

基于近红外光谱和 SIMCA 算法的烟叶部位相似性分析

于春霞¹, 马翔², 张晔晖¹, 李军会^{1*}, 赵龙莲¹, 许莉¹, 温亚东², 王毅², 张录达³

1 中国农业大学信息与电气工程学院, 北京 100083

2 玉溪红塔集团技术中心, 云南 玉溪 653100

3 中国农业大学理学院, 北京 100193

摘要 烟叶的外观特征能反映其内在品质, 不同部位、品种、成熟度的烟叶在外观上表现出来的特征为烟叶生产加工提供了标准和依据。烟叶根据其在植株上着生位置不同可分为5个部位, 分别为: 顶叶、上二棚叶、腰叶、下二棚叶和脚叶。文章以2008年云南玉溪地区K326品种的5个部位各100份烟叶样品为实验材料, 采集了500个近红外光谱数据, 应用基于SIMCA算法的相似性分析数学模型, 对不同部位烟叶的近红外光谱进行了相似性分析。实验结果表明, 基于近红外光谱的烟叶部位相似性分析结果与实际烟叶部位之间的相似性程度是相符的, 部位之间相距越大其相似性越差, 同时相邻部位之间的相似程度是不同的。该研究对基于SIMCA算法的主成分确定进行了讨论, 并实现了类别间相似性的量化分析, 对于烟叶复烤配方的替代、调整以及进行工业分级的评价等具有指导意义。

关键词 烟叶部位; 相似性分析; SIMCA; 近红外

中图分类号: O657.3 文献标识码: A DOI: 10.3964/j.issn.1000-0593(2011)04-0924-04

引言

烟叶的外观特征能反映其内在品质, 不同部位、品种、成熟度的烟叶在外观上表现出来的特征为烟叶的工业分级提供了标准和依据^[1, 2]。在烟叶的质量管理中, 部位的特征是十分重要的。根据其在植株上着生位置不同烟叶可分为顶叶(T)、上二棚叶(B)、腰叶(C)、下二棚叶(X)和脚叶(P)5个部位。不同的烟叶部位在物理特性、化学成分以及评吸特性等方面都有着各自的特点^[3, 4], 比如从评吸特性上看, 越往顶部香气量越饱满, 香气浓度越往顶部越高, 劲头和刺激性也相应增加。在实际的烟叶生产加工过程中, 有时并不需要对烟叶的不同外部特征进行分类识别, 只要得到不同分类情况下类间的品质相似性关系, 以可靠的相似性分析结果为依据, 制定出符合烟叶产品内在品质特性规律的类别组合或替代, 对烟叶产品质量的提升、生产成本的降低以及复烤配方的设计与替代等都具有重要作用。本文以烟叶部位这样一种具有相对较明确的相似性关系的烟叶特征为例来验证基于近红外光谱技术和SIMCA算法来体现类间品质相似性关系的可靠性, 并期望得到更多有用的烟叶部位之间的相似性信息。

近红外光谱分析技术具有样品前处理简单、分析速度快、精度高、成本低等优点, 被广泛应用于农业、食品、石化、医疗等领域^[5-9]。近几十年来, 近红外光谱技术结合化学计量学方法越来越多的应用于烟草质量分析检测中^[10-13], SIMCA算法在近几年也开始用于烟草质量控制^[14], 例如: 王家俊等将基于FT-NIR分析技术的SIMCA建模用于卷烟配方过程质量监测等。本文将SIMCA算法应用于烟叶近红外光谱数据的相似性分析尚未见报道。

1 原理与算法

SIMCA (soft independent modeling of class analogy), 1976年由瑞典化学家Wold提出, 意为各类样本独立地建立局部“软”模式, 是一种类模型方法, 即对每类构造一个主成分分析的数学模型, 在此基础上进行样本的分析^[15-17]。

在SIMCA算法描述中对于第 q 类样本, 其主成分模型可表示为

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

式中, $\bar{x}_j^q$ 为类 q 中变量 j 的均值; A_q 为类 q 中主成分的数目;

收稿日期: 2010-07-02, 修订日期: 2010-10-06

基金项目: 国家“十一五”科技支撑计划项目(2006BAD11A12)和红塔集团科技项目资助

作者简介: 于春霞, 女, 1985年生, 中国农业大学信息与电气工程学院硕士研究生 e-mail: ychxkuailc@126.com

* 通讯联系人 e-mail: caunir@cau.edu.cn

t_{ia}^q 为样本 i 对主成分 a 的得分; l_{ja}^q 为变量 j 对主成分 a 的载荷; e_{ij}^q 为第 i 样本第 j 变量的残余误差。

于是, 类 q 的总残余方差为

$$S_0^2 = \sum_{i=1}^n \sum_{j=1}^p \frac{e_{ij}^2}{(n-A_q-1)(p-A_q)}$$

对于样本 i , 残余方差为

$$S_i^2 = \sum_{j=1}^p \frac{e_{ij}^2}{(p-A_q)}$$

在 n 维空间中类间的分离程度可利用每类的残余方差来确定。定义类 p 和类 q 间的距离为

$$D_{pq}^2 = \frac{S_{pq}^2 + S_{qp}^2}{(S_0^2)_p + (S_0^2)_q}$$

这里 S_{pq}^2 为用类 q 的模型拟合类 p 中各点所得的残余方差; S_{qp}^2 则为用类 p 模型拟合类 q 中各点的残余方差。 D_{pq}^2 值越小, 类间的相似程度越高。

由于 D^2 值是一个描述类间分离程度即表示类间差异性的一个指标数据, 且是一个非归一化数据, 其直观易读性不够好, 因此本文提出了在某一主成分下一种归一化的表示类间相似性的表达方法, 定义的类 p 和类 q 间的相似度公式为

$$\text{Sim}_{pq} = 1 - \frac{D_{pq}^2}{\sum_{k=1}^{p-1} \sum_{q'=1}^n D_{pq'}^2}$$

式中 n 为样本的类别数, Sim 值范围为 $0 < \text{Sim} < 1$, Sim 值越大表明相似性越高。

2 材料与方法

2.1 实验材料

2008 年云南省玉溪地区 K326 品种 5 个部位各 100 份烟叶样品, 共计 500 份。

2.2 实验仪器与测试方法

实验所用仪器为 MPA 型傅里叶变换近红外光谱仪(德国 Bruker 公司生产); 光谱谱区扫描范围为 $4\,000 \sim 12\,000\text{ cm}^{-1}$; 分辨率为 8 cm^{-1} ; 扫描次数为 64 次。

2.3 数据处理方法

光谱预处理方法采用一阶导数加 15 点平滑, 谱区范围选择 $4\,000 \sim 8\,000\text{ cm}^{-1}$ 。由于在基于 SIMCA 算法建立相似性分析模型时没有较好的优化确定主成分数目的方法, 因此本文提出了一种基于实验数据和经验的主成分数的确定方法, 其具体确定方法为: 得到的相邻主成分数下的相似性分析结果无显著差异的前提下, 采用的主成分数少为原则。

3 结果与分析

对 500 个烟叶样品的近红外光谱数据建立的相似分析数学模型中, 得到的 1~8 个主成分数下不同部位之间的 D^2 值见表 1。相邻主成分数下对应的两列 D^2 值之间的 R^2 值见表 2, R^2 值越接近于 1, 表示相邻主成分数下部位之间的 D^2 值差距越小, 当 R^2 值大于 0.99 时认为两个主成分数下得到的相似性分析结果没有显著性差异。

由表 2 中可以看出, 4~5 个主成分数之间对应的 R^2 值已达到 0.99 以上。遵循相邻主成分数下得到的相似性分析结果没有显著性差异和主成分个数尽量少的原则, 本文推荐主成分数为 4 下得到的 D^2 值来求解类间相似度值, 其对应的各烟叶部位间的相似度值见表 3。

由表 3 中的相似度数据可以得出:

- ①顶叶(T)与上二棚叶(B)之间的部位相似性最高。
- ②顶叶(T)与脚叶(P)之间的部位相似性最低。
- ③在相邻部位的相似性结果中腰叶(C)和下二棚叶(X)

Table 1 D^2 between tobacco plant parts with different PC numbers

PC	1	2	3	4	5	6	7	8
T-P	5.85	11.09	8.87	3.77	3.19	2.84	3.10	3.16
P-B	3.77	8.97	5.56	3.07	2.70	2.56	2.80	2.90
X-T	4.49	7.91	6.35	2.66	2.34	2.50	2.76	2.82
P-C	1.83	3.29	3.13	2.19	2.09	1.97	2.07	2.12
X-B	2.80	6.11	3.96	2.18	1.95	2.12	2.36	2.49
X-C	1.38	2.12	2.02	1.52	1.45	1.53	1.62	1.70
T-C	2.24	2.74	1.87	1.32	1.43	1.50	1.52	1.51
X-P	1.14	1.22	1.36	1.32	1.36	1.39	1.46	1.41
G-B	1.40	2.02	1.44	1.19	1.30	1.40	1.40	1.43
T-B	1.28	1.16	1.16	1.14	1.18	1.22	1.27	1.29

Table 2 R^2 between two columns of D^2 corresponding to adjacent PC numbers

	1-2	2-3	3-4	4-5	5-6	6-7	7-8
R^2	0.969	0.976	0.974	0.997	0.979	0.998	0.998

Table 3 Sim between two tobacco plant parts with PC numbers of 4

T-P	P-B	X-T	P-C	X-B	X-C	T-C	X-P	G-B	T-B
0.46	0.56	0.62	0.69	0.69	0.78	0.81	0.81	0.83	0.84

的相似性最低,甚至低于跨一个部位的腰叶(C)和顶叶(T)之间的相似性。

④上二棚叶(B)和腰叶(C)与下二棚叶(X)和脚叶(P)之间的相似程度都高于腰叶(C)和顶叶(T)之间的相似性,而低于顶叶(T)与上二棚叶(B)之间的相似性。

从总体上来看,相邻部位之间的相似性高于跨一个部位的相似性,跨一个部位的相似性高于跨两个部位的相似性,依此类推。这一结果与烟叶部位相似性的实际规律是一致的。

4 结果与讨论

本文以同一地区同一品种不同部位烟叶的近红外光谱为研究对象,对烟叶部位之间存在的品质特征的相似性进行了分析。分析结果表明,基于近红外光谱和 SIM CA 算法得到

的部位之间的相似性关系与实际烟叶部位的相似性关系一致,并得到了一组可靠的量化的烟叶部位间相似度数据。对于烟叶生产加工过程中工业分级的评价以及复烤配方的设计与调整具有重要意义。例如,通过不同年份间烟叶部位相似度数据的对比可以分析烟叶年份之间的差异,从而对复烤配方进行必要的调整,稳定烟叶产品的质量。本文中所采用的相似性分析方法可以推广应用到其他类别间的相似性分析,譬如不同产区、不同品种之间等,同时对其他农产品的类别特征的相似性分析也具有参考价值。

致谢: 本文中的研究工作都是在“红塔集团(技术中心) — 中国农大(信电学院) 烟叶近红外技术联合实验室”以及“中国农业大学信息与电气工程学院近红外光谱分析实验室”中进行的,在此对实验室中所有同仁所做的长期的、大量的研究工作表示感谢。

References

- [1] SHU Rui-xin, WANG Guo-dong, ZHANG Jian-ping, et al(束茹欣, 王国东, 张建平, 等). Tobacco Science & Technology(烟草科技), 2006, (8): 12.
- [2] XIE Juan, LU O Jian-qun, YAO He-ming, et al(谢娟, 罗建群, 姚鹤鸣, 等). Tobacco Science & Technology(烟草科技), 2008, (7): 42.
- [3] Ni Lijun, Zhang Liguang, XIE Juan, et al. Analytica Chimica Acta, 633(2009): 53.
- [4] ZHANG Jian-ping, CHEN Jiang-hua, SHU Rui-xin, et al(张建平, 陈江华, 束茹欣, 等). Acta Tabacaria Sinica(中国烟草学报), 2007, 13(5): 1.
- [5] Rao Yit-lan, Xiang Bing-ren, Zhou Xiaohua, et al. Journal of Food Engineering, 2009, 93(2): 249.
- [6] YAN Yan-lu, et al(严衍禄, 等). Basis and Application of Near Infrared Spectroscopy(近红外光谱分析基础与应用). Beijing: China Light Industry Press(北京: 中国轻工业出版社), 2005. 286.
- [7] HAN Liang-liang, MAO Pei-sheng, WANG Xin-guo, et al(韩亮亮, 毛培胜, 王新国, 等). J. Infrared Millin. Waves(红外与毫米波学报), 2008, 27(2): 86.
- [8] NIE Lixing, WANG Gang-li, LI Zhi-meng, et al(聂黎行, 王钢力, 李志猛, 等). J. Infrared Millin. Waves(红外与毫米波学报), 2008, 27(3): 205.
- [9] CHU Xiao-li, XU Yit-peng, LU Wan-zhen(褚小立, 许育鹏, 陆婉珍). Chinese Journal of Analytical Chemistry(分析化学) 2008, 36(5): 702.
- [10] MA Xiang, WEN Ya-dong, WANG Yi, et al(马翔, 温亚东, 王毅, 等). Tobacco Science & Technology(烟草科技), 2007, (3): 16.
- [11] He Z, Lian W, Wu M, et al. Journal of Near Infrared Spectroscopy, 2006, 14(1): 45.
- [12] McClure W F, Hamid Abdul, Giesbrecht F G, et al. Applied Spectroscopy, 1984, 38(3): 322.
- [13] WANG Dong-dan, LI Tian-fei, WU Yu-ping, et al(王东丹, 李天飞, 吴玉萍, 等). Journal of Yunnan University(云南大学学报), 2001, 23(2): 135.
- [14] WANG Jia-jun, LI Juan(王家俊, 李娟). Tobacco Science & Technology(烟草科技), 2008, (3): 5.
- [15] Panmanas Sirisomboon, Yuki Hashimoto, Munehiro Tanaka. Journal of Food Engineering, 2009, 93: 502.
- [16] XU Lu, SHAO Xue-guang(许禄, 邵学广). Methods of Chemometrics(化学计量学方法). Beijing: Science Press(北京: 科学出版社), 2004.
- [17] NI Yong-nian(倪永年). Applications of Chemometrics to Analytical Chemistry(化学计量学在分析化学中的应用). Beijing: Science Press(北京: 科学出版社), 2004.

Tobacco Plant Parts Similarity Analysis Based on Near-Infrared Spectroscopy and SIMCA Algorithm

YU Chun-xia¹, MA Xiang², ZHANG Ye-hui¹, LI Jun-hui^{1*}, ZHAO Long-lian¹, XU Li¹, WEN Ya-dong², WANG Yi², ZHANG Lu-da³

1. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

2. Technology Center of Yuxi Hongta Group, Yuxi 653100, China

3. College of Science, China Agricultural University, Beijing 100193, China

Abstract The appearance features of tobacco reflect its inner quality. Many factors, such as different plant parts, variety and maturity, provide standard and foundation for tobacco production processing. According to the different position of tobacco plant parts, tobacco plants leaves can be divided into five parts as tip, upper-middle, middle, lower-middle and priming leaf respectively. Five hundred tobacco leaf samples (100 each for one of five tobacco plant parts) from Yunnan province in 2008 were collected using near infrared spectroscopy, which all belong to tobacco varieties of K326. The similarity analysis of tobacco plant parts was carried out using mathematical model of SIMCA similarity analysis. The conclusion showed that the tobacco plant parts similarity results based on near-infrared spectroscopy corresponded to the relative tobacco plant parts in Yunnan province. The farther two tobacco plant parts were away from each other, the lower the similarity of corresponding parts was. And the similarity results of adjacent tobacco plant parts were different. The study discussed a method of confirming PC numbers and realized the quantitative similarity analysis between classes. It is instructive in replacement or adjustment of tobacco leaf blending and evaluation of tobacco industrial grading.

Keywords Tobacco plant parts; Similarity analysis; SIMCA; Near-infrared

(Received Jul. 2, 2010; accepted Oct. 6, 2010)

* Corresponding author

新书快讯:《傅里叶变换红外光谱分析》(第2版) 简介

由北京大学翁诗甫教授级高工编著的《傅里叶变换红外光谱分析》(第2版)一书已于2010年5月由化学工业出版社出版。

本书系统地介绍了红外光谱的基本概念、傅里叶变换红外光谱学的基本原理、傅里叶变换红外光谱仪的结构、傅里叶变换红外光谱仪附件原理和使用技术、红外光谱样品制备和测试技术、红外光谱数据处理技术、红外光谱的定量分析和未知物的剖析、基团的振动频率分析以及红外光谱仪的保养和维护技术。

本书可供教学、科研、厂矿企业、分析测试部门从事红外光谱分析测试工作者学习参考。本书既可作为红外光谱培训班的教学用书,也可作为高等院校与红外光谱相关的各学科教师、研究生和本科生的教学或参考用书。

本书为大32开,389页,书号978-7-122-07638-0,定价38元。由化学工业出版社出版(100011北京市东城区青年湖南街13号),详情可登录www.cip.com.cn查询。联系电话:010-64518888,64518899。