SIMCA模式识别方法在近红外光谱识别茶叶中的应用

陈全胜[1]，赵杰文[1]，张海东[1,2]，刘木华[1,3]
(1.江苏大学生物与环境工程学院，江苏 镇江 212013；2.云南农业大学工程技术学院，云南 昆明 650201；3.江西农业大学工学院，江西 南昌 330045)

摘 要：茶叶快速准确识别方法研究是当前茶叶行业亟待解决的一项重大课题。本研究采用一种近红外光谱结合SIMCA模式识别方法对茶叶进行识别与分类。研究结果表明，选取6500～5300cm^{-1}波长范围内的光谱，通过标准归一化(SNV)预处理后，利用SIMCA的模式识别方法分别为龙井、碧螺春、祁红和铁观音等四类茶叶建立了类模型。主成分数分别为4 、5 、2 和3时，类模型对未知样本的识别效果最佳。在α=5%的显著性水平下，四类模型的对未知茶叶样本的识别率分别是90%、80%、100%和100%，拒绝率全是100%。本论文为快速准确识别茶叶提供了一种新思路。

关键词：茶叶；近红外光谱；SIMCA；识别

Application of Near Infrared Reflectance Spectroscopy to the Identification of Tea Using SIMCA Pattern Recognition Method

CHEN Quan-sheng[1], ZHAO Jie-wen[1], ZHANG Hai-dong[1,2], LIU Mu-hua[1,3]
(1. School of Biological and Environmental Engineering, Jiangsu University, Zhenjiang 212013, China;
2. Faculty of Engineering and Technology, Yunnan Agricultural University, Kunming 650201, China;
3. Engineering College, Jiangxi Agricultural University, Nanchang 330045, China)

Abstract: It is an urgent affair to think up a quick and precise method in the identification of tea varieties. A rapid tea identification method by near infrared reflectance spectroscopy coupled with pattern recognition based on SIMCA was proposed in this paper. In the spectra region between 6500cm^{-1} and 5300cm^{-1}, four predictive models of Longjing tea, Biluochun tea, Qihong tea and Tieguanyin tea were built separately by the standard normal variate (SNV) preprocessing method with SIMCA pattern recognition method. The results showed that four models are the best when 4, 5, 2 and 3 principal components were used separately in building models. Under the α=5% significance level, the identification rates of four models for the unknown samples are 90%, 80%, 100% and 100% in turn by means of NIR wave lengths, while, the rejection rates of four models are all 100%. A new idea by the quick and precise identification of tea was offered in this paper.

Key words: tea; near-infrared spectroscopy; Soft Independent Modelling of Class Analogy (SIMCA); identification
中图分类号：0657.33 文献标识码：A 文章编号：1002-6630(2006)04-0186-04

中国是茶叶的故乡，盛产许多品种的茶叶。目前中国的茶叶市场相对混乱，特别在名优茶市场，以次充好以假乱真的现象比较严重，这既损害了消费者的利益，也不利于中国茶叶品牌的保护。研究快速、准确的识别方法，对于维护中国茶叶品牌，提高茶叶品质有着直接的现实意义。

传统的茶叶识别方法是感官评定法和化学方法。其中，感官评定的结果受人为因素和外界环境的干扰很大，影响到结果的客观性；化学方法虽然能够准确地识别茶叶，但是繁琐的步骤和昂贵的费用使它不能应用到

收稿日期：2005-06-20
基金项目：国家高技术“863”计划资助项目(2002AA248051)；国家自然科学基金资助项目(30370813)
作者简介：陈全胜(1973-)，男，博士研究生，主要从事食品与农产品无损检测研究。

茶叶的快速识别上。近红外漫反射光谱(NIR)分析具有速度快、成本低以及结果重现性好等优点。国内外学者先后利用近红外光谱方法定性和定量地分析了茶叶中蛋白质、咖啡碱、氨基酸、多酚类以及水分的含量[1~3]，但是近红外光谱方法在茶叶识别上的应用研究还很少。鉴于此，本研究尝试了将近红外光谱结合模式识别的方法应用到茶叶的快速识别中，该方法在石油[4]和中草药[7,8]的识别和分类上得到了许多成功地应用。希望这种方法能成为一种茶叶快速识别的新技术得到广泛地应用。

1 材料与方法

1.1 材料

实验所用的材料是龙井、碧螺春、祁红和铁观音四种中国名茶。为了取样均匀，实验前先将每一个品种的茶叶分别用咖啡粉碎机粉碎过筛，然后在每一个品种的茶叶中，按照四分法原则，随机称取10g作为一个样本，其中，龙井、碧螺春和铁观音分别取30个样本，祁红取18个样本。实验中所用茶叶均来自安徽农业大学茶叶系，出产日期都在2004年5～7月份。

1.2 光谱采集

实验所用的近红外检测系统主要是近红外光谱仪(Nexus 670 FT-IR 美国Nicolet公司)。扫描范围：11000～3800cm^{-1}；扫描次数：64次；分辨率：4cm^{-1}。实验时，保持室内的温度和湿度基本一致，将样本倒入样品杯中，充分压实。每一个样本在不同时间，不同位置分别采集三次，取3次采集的平均值作为该样本的原始光谱。

1.3 分析方法

模式识别一般是根据“物以类聚”的原则进行样本的分类,目前所采用的方法主要有马氏距离法、线性学习机法、K-均值法及SIMCA 等方法[5,6]。由于茶叶近红外光谱特征变量数多，茶叶类型复杂，本研究最终选用SIMCA 模式识别方法。

SIMCA (Soft Independent Modeling of Class Analogy)方法实际上是相似分析方法，该方法在光谱[8]、色谱[9]的定性分析中得到了广泛的应用。在本研究中，SIMCA模式识别方法首先针对每一类样品的光谱数据矩阵进行主成分分析，建立主成分回归类模型，然后依据该模型对未知样品进行分类，即分别试探将该未知样本与各样本的类模型进行拟合，以确定未知样本类别。具体的分析原理和步骤如下。

(Ⅰ)建立类的主成分回归模型，对于第q类样本中的第k个样本矢量$X_{ik}^{(q)}$可用如公式(1)的主成分的回归模型表示：

$$x_{ik}^{(q)} = a_i^{(q)} + \sum_{a=1}^{A_q} \beta_{ia}^{(q)} \theta_{ak}^{(q)} + \varepsilon_{ik}^{(q)} \qquad (1)$$

式中，$a_i^{(q)}$：变量的均值；A_q：主成分数；$\beta_{ia}^{(q)}$：变量i在主成分a上的载荷；$\theta_{ak}^{(q)}$：样本k关于主成分a的得分；$\varepsilon_{ik}^{(q)}$：偏差。

(Ⅱ)用所建的q类模型拟合未知样本p，用拟合残差$S_p^{(q)^2}$表示未知样本p与q类模型的相似性，计算q类模型的总体偏差$S_0^{(q)^2}$和拟合残差$S_p^{(q)^2}$，分别如公式(2)和公式(3)所示：

$$S_0^{(q)^2} = \sum_{k=1}^{n_q} \sum_{i=1}^{m} (\varepsilon_{ik})^2 / [(n_q - A_q - 1)(m - A_q)] \qquad (2)$$

$$S_p^{(q)^2} = \sum_{i=1}^{m} (\varepsilon_{ip})^2 / (m - A_q) \qquad (3)$$

式中，n_q：第q类模型的样本数目；m：变量数；ε_{ip}：偏差。

(Ⅲ)由公式(4)和(5)计算值与临界值F_0，通过F显著性检验判断未知样本p是否属于该类模型。如果$F < F_0$则样本p属于该类模型；否则样本p不属于该类模型。

$$F = S_p^{(q)^2} / S_0^{(q)^2} \qquad (4)$$

$$F_0 = F_\alpha((m - A_q), (n_q - A_q - 1)(m - A_q)) \qquad (5)$$

式中，α：显著性水平；$(m-A_q)$，$(n_q-A_q-1)(m-A_q)$为F分布的自由度。

所有的数据分析都是基于TQ Analysis V6(Nicolet近红外系统自带)、Matlab V6.5 和Unscrambler V9.2(CAMO A/S)的软件平台。

2 结果与分析

2.1 波长范围的选择

图1是四种茶叶的原始光谱图(a)和二阶导数光谱图(b)，从图1中可以看出原始光谱在波数为5155cm^{-1}(波长1940nm)的附近有一个明显的吸收峰，二阶导数光谱在波数为5155cm^{-1}(波长1940nm)和6944cm^{-1}(波长1440nm)的附近有明显的波动。因为纯水中的O-H伸缩振动的一级基频区在6944cm^{-1}(波长1440nm)附近，它的一个合频区在5155cm^{-1}(波长1940nm)附近，在这两个波长附近是水分吸收的敏感区，从图1中可以看出在这两个区域，水分对茶叶的近红外图谱的影响还是很大的。本研究所用的样本都是干茶(一般干茶中的水分含量在5%左右)，为了减少水分的影响，选择光谱波长范围尽量避开水分吸收峰的特征波长区。本研究有比较地选用了各段的波长进行了分析，结果显示选用6500～5300cm^{-1}(波长1538～1818nm)范围内的光谱数据既避开了水分的影响且取得了较好的实验结果。

2.2 数据的预处理

实验中，茶叶样本粒径的大小和均匀度不能保证完全一致，这些都对光的漫反射有一定影响；同时样本的密实度也影响了光在样品中的传播。因此，需要对样

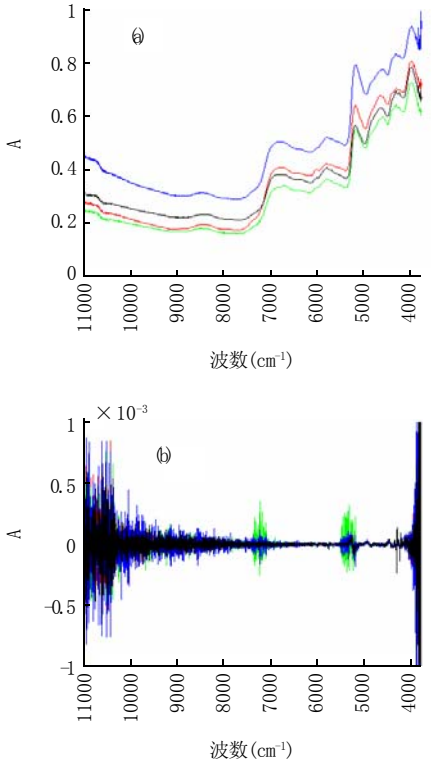

图1 四种茶叶的原始光谱(a)图和二阶导数光谱(b)图

Fig.1 The raw spectra (a) and second derivative spectra (b) of four categories tea

本的原始光谱数据预处理。实验中，运用了多元散射校正(MSC)、标准归一化(SNV)、一阶导数和二阶导数等四种预处理方法，通过对比发现，用SNV或和MSC的方法明显优于一阶导数和二阶导数预处理方法。本实验最终采用了标准归一化(SNV)的预处理方法。

2.3 模型建立与主成分数确定

在6500～5300cm⁻¹波数范围内，分别截取这108个茶叶样本的近红外光谱，并对这些光谱数据组成的矩阵进行主成分分解，以其前3个得分向量作图，结果见图2。从图2中可以看出，这些茶叶有明显的聚类趋势，因此本研究在主成分分解的基础上，分别在龙井、碧螺春、祁红和铁观音等4类茶叶样品中随机挑选20、20、10和20个样本作为训练集，对每一类茶叶样品分别建立SIMCA模型，剩下的38个样本作为预测样本，用来检验模型的可靠性。

按照公式(1)对已知样本进行主成分分解，通过交互验证来确定上述4类茶叶模型的最佳主成分数，即在预测残差平方和(PRESS)变化不大的情况下选取比较小的主成分数。图4表示在不同的主成分数下，4类模型的预测残差平方和(PRESS)如图3所示。由图3可以看出：龙井、碧螺春、祁红和铁观音等四类模型选取的最佳主成分数

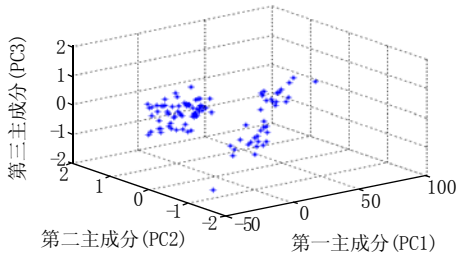

图2 所有茶叶样本的前3个主成分得分示意图

Fig.2 Score cluster plot using top three principal components (PCs) for all tea samples

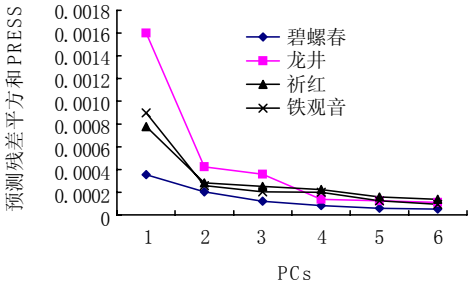

图3 PRESS值与模型主成分数之间的关系

Fig.3 The relationship between PRESS and principal components in model

分别为4、5、2、3。

2.4 训练和预测的结果

图4是训练集光谱矩阵第一主成分和第二主成分得分图，表明校正集中的样本点在该二维平面上的投影。训练集光谱矩阵的第一主成分与第二主成分的方差贡献率分别为69%与21%，累加方差贡献率达到了90%，所以样本点在该二维平面上的投影分布可以充分表征样本在超维空间中的分布特征。从图4中可以看出，四种茶叶基本上可以分别开来。龙井茶与碧螺春茶之间的距离相对较近，其次它们与铁观音茶距离相对较远，而与祁红的距离相对最远。图4中四种茶叶的分布现象可以用以下原因解释：首先是加工工艺的原因，龙井茶与碧螺春茶同属于绿茶，铁观音属于乌龙茶，祁红属于红茶，绿茶为非发酵茶，乌龙茶是轻发酵茶，红茶属于全发酵茶；加工工艺不同，它们内部成分(如多酚类和芳香类物质)的氧化降解程度也不一样，而这些内部物质大都在近红外光谱上会有不同的吸收峰。内部物质含量不同，它在近红外光谱上的吸收峰也有一定的差异，差异越大，它们在空间的距离就越大，内部品质也就越大。其次是地理位置、气候、土壤等外在条件的原因，这些外在条件对茶叶内部品质也存在一定的影响，虽然没有加工工艺对茶叶内部品质影响明显，但是这种差别在近红外光谱中还是可以表现出来，所以碧螺春茶和龙井茶在空间也存在一定的距离。从以上的分析中也

就可以推断：四种茶中，龙井茶与碧螺春茶的品质相对一致，其次与铁观音茶，与祁红的品质相差最大。

在显著性水平α=1%条件下，所建立的预测模型性能最佳训练和预测的结果如表1所示。从表1中可以看出模型在训练过程中只有龙井模型的识别率为90%，其余模型的识别率都达到100%。而在预测时，只有碧螺春模型的识别率是80%，其余都为100%，对于SIMCA模型对非己类样本的拒绝率，四类模型都达到100%。

表1 训练和预测的结果

Table 1 The results of experiments by calibration and prediction

茶叶品种	训练集		预测集	
	识别率(%)	拒绝率(%)	识别率(%)	拒绝率(%)
龙井	90	100	100	100
碧螺春	100	100	80	100
祁红	100	100	100	100
铁观音	100	100	100	100

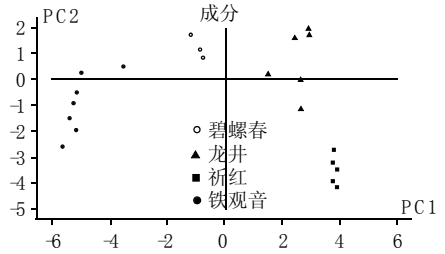

图4 训练集样本第一主成分和第二主成分得分图

Fig.4 Score cluster plot using first and second principal component (PC) for samples in calibration set

3 结论与展望

本实验利用近红外光谱方法识别了四种茶叶，在主成分分析的基础上利用SIMCA模式识别原理对四种茶叶分别建立了类模型，模型基本能正确识别这四种茶叶，在对其它类型茶叶识别时，拒绝率达到100%。结论充分表明了近红外光谱结合SIMCA模式识别方法在茶叶分类识别中的可行性。与人工感官识别方法相比，本文所提出的茶叶类型识别方法具有识别准确率高、自动化程度强、识别方法可扩充性好和适用范围广等优点。

由于茶叶样本受到储存时间和储存条件影响，内部成分将会有不同程度的变化。所以在SIMCA方法建立识别茶叶的类模型时，要充分考虑到训练集样本和预测集样本的一致性。另外，在SIMCA模式识别中，茶叶类模型的建立基本是利用了线形判别的方法，茶叶识别的结果不一定能做到100%的正确。所以在以后的工作中还可以尝试利用其它非线性的模式识别方法来完善模型，使结果更准确。

参考文献：

[1] J Lupaert, M H Zhang, D L Massart. Feasibility study for the using near infrared spectroscopy in the qualitative and quantitative of green tea, *Camellia sinensis (L.)*[J]. Analytica Chemica Acta, 2003, 487(2): 303-312.

[2] M H Zhang, J Lupaert, Q S Xu, et al. Determination of total antioxidant capacity in green tea by NIRS and multivariate calibration[J]. Talanta, 2004, 62(1): 25-35.

[3] H Schulz, U H Engelhardt, A Wengent, et al. Application of NIRS to the simultaneous prediction alkaloids and phenolic substance in green tea leaves[J]. J Agric Food Chem, 1999, 47: 5064-5067.

[4] 王丽, 卓林, 何鹰, 等. 近红外光谱技术识别海面溢油[J]. 光谱学与光谱分析, 2004, 24(12): 1537-1539.

[5] 陆婉珍, 袁洪福, 徐广通, 等. 现代近红外光谱分析[M]. 北京: 中国石化出版社, 2000. 188.

[6] 许禄, 邵学广. 化学计量学方法[M]. 北京: 科学出版社, 2004.

[7] 周群, 孙素琴, 梁曦云. 枸杞产地的红外指纹图谱与聚类分析方法研究[J]. 光谱学与光谱分析, 2003, 23(3): 50-511.

[8] A Candofi, R De Maesschalck, D L Massart, et al. Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA[J]. Journal of Pharmaceutical and Biomedical Analysis, 1999, 19: 923-935.

[9] 刘颖荣, 许育鹏, 杨海鹰, 等. 汽油样品类型的模式识别研究与应用[J]. 色谱, 2004, 22(5): 482-485.

信 息