEP	0.92	0.39	1.17	0.63	0.16	0.11
D/ %	1.274	0.711	1.972	0.843	0.829	0.768

参 考 文 献

[1] LU Wan-zhen, YUAN Hong-fu, XU Guang-tong(陆婉珍, 袁洪福, 徐广通, 等). Modern Near-Infrared Spectroscopic Analysis Technique(现代近红外光谱分析技术). Beijing: China Petrochemical Press(北京: 中国石化出版社), 2000. 6.

[2] YAN Yan-lu, ZHAO Long-lian, LI Jun-hui, et al(严衍禄, 赵龙莲, 李军会, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2000, 20(6): 777.

[3] WANG Fang, CHEN Da, SHAO Xue-guang(王 芳, 陈 达, 邵学广). Tobacco Science & Technology(烟草科技), 2002, (5): 23.

[4] MA Xiang, WANG Yi, WEN Ya-dong, et al(马 翔, 王 毅, 温亚东, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2004, 24(4): 444.

[5] ZHANG Lu-da, ZHAO Li-li, ZHAO Long-lian, et al(张录达, 赵丽丽, 赵龙莲, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(8): 1227.

[6] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立, 袁洪福, 陆婉珍). Progress in Chemistry(化学进展), 2004, 16(4): 535.

[7] Forina M, Casolino C, Pizarro Millan C. Journal of Chemometrics, 1999, 13(2): 165.

[8] Brandye M Smith, Paul J Gemperline. Analytica Chimica Acta, 2000, 423(2): 167.

[9] Vigneau E, Qannari E M. Communications in Statistics-Simulation and Computation, 2003, 32(4): 1131.

[10] Vigneau E, Sahmer K, Qannariand E M, et al. Journal of Chemometrics, 2005, 19(3): 122.

[11] Heronides Adonias Dantas Filho, Roberto Kawakami Harrop Galvao, Mário Cesar Ugulino Araújo, et al. Chemometrics and Intelligent Laboratory Systems, 2004, 72(1): 83.

[12] 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.

[13] Donald A Burns, Emil W Ciurczak. Handbook of Near-Infrared Analysis, New York: Marcel Dekker, Inc., 1992.

Applied Study on Clustering of Variables around Latent Components Method in Wavelength Region Selection with Near-Infrared Spectroscopy

BAO Feng-wei[1], PENG Qian-rong[2*], LIU Jing-yan[3], CAI Yuan-qing[2], MAO Han-bing[2], TANG Ke[2], LÜ Yan-wen[2]

1. College of Chemical Engineering, Guizhou University, Guiyang 550003, China
2. Technology Center, China National Tobacco Guizhou Industrial Corporation, Guiyang 550003, China
3. College of Bioscience and Bioengineering, Hebei University of Science and Technology, Shijiazhuang 050018, China

Abstract The present paper introduced the principle of clustering of variables around latent components method, and used this method in selecting spectrum range of the NIR quantitative analysis models. Taking tobacco samples as experiment materials, we dealed with 107 sample spectra, divided the spectra into 5 clusters, and explained the information reflected by each of these 5 clusters in terms of chemistry. On this basis, we chose the corresponding wavelength range to set up the quantitative models of the total sugar, reducing sugar and nicotine by PLS method. Compared with the model based on the full NIR spectral range, Rtraining of the models based on the chosen spectral range rose from 0.977 1, 0.917 2 and 0.987 4 to 0.995 5, 0.975 1 and 0.994 4; Rtest rose from 0.977 8, 0.941 2 and 0.993 2 to 0.992 7, 0.967 9 and 0.994 0; RMSECV dropped from 1.09, 1.43, 0.14 to 1.05, 1.05 and 0.13, RMSEP dropped from 0.92, 1.17 and 0.16 to 0.39, 0.63 and 0.11 and the D value dropped from 1.274 %, 1.972 % and 0.829 % to 0.711 %, 0.843 % and 0.768 % for the total sugar, reducing sugar and nicotine, respectively. These data indicated that this method can improve the forecasting precision and stability of the model, so offers certain guidance on practical application.

Keywords Near infrared spectroscopy; Clustering of variables around latent components; Wavelengths selection

(Received May 10, 2006; accepted Aug. 20, 2006)

* Corresponding author