

基于近红外的多相偏最小二乘模型组合分析 实现茶叶原料品种鉴定与溯源的研究

周健¹, 成浩^{1*}, 曾建明¹, 王丽鸳¹, 韦康¹, 贺巍², 王伟峰¹, 刘栩¹

1. 国家茶树改良中心, 中国农业科学院茶叶研究所, 浙江 杭州 310008

2. 南京农业大学茶学系, 江苏 南京 210095

摘要 为实现茶叶的品种溯源, 以近红外光谱分析技术为基础, 尝试采用 PLS、欧氏距离等方法的组合来实现茶叶的特殊原料品种的鉴别、具体品种识别分析。结果表明: 研究中以 4 个品种(龙井 43, 群体种、迎霜和乌牛早)的茶树鲜叶为制成的茶叶样本为材料, 分别建立了 4 个不同原料加工成的茶叶样本的 PLS 鉴定模型, 对定标集样本识别的准确率分别为 89.8%, 90.9%, 96.1% 和 99.5%, 验证集的识别准确率分别为 87.1%, 84.2%, 96.1% 和 97.5%; 而采用 4 个品种鉴定模型的组合分析“初次识别”并且结合欧氏距离法进行“二次识别”后, 最终对 4 种不同品种茶叶样本的原料品种的识别准确率分别为 90.3% (定标集) 和 83.5% (验证集)。该研究为实现特殊品种加工成茶叶的鉴定以及成品茶的品种溯源提供了一种参考方法。

关键词 茶; 品种; 溯源; 近红外; 偏最小二乘法; 欧氏距离

中图分类号: O657.3 文献标识码: A DOI: 10.3964/j.issn.1000-0593(2010)10-2650-04

引言

当前, 农产品质量安全问题日益突出。农产品质量追溯是农产品质量体系建设的一个重要内容。建立我国农产品质量的可追溯制度, 有利于控制农产品(食品)质量, 并可及时、有效地处理质量问题, 追究责任, 最终提高安全水平。但是可追溯制度是建立在已经建立一个覆盖农产品(食品)从初级产品到最终消费品各个阶段资料的信息库的基础上通过良好的标识管理来实现的^[1,2]。而针对那些缺乏标识信息或者尚未实现详细资料登记的农产品, 要实现其产品的产地、品种等相应生产信息溯源, 则需要建立一种可靠准确的鉴别分析方法, 从而在无明显资料的条件下实现其可追溯性。

就茶叶来看, 目前茶叶还尚未全面建立有效的可追溯系统, 因此建立一种针对茶叶产地、原料品种等生产信息的鉴别分析技术, 对于实现茶叶产地、品种的溯源, 加强地理标志产品以及茶树新品种的保护, 都具有非常重要的意义。对于茶叶来说, 品种是影响茶叶质量的非常重要的因素之一。由于品种适制性等原因, 象龙井茶之类的传统名茶的品质与

其原料的品种关系密切, 不同品种制作的产品往往会有较大的品质差异, 从而对其经济价值也产生较大的影响。因此品种溯源是茶叶可追溯制度中的重要内容^[3]。

与其他方法相比, 近红外光谱分析技术除了具有快速、非破坏性、无污染等优点外, 它还具有信息量大、光谱特性稳定、重现性好的特点^[4], 在品种分析^[5], 产地识别^[6]等方面具有广泛的应用前景。作者曾经尝试利用近红外技术实现了对特殊产地、品种茶叶的识别^[7,8]。在之前研究的基础上, 本研究进一步以近红外技术为基础, 利用不同的化学计量学方法如 PLS^[9,10]、欧氏距离^[12]等方法的结合, 来尝试建立一种可对市场上的商品茶的原料品种实现准确识别的方法, 为今后实现茶叶产品的特殊品种鉴定, 品种溯源等奠定基础。

1 原料与方法

1.1 实验材料

实验采用的 838 份样本全部在中国的浙江省采集。均为按照龙井茶加工工艺生产的龙井茶及扁形茶, 其中包含西湖龙井 232 份, 钱塘龙井 204 份, 越州龙井 240 份, 以及其他

收稿日期: 2009-12-16, 修订日期: 2010-03-22

基金项目: 国家科技基础条件平台项目(2005DKA1002), 现代农业产业技术体系建设专项资金和浙江省科技计划项目重大科技专项优先主题农业项目(2008C12070)资助

作者简介: 周健, 1979 年生, 中国农业科学院茶叶研究所助理研究员 e-mail: zjph263@126.com

* 通讯联系人 e-mail: chenghao@mail.tricaas.com

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

地区的扁形茶样本 162 份, 样本的原料为龙井 43、迎霜、乌牛早、以及群体种等几个主要栽培品种(表 1)。样本均从茶农或者茶叶生产厂家直接采集, 为研究者指定相应的茶树品种后由企业或茶农将该品种的鲜叶原料直接炒制成品茶后采集。样本采用三分法分为 2 部分: 即样本随机排列后每 3 个样本中的 2 个选入定标样本用于建立识别模型。剩下的一个样本选入验证集作为未知样本来对识别模型进行验证, 因此包括 559 个定标集样本和 279 个验证集样本。

Table 1 Origin, material cultivar and number of samples in research

	龙井 43	乌牛早	迎霜	群体种	合计
非龙井地区	42	17	33	70	162
钱塘龙井产区	59	37	30	78	204
西湖龙井产区	110	0*	0	122	232
越州龙井产区	80	27	50	83	240
合计	291	81	113	353	838

* : 西湖龙井地区茶树的栽培品种只有龙井 43 与群体种, 故该地区只采集这两个品种为原料的茶叶样本。

1.3 数据分析

1.3.1 4 种不同原料茶叶的同步识别

分别对以龙井 43、群体种、迎霜和乌牛早为原料制成的样本赋值, 其中龙井 43 样本为 1, 群体种样本为 2, 迎霜样本为 3, 乌牛早样本为 4, 以定标集样本作为建模样本, 利用 PLS 法建立 4 种不同原料茶叶的同步识别模型, 根据模型的计算结果, 按照 0~1.5 的识别归类为龙井 43 样本、1.5~2.5 的识别归类为群体种样本、2.5~3.5 的归类为迎霜样本、3.5~4.5 的归类为乌牛早样本的识别标准, 分别对定标集的样本和验证集的样本进行原料品种的识别, 计算模型对定标集和验证集样本原料的分析准确率, 作为模型的评价标准。比较不同的光谱预处理方法(smoothering、MSC、一阶导数、二阶导数)对模型的优化效果, 确定建立模型的最佳条件。

1.3.2 针对单个特定原料品种的茶叶样本鉴定模型的构建

龙井 43 品种样本鉴定模型的构建: 分别对龙井 43 品种样本赋值为 1, 其他样本赋值为-1, 以定标集样本的光谱作为建模样本, PLS 法建立针对龙井 43 品种样本的鉴别模型, 根据模型的计算结果, 按照计算结果为 0~2 的归类为龙井 43 品种样本, -2~0 的归类为非龙井 43 品种样本, 根据此标准分别对定标集的样本和验证集的样本的原料是否为龙井 43 进行判别分析, 计算模型对定标集和验证集样本识别准确率, 作为评价模型的优劣标准。

其他 3 种原料品种制成的茶叶也按照上述过程建立其特定的识别模型, 以模型对定标集的样本和验证集的样本的识别准确率进行作为模型的评价指标。并且在模型的建立过程中利用不同的光谱预处理方法(smoothering、MSC、一阶导数、二阶导数等)对光谱进行预处理, 比较不同预处理条件下构建的模型的识别效果, 分别确定 4 个模型的最佳条件。最终建立 4 个分别针对龙井 43、群体种、迎霜和乌牛早的鉴别模

型, 以分别实现对以这 4 种品种为原料制成的茶叶的鉴定。

1.3.3 单独模型组合分析(“初次识别”)识别茶叶原料品种

按照 1.3.2 中的定义标准, 4 个模型对于 4 种原料制成的茶叶的标准分析结果应如表 2 所示。分别以构建的龙井 43、群体种、迎霜和乌牛早为原料制成的茶叶的鉴定模型对样本进行预测分析, 将模型对样本的分析结果与表 2 的 4 个原料品种的识别标准进行比对来预测样本的具体品种。经过比对样本可总体分为 3 类, 除正确识别样本(4 个模型的分析结果为一正三负, 识别正确)和错误识别样本(4 个模型的分析结果为一正三负, 但是识别错误)外, 还有未能分类样本(4 个模型的分析结果为二正二负, 三正一负, 四正, 四负等 4 类, 未能识别)。统计经过“初次识别”定标集和验证集样本的正确识别数, 计算识别准确率, 以此评价模型组合分析(“初次识别”)对样本具体原料品种的识别效果。

Table 2 Standard result of 4 teas predicted by 4 models

	龙井 43 模型	群体种模型	迎霜模型	乌牛早模型
龙井 43	+	-	-	-
群体种	-	+	-	-
迎霜	-	-	+	-
乌牛早	-	-	-	+

+ : 模型预测结果为> 0, < 2; - : 模型预测结果为> - 2, < 0

1.3.4 对初次识别中的“未能分类样本”利用欧氏距离法进行“二次识别”

对初次识别后的未能正确识别的样本利用其分别与龙井 43、群体种、迎霜和乌牛早的样本标准坐标间的欧氏距离的大小来进行进一步识别。具体过程如下: 按照 1.3.2 中的赋值标准, 由龙井 43、群体种、迎霜和乌牛早的 4 个鉴别模型分别对样本进行分析, 每个可形成一个在四维欧氏空间的坐标, 而 4 种茶叶(龙井 43 茶叶样本, 群体种茶叶样本、迎霜茶叶样本和乌牛早茶叶样本)在四维欧氏空间的标准坐标应分别为(+1, -1, -1, -1), (-1, +1, -1, -1), (-1, -1, +1, -1), (-1, -1, -1, +1)。利用 4 个模型对样本进行计算分析获得的坐标为样本坐标, 计算样本坐标与 4 种茶叶的样本标准坐标间的欧氏距离, 当样本坐标与某原料品种样本标准坐标间的欧氏距离最小时, 则判定其为样本为该原料品种制成。其中待识别样本坐标与特定原料茶叶样本标准坐标间欧氏距离的计算方法如下

$$D = \sqrt{\sum (X_{is} - X_{in})^2} \quad i = 1, 2, 3, 4$$

X_{is} 为 4 个模型计算获得的样本在四维欧氏空间的坐标, X_{in} 为特定品种的茶叶样本在四维欧氏空间的基准坐标。

1.4 软件

本研究中光谱的采集在近红外仪的自带软件 OPUS QUANT 2 上进行。PLS 法建立不同原料品种的识别模型及其对定标集和验证集样本的预测分析的运算过程在软件 Unscrambler 9.0 上进行。欧氏距离的计算以及样本的特殊原料品种的鉴定、具体品种识别分析在软件 Excel 2003 中进行。

2 结果与分析

2.1 4 种茶叶的原料品种同步分析

首先按照 1.3.1 中的步骤尝试采用 PLS 法构建一个同步分析模型来同时区分 4 种不同原料品种的茶叶。表 3 中显示了不同的预处理方法对构建 4 种不同原料品种的茶叶的同步分析模型的影响及模型对样本的识别效率。从表中可以看出: 经过对光谱进行预处理后, 模型的识别效率没有明显提高, 其中利用原始光谱构建的模型对定标集的 559 个样本和验证集的 279 个样本的识别准确率为 69.8% 和 63.4%。而利用 MSC, Smoothing, 一阶导数、二阶导数等预处理后构建模型的对定标集的识别准确率分别为 72.6%, 71.2%, 45.6% 和 44.0%, 对验证集样本的识别准确率分别为 65.6%, 64.5%, 44.4% 和 39.1%。因此可以看出, 光谱预处理不能有效提高同步分析模型的识别效率, 相反经过导数(一阶、二阶导数)处理后的模型的识别效率大为降低。而最终同步分析模型对样本的原料品种识别的准确率最高也仅为 MSC 预处理条件下的 72.6% (定标集) 与 65.6% (验证集), 识别效果不理想。

Table 3 Establishment of a synchronous analysis model for recognizing 4 teas

	建模因子	R^2	定标集样本识别准确率/%	验证集样本识别准确率/%
原始光谱	17	0.819	69.8	63.4
MSC	16	0.820	72.6	65.6
Smoothing	18	0.930	71.2	64.5
一阶导数	6	0.583	45.6	44.4
二阶导数	3	0.416	44.0	39.1

2.2 特定原料品种的茶叶样本鉴定模型的构建

由于利用一个同步分析模型对 4 种不同原料的茶叶进行同时区分的效果不佳, 继续尝试对不同原料的茶叶分别建立识别模型来实现对特定原料制成的茶叶的鉴别。经过对不同的预处理方法对分别构建龙井 43、群体种、迎霜和乌牛早为原料的 4 种茶叶的鉴定模型效果的比较, 确定了模型建立的最佳条件(表 4)。利用这 4 个模型分别对定标集和验证集的样本是否属于龙井 43、群体种、迎霜和乌牛早为原料制成按照 1.3.3 中的方法分别进行鉴定。结果表明: 4 个模型的对定标集样本的识别准确率分别为: 89.8%, 90.9%, 96.1% 和 99.5%。对验证集样本的识别准确率为 87.1%, 84.2%, 96.1% 和 97.5%。说明利用 PLS 法可建立针对某个特定原料品种制成的茶叶样本的鉴别模型, 具有良好的识别效果, 可用于特定品种制成的茶叶的鉴别。

2.3 利用 4 个原料品种的识别模型进行组合分析对茶叶具体品种进行“初次识别”

在建立针对 4 个不同原料品种的茶叶的鉴别模型后, 按照 1.3.3 中的识别标准利用这 4 个模型进行组合分析来进行茶叶具体原料品种的识别(“初次识别”)。结果发现: 模型组合分析对茶叶样本原料品种的识别准确率(见表 5)分别达到

了 83.4% (定标集) 和 74.2% (验证集)。与同步分析的 72.6% (定标集) 和 65.6% (验证集) 相比, 利用多品种模型组合分析进行茶叶原料品种的识别显然具有更良好的效果。

Table 4 Optimization of 4 models for recognizing of samples separately

模型	最佳预处理方法	建模因子	R^2	识别正确率/%	
				定标集	验证集
龙井 43	Smoothing	16	0.733	89.8	87.1
群体种	MSC	17	0.805	90.9	84.2
迎霜	原始光谱	16	0.813	96.1	96.1
乌牛早	MSC	19	0.905	99.5	97.5

Table 5 The identification result of tea material cultivar with first recognition

	定标集	验证集
样本数量	559	279
“初次识别”准确的样本数	466	207
“初次识别”准确率/%	83.4	74.2

2.4 “初次识别”后未能识别样本的“二次识别”

按照 1.3.4 对“初次识别”后的未识别样本继续进行“二次识别”。结果发现: 经过“初次识别”后的定标集的 559 个样本中错误识别样本为 37 个, 未能分类样本为 56 个, 其中 39 个可通过“二次识别”后实现正确的原料品种识别, 最终识别错误的样本数为 54 个; 而验证集的 279 个样本中错误识别样本为 34 个, 未能分类样本为 38 个, 其中 26 个可通过“二次识别”后实现正确的原料品种的识别, 最终识别错误的样本数为 46 个。因此经过二次识别后定标集和验证集的最终的识别的准确率分别为 90.3% 和 83.5%, 比“初次识别”的分别提高了 6.9% 和 9.3% (表 6), 说明了利用欧氏距离法进行二次识别后可有效提高识别的准确率。

Table 6 The identification result of samples from different area with second recognition by Euclidean distance method

	定标集	验证集
初次识别后错误样本数	37	34
初次识别后未识别样本数	56	38
二次识别可正确识别样本数	39	26
初次识别后未识别样本可经二次识别正确识别的比例/%	69.6	68.4
二次识别后识别错误样本数	54	46
最终识别正确率/%	90.3	83.5

3 结论

本研究证实了利用近红外技术结合 PLS、欧氏距离等分析方法实现茶叶特定原料品种的鉴别以及茶叶原料品种溯源识别的可行性。研究中利用 PLS 法分别建立了针对龙井 43、群体种、迎霜以及乌牛早的 4 个品种样本的鉴别模型, 对定

标集的 559 个样本的识别准确率分别为 89.8%, 90.9%, 96.1% 和 99.5%, 而对验证集的 279 个未知样本的识别准确率分别为 87.1%, 84.2%, 96.1% 和 97.5%, 实现了对特定原料品种加工成的茶叶的识别。在此基础上, 利用龙井 43、群体种、迎霜以及乌牛早的 4 个识别模型的组合分析以及欧氏距离法的二次判别后, 可将由龙井 43、群体种、迎霜以及乌牛早为原料加工成的茶叶分别区分开来, 定标集与验证集

的识别准确率最终分别为 90.3% 和 83.5%, 证实了利用近红外技术结合 PLS、欧氏距离等分析方法实现了对茶叶的原料品种的溯源识别的可行性。本研究的结果证实利用近红外技术结合 PLS、欧氏距离等分析方法的组合, 不仅可以有效鉴别某个特定原料加工成的茶叶, 而且可利用多个品种鉴别模型的组合分析实现对不同原料品种加工成的茶叶的识别, 为最终实现茶叶的品种溯源提供理论依据。

参 考 文 献

- [1] DU Guoming (杜国明). Guangdong Agricultural Sciences (广东农业科学), 2008, (2): 101.
- [2] LI Guangling, ZHANG Lili, WU Yanning, et al (李广领, 张利丽, 吴艳兵, 等). Hunan Agricultural Sciences (湖南农业科学), 2009, (2): 120.
- [3] CHENG Hao, WANG Liyuan, ZHOU Jian, et al (成浩, 王丽鸳, 周健, 等). Scientia Agricultura Sinica (中国农业科学), 2008, 41 (8): 2413.
- [4] Williams P, Norris K. Near infrared Technology in the Agricultural and Food Industries (Second Edition). Minnesota, USA: The American of Cereal Chemists, Inc. St. Paul, 2001.
- [5] CHEN Quansheng, ZHAO Jiewen, ZHANG Haodong, et al (陈全胜, 赵杰文, 张海东, 等). Food Science (食品科学), 2006, 27(4): 186.
- [6] ZHANG Xiaohui, LIU Jianxue (张晓慧, 刘建国). Laser & Infrared (激光与红外), 2008, 38(4): 342.
- [7] ZHOU Jian, CHENG Hao, HE Wei, et al (周健, 成浩, 贺巍, 等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2009, 29(5): 1251.
- [8] ZHOU Jian, CHENG Hao, YE Yang, et al (周健, 成浩, 叶阳, 等). Acta Optica Sinica (光学学报), 2009, 29(4): 1117.
- [9] Ortiz C, Zhang D, Xie Y, et al. Anal. Biochem., 2004, 332: 245.
- [10] Cozzolino D, Chree A, Scaife J R, et al. J. Agri. Food. Chem., 2005, 53(11): 4459.
- [11] Lutz U, Lutz R W, Lutz W K. Anal. Chem., 2006, 78(13): 4564.
- [12] Koolgaard J P, Lawoko C R O. Commun. Statist. - Theory Methods A, 1996, 25: 2989.

Study on Identification and Traceability of Tea Material Cultivar by Combined Analysis of Multi-Partial Least Squares Models Based on Near Infrared Spectroscopy

ZHOU Jian¹, CHENG Hao^{1*}, ZENG Jianming¹, WANG Liyuan¹, WEI Kang², HE Wei², WANG Weifeng², LIU Xu²

1. National Center for Tea Improvement, Tea Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou 310008, China

2. Department of Tea Science, Nanjing Agricultural University, Nanjing 210095, China

Abstract The present study attempted to achieve the identification and traceability of tea material cultivar through combination of multi-partial least squares models and Euclidean distance, etc. The results indicate that with the samples manufactured with tea fresh leaves of cultivar Longjing 43, Qunti, Yingshuang and Wuniuzao as the analysis objects, 4 models were established in this study so as to identify the tea with material cultivar being tea fresh leaves of cultivar Longjing 43, Qunti, Yingshuang and Wuniuzao separately by PLS. Their accuracy rate of identification of samples in the calibration set were 89.8%, 90.9%, 96.1% and 99.5%, respectively, while those in test set were 87.1%, 84.2%, 96.1% and 97.5%, respectively. After the “first identification” through the combined analysis of the four models for identification of tea material cultivar and the “second identification” adopting the Euclidean distance, the accuracy rate of material cultivar recognition for the tea samples was 90.3% (calibration set) and 83.5% (test set), respectively. This study provided a reference method for the identification of tea manufactured with a specific material cultivar and the material cultivar traceability of the manufactured tea.

Keywords Tea; material cultivar; Traceability; Near infrared; Partial least squares; Euclidean Distance

* Corresponding author

(Received Dec. 16, 2009; accepted Mar. 22, 2010)